CENTRO DE CIENCIAS BÁSICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

TESIS

DISEÑO Y DESARROLLO DE UN AGENTE DE INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ENTORNOS DIGITALES

PRESENTA

Manuel González Martínez

PARA OBTENER EL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

TUTORAS

Dra. Aurora Torres Soto
Dra. María Dolores Torres Soto

INTEGRANTE DEL COMITÉ TUTORAL

Dr. Eduardo Emmanuel Rodríguez López

Aguascalientes, Ags., 13 de junio de 2026

CARTA DE VOTO APROBATORIO

MTRO. GUILLERMO DOMÍNGUEZ AGUILAR
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

_____
P R E S E N T E

Por medio del presente como ***DIRECTORA*** designada del estudiante **MANUEL GONZÁLEZ MARTÍNEZ** con ID 2*28552* quien realizó *la tesis* titulada: **DISEÑO Y DESARROLLO DE UN AGENTE DE INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ENTORNOS DIGITALES**, un trabajo propio, innovador, relevante e inédito y con fundamento en la fracción IX del Artículo 43 del Reglamento General de Posgrados, doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que ***él*** **pueda** continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 15 de JUNIO de 2026.

DRA. AURORA TORRES SOTO
Directora

c.c.p.- Interesado
c.c.p.- Coordinación del Programa de Posgrado

CARTA DE VOTO APROBATORIO

MTRO. GUILLERMO DOMÍNGUEZ AGUILAR
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

_____
P R E S E N T E

Por medio del presente como ***CODIRECTORA*** designada del estudiante **MANUEL GONZÁLEZ MARTÍNEZ** con ID 2*28552* quien realizó *la tesis* titulada: **DISEÑO Y DESARROLLO DE UN AGENTE DE INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ENTORNOS DIGITALES**, un trabajo propio, innovador, relevante e inédito y con fundamento en la fracción IX del Artículo 43 del Reglamento General de Posgrados, doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que ***él*** **pueda** continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
“Se Lumen Proferre”
Aguascalientes, Ags., a 15 de JUNIO de 2026.

DRA. MARIA DOLORES TORRES SOTO
Codirectora

c.c.p.- Interesado
c.c.p.- Coordinación del Programa de Posgrado

MTRO. GUILLERMO DOMÍNGUEZ AGUILAR
DECANO DEL CENTRO DE CIENCIAS BÁSICAS

_____

P R E S E N T E

Por medio del presente como ***ASESOR*** designado del estudiante ***MANUEL GONZÁLEZ MARTÍNEZ*** con ID ***228552*** quien realizó la tesis titulada: ***DISEÑO Y DESARROLLO DE UN AGENTE DE INTELIGENCIA ARTIFICIAL PARA LA DETECCIÓN DE NOTICIAS FALSAS EN ENTORNOS DIGITALES***, un trabajo propio, innovador, relevante e inédito y con fundamento en la fracción IX del Artículo 43 del Reglamento General de Posgrados, doy mi consentimiento de que la versión final del documento ha sido revisada y las correcciones se han incorporado apropiadamente, por lo que me permito emitir el **VOTO APROBATORIO**, para que él pueda continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
“Se Lumen Proferre”
Aguascalientes, Ags., a 15 de junio de 2026.

Eduardo Emmanuel Rodríguez López
Asesor de tesis

c.c.p.- Interesado
c.c.p.- Coordinación del Programa de Posgrado

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

DICTAMEN DE LIBERACIÓN ACADÉMICA
PARA INICIAR LOS TRÁMITES DEL EXAMEN DE GRADO

POSGRADOS UAA

Fecha de dictaminación (dd/mm/aaaa): 20/07/2026

NOMBRE: MANUEL GONZÁLEZ MARTÍNEZ **ID** 228552

PROGRAMA: MAESTRÍA EN CIENCIAS CON OPCIONES A LA COMPUTACIÓN, MATEMÁTICAS APLICADAS **LGAC (del posgrado):** Inteligencia Artificial

MODALIDAD DEL PROYECTO DE GRADO: Tesis Tradicional (X) *Tesis por artículos científicos ( ) **Tesis por Patente ( ) Trabajo Práctico ( )

TITULO: Diseño y desarrollo de un agente de Inteligenia Artificial para la detección de noticias falsas en entornos digitales.

IMPACTO SOCIAL (señalar el impacto logrado): Contribuye a combatir la desinformación digital mediante un agente de IA auditable que apoya la verificación crítica de contenidos y promueve el uso responsable de la tecnología en la sociedad.

INDICAR SEGÚN CORRESPONDA: **SI, NO, NA (No Aplica)**

	Elementos para la revisión académica del trabajo de tesis o trabajo práctico:
SI	El trabajo es congruente con las LGAC del programa de posgrado
NO	La problemática fue abordada desde un enfoque multidisciplinario
SI	Existe coherencia, continuidad y orden lógico del tema central con cada apartado
SI	Los resultados del trabajo dan respuesta a las preguntas de investigación o a la problemática que aborda
SI	Los resultados presentados en el trabajo son de gran relevancia científica, tecnologíca o profesional según el área
SI	El trabajo demuestra más de una aportación original al conocimiento de su área
SI	Las aportaciones responden a los problemas prioritarios del país
SI	Generó transferencia del conocimiento o tecnológica
SI	Cumple con la ética para la investigación (reporte de la herramienta antiplagio)
	El egresado cumple con lo siguiente:
SI	Cumple con lo señalado por el Reglamento General de Posgrados
SI	Cumple con los requisitos señalados en el plan de estudios (créditos curriculares, optativos, actividades complementarias, estancia, predoctoral, etc.)
SI	Cuenta con los votos aprobatorios del comité tutorial
NA	Cuenta con la carta de satisfacción del Usuario (En caso de que corresponda)
SI	Coincide con el título y objetivo registrado
SI	Tiene congruencia con cuerpos académicos
SI	Tiene el CVU de la SECIHTI actualizado
NA	Tiene el o los artículos aceptados o publicados y cumple con los requisitos institucionales (en caso de que proceda)
	***En caso de Tesis por artículos científicos publicados** (completar solo si la tesis fue por artículos)*
NA	Aceptación o Publicación de los artículos en revistas indexadas de alto impacto según el nivel del programa
NA	El (la) estudiante es el primer autor(a)
NA	El (la) autor(a) de correspondencia es el Director (a) del Núcleo Académico
NA	En los artículos se ven reflejados los objetivos de la tesis, ya que son producto de este trabajo de investigación.
NA	Los artículos integran los capítulos de la tesis y se presentan en el idioma en que fueron publicados
	***En caso de Tesis por Patente**
NA	Cuenta con la evidencia de solicitud de patente en el Departamento de Investigación (anexarla al presente formato)

Con base en estos criterios, se autoriza continuar con los trámites de titulación y programación del examen de grado:

Sí X

No

FIRMAS

Elaboró:

*NOMBRE Y FIRMA DEL(LA) CONSEJERO(A) SEGÚN LA LGAC DE ADSCRIPCION: Dr. Rogelio Salinas Gutierres

** En caso de conflicto de intereses, firmará un revisor miembro del NA de la LGAC correspondiente distinto al director o miembro del comité tutorial, asignado por el Decano.*

NOMBRE Y FIRMA DEL COORDINADOR DE POSGRADO: Dra. Mariana Alfaro Gómez

Revisó:

NOMBRE Y FIRMA DEL SECRETARIO DE INVESTIGACIÓN Y POSGRADO: Dra. Iliana Ernestina Medina Ramirez

Autorizó:

NOMBRE Y FIRMA DEL DECANO: Mtro. Guillermo Domínguez Aguilar

Nota: procede el trámite para el Depto. de Apoyo al Posgrado

En cumplimiento con el Art. 24 fracción V del Reglamento General de Posgrado, que a la letra señala entre las funciones del Consejo Académico: Proponer criterios y mecanismos de selección, permanencia, egreso y titulación de estudiantes para asegurar la eficiencia terminal y la titulación y el Art. 28 fracción IX, atender, asesorar y dar el seguimiento del estudiantado desde su ingreso hasta su titulación.

Agradecimientos

Agradezco a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por el apoyo económico otorgado mediante beca de posgrado nacional, que hizo posible la realización de mis estudios y el desarrollo de esta investigación.

Agradezco a la Universidad Autónoma de Aguascalientes, al Centro de Ciencias Básicas y al Departamento de Ciencias de la Computación por el marco académico, los recursos institucionales y las condiciones que permitieron llevar a cabo este trabajo recepcional.

Agradezco a la Dra. Aurora Torres Soto y a la Dra. María Dolores Torres Soto por su dirección, revisión crítica y acompañamiento durante la elaboración de la tesis.

Agradezco al Dr. Eduardo Emmanuel Rodríguez López no sólo por sus observaciones y apoyo como integrante del Comité Tutoral, sino como amigo y piedra angular durante mi formación académica. Su esfuerzo y compromiso han sido un referente para mí del tipo de persona que quiero llegar a ser. Tipazo.

Agradezco de todo corazón a mi pareja, Nubia Itzel, por su compañía durante estos años. Me escuchó cuando el trabajo pesaba, me animó cuando dudaba y estuvo presente en los momentos en que más necesitaba apoyo. Sin ella no estaría donde estoy y no sería quien soy.

Finalmente quiero agradecer a mi señora madre por siempre cuidar y apoyar a su joven hijo, a mi papá y a mis hermanas, son la mejor familia que hubiera podido tener.

Índice general

Índice de tablas

Índice de figuras

Acrónimos

Sigla	Descripción
API	Interfaz de programación de aplicaciones (*application programming interface*).
C2PA	*Coalition for Content Provenance and Authenticity*, estándar abierto para registrar el origen y las transformaciones de un contenido digital.
F1	Media armónica de precisión y exhaustividad (*recall*).
GGUF	Formato de archivo para distribuir y ejecutar localmente modelos de lenguaje cuantizados.
LLM	Modelo de lenguaje de gran escala (*large language model*).
NLI	Inferencia de lenguaje natural (*natural language inference*).
OCR	Reconocimiento óptico de caracteres (*optical character recognition*).
QAT	Entrenamiento consciente de la cuantización (*quantization-aware training*).
RAG	Generación aumentada por recuperación (*retrieval-augmented generation*).

Resumen en español

Este estudio diseña, implementa y evalúa un agente conversacional de inteligencia artificial para el análisis auditable de noticias falsas en entornos digitales. El problema abordado no se reduce a clasificar contenidos como verdaderos o falsos, sino a hacer explícito qué afirmación fue analizada, qué evidencia se recuperó, qué fuentes fueron usadas, qué comparaciones claim-evidence sostienen la respuesta y qué limitaciones permanecen abiertas. El prototipo se organiza bajo un principio de auditabilidad: el modelo de lenguaje no opera como juez autónomo de verdad, sino como componente de extracción, normalización, orquestación y redacción restringida por un contrato de evidencia. La metodología se organiza como un ciclo local de diseño, implementación, evaluación y revisión de calidad. La evaluación separa métricas primarias de auditabilidad y métricas secundarias de clasificación. En la evaluación pública extendida de 60 muestras, el agente obtuvo retrieval hit rate 1.0, traceability rate 1.0, unsupported-statement rate 0.0 y decisive-source rate 0.6. En clasificación secundaria, el comparador estricto basado en Gemma 4 E4B quedó por arriba: el agente obtuvo accuracy 0.4833, macro-F1 0.3454 y weighted-F1 0.47, frente a 0.5, 0.4633 y 0.5121 del comparador. Estos resultados respaldan la contribución como infraestructura reproducible de revisión, no como reemplazo del verificador humano ni como clasificador líder en benchmarks. El trabajo concluye con límites técnicos en NLI, agregación conservadora, evidencia multimodal y latencia, además de una agenda de mejora para sistemas conversacionales auditables.

Palabras clave: Desinformación; noticias falsas; agentes conversacionales; auditabilidad; fact-checking; recuperación de evidencia; NLI; grounding; trazabilidad; modelos de lenguaje.

Abstract

This thesis designs, implements, and evaluates an AI-based conversational agent for auditable analysis of fake news in digital environments. The problem is not framed as a single true-or-false classification task, but as the need to expose which claim was analyzed, which evidence was retrieved, which sources were used, which claim-evidence comparisons support the public response, and which limitations remain unresolved. The prototype is guided by a principle of auditability: the language model is not treated as an autonomous truth judge, but as a component for extraction, normalization, orchestration, and constrained drafting based on a structured evidence contract. The methodology follows a local cycle of design, implementation, evaluation, and documented review. Evaluation separates primary auditability metrics from secondary classification metrics. In the extended public evaluation over 60 samples, the agent reached retrieval hit rate 1.0, traceability rate 1.0, unsupported-statement rate 0.0, and decisive-source rate 0.6. In secondary classification, the strict Gemma 4 E4B comparator performed better: the agent reached accuracy 0.4833, macro-F1 0.3454, and weighted-F1 0.47, compared with 0.5, 0.4633, and 0.5121 for the comparator. These results support the contribution as a reproducible review infrastructure rather than a replacement for human verification or a benchmark-leading classifier. The thesis closes with technical limitations in NLI, conservative aggregation, multimodal evidence, and latency, and outlines future work for auditable conversational systems.

Keywords: misinformation; fake news; conversational agents; auditability; fact-checking; evidence retrieval; NLI; grounding; traceability; language models.

1. Introducción

1.1. Contexto

La circulación de información en entornos digitales ha transformado la manera en que las personas se enteran de acontecimientos políticos, sociales, científicos y culturales. La misma infraestructura que permite acceso inmediato a fuentes diversas también facilita la propagación rápida de contenido engañoso, impreciso o directamente falso. En este contexto, el problema no se limita a distinguir entre una noticia verdadera y una falsa. Muchas entradas reales combinan afirmaciones parciales, imágenes reutilizadas, titulares ambiguos, fuentes de calidad desigual, fechas desplazadas y contextos incompletos. Una herramienta útil para analizar desinformación debe preservar esa complejidad en vez de reducirla prematuramente a una etiqueta opaca.

Los modelos de lenguaje han ampliado las posibilidades de construir asistentes capaces de resumir, buscar, comparar y explicar información. Sin embargo, una respuesta conversacional fluida puede ocultar el proceso que la produjo. Cuando un sistema afirma que una noticia es falsa, verdadera o inconclusa, el usuario necesita saber qué fue verificado, qué evidencia se encontró, qué evidencia no se encontró, qué fuente fue usada y qué parte de la respuesta proviene de inferencia o de una limitación. Sin esa trazabilidad, la automatización puede trasladar el problema: en lugar de confiar ciegamente en una noticia, el usuario termina confiando ciegamente en una respuesta generada.

Este estudio parte de una postura conservadora: un agente de inteligencia artificial no debe actuar como juez absoluto de verdad. Su papel metodológico es orquestar técnicas existentes, registrar un rastro de evidencia y redactar una respuesta que haga visibles sus propios límites. Por ello, el proyecto no busca entrenar un nuevo modelo, crear un conjunto de datos propio ni competir por liderazgo en métricas de clasificación. El objetivo es diseñar y evaluar una arquitectura aplicada para análisis auditable de desinformación.

1.2. Problema de investigación

El problema central consiste en diseñar, implementar y evaluar un agente conversacional capaz de analizar texto, enlaces e imágenes mediante un flujo de evidencia trazable. El agente debe combinar técnicas existentes de detección de desinformación, recuperación de evidencia, comparación semántica, señales de imagen y generación de respuestas explicables, sin permitir que el modelo de lenguaje invente evidencia o com-

plete huecos con conocimiento no recuperado.

La delimitación inicial define el problema en términos de evidencia y auditabilidad: el sistema debe registrar afirmaciones, fuentes, técnicas aplicadas, resultados, incertidumbre y limitaciones. Esta definición desplaza el énfasis desde la clasificación aislada hacia la reproducibilidad del análisis. Una etiqueta puede ser útil, pero sólo si puede conectarse con la evidencia que la sostiene o con la limitación que impide sostenerla.

En consecuencia, la detección de noticias falsas que anuncia el título se entiende en este trabajo en un sentido preciso: no como la emisión de un veredicto cerrado de verdad o falsedad, sino como un proceso de análisis auditable en el que el sistema detecta qué afirmación está en disputa, qué evidencia la sostiene o la contradice y qué límites condicionan la conclusión. Bajo esta lectura, la clasificación es un resultado secundario subordinado a la trazabilidad de la evidencia, y no el objetivo último del agente.

1.3. Objetivo general

Diseñar, desarrollar y evaluar un agente conversacional orientado a producir análisis revisables de contenidos digitales potencialmente desinformativos, priorizando trazabilidad, evidencia y límites explícitos sobre la decisión automática de verdad.

1.4. Objetivos específicos

1. Analizar el problema de revisión que surge cuando una respuesta conversacional presenta conclusiones sobre desinformación sin mostrar con claridad evidencia, alcance e incertidumbre.

2. Definir los elementos mínimos que debe conservar un análisis revisable: afirmación analizada, fuentes consultadas, evidencia disponible, relación entre evidencia y afirmación, incertidumbres y límites.

3. Diseñar y desarrollar un prototipo de agente conversacional que organice el análisis en etapas trazables y formule respuestas condicionadas por la evidencia disponible.

4. Evaluar el prototipo comparando respuestas con y sin rastro explícito de evidencia, considerando trazabilidad, soporte de afirmaciones, claridad de límites y compatibilidad de etiqueta.

5. Identificar los límites técnicos y metodológicos del enfoque ante evidencia parcial, ambigua, multimodal o insuficiente, y derivar líneas de mejora.

1.5. Preguntas de investigación

La pregunta central de investigación es:

¿Cómo puede un agente conversacional de inteligencia artificial apoyar la revisión crítica de contenidos digitales potencialmente desinformativos cuando la confianza depende de evidencia trazable y no sólo de una etiqueta final?

De esa pregunta central se derivan cinco preguntas específicas, cada una vinculada con un objetivo específico:

1. ¿Qué condiciones hacen que una respuesta conversacional sobre desinformación digital sea difícil de revisar o auditar?
2. ¿Qué información necesita conservarse para que el análisis de una afirmación digital pueda ser revisado críticamente por una persona?
3. ¿Cómo puede organizarse un agente conversacional para producir respuestas útiles sin presentarse como autoridad final de verdad?
4. ¿Qué aporta un rastro explícito de evidencia frente a una respuesta conversacional que sólo entrega una explicación final?
5. ¿Qué límites aparecen cuando el contenido analizado combina texto, enlaces, imágenes o evidencia incompleta?

1.6. Hipótesis de trabajo

Las hipótesis se formulan antes del diseño experimental y funcionan como puente entre el problema, la literatura y la evaluación. No anticipan resultados; delimitan qué cambios deberían observarse si un agente conversacional conserva evidencia estructurada en lugar de responder sólo con texto generado.

1. H1. Un expediente estructurado de evidencia debe mejorar la trazabilidad de las respuestas frente a una respuesta conversacional sin contrato explícito.
2. H2. Separar la recuperación de evidencia debe mejorar la cobertura de fuentes pertinentes, aunque no garantice siempre una fuente decisiva.
3. H3. Incorporar comparación entre afirmación y evidencia mediante señales de *stance* y NLI debe reducir etiquetas sobreconfiadas cuando la evidencia es débil, neutral o contradictoria.

4. H4. Registrar señales visuales como OCR, metadatos, procedencia y disponibilidad debe mejorar la explicación multimodal sin convertir esas señales en veredictos automáticos.

5. H5. Las métricas de auditabilidad, atribución y soporte de afirmaciones deben describir mejor la contribución del prototipo que accuracy o F1 por sí solas.

1.7. Alcance y restricciones

El alcance incluye texto, enlaces e imágenes. Excluye video, forense avanzado de deepfakes, pruebas con usuarios, creación de conjuntos de datos propios, entrenamiento o ajuste fino de modelos y cualquier afirmación de liderazgo en benchmarks. Las señales de imagen, procedencia, C2PA, metadatos o marcas se tratan como contexto, no como prueba directa de verdad o falsedad. La ausencia de evidencia nunca se convierte automáticamente en falsedad.

Estas restricciones no son limitaciones accidentales, sino decisiones metodológicas. El objetivo es aislar el valor del flujo de evidencia: si el sistema aporta algo, debe verse en la calidad del rastro, no en una promesa de omnisciencia.

La exclusión de pruebas con usuarios merece una justificación específica. La pregunta central de esta tesis es si un agente puede producir un expediente trazable y auditable, una propiedad del artefacto que se mide sobre los objetos persistidos —contratos, evidencia, comparaciones y límites— sin requerir participantes. Un estudio con usuarios respondería una pregunta distinta y posterior: si ese expediente reduce el tiempo de revisión o mejora las decisiones de un verificador humano. Esa evaluación exige diseño experimental con sujetos, control de sesgos, consentimiento y validación de instrumentos, condiciones que constituyen un trabajo en sí mismo. Incorporarla aquí diluiría el foco y mezclaría la validación técnica del expediente con la validación de su utilidad práctica. Por ello la utilidad para un revisor real se mantiene como hipótesis a validar en trabajo futuro y no como afirmación del estudio.

1.8. Contribuciones

El estudio aporta:

1. Una arquitectura orientada a la auditabilidad para análisis conversacional de desinformación.

2. Un contrato de evidencia que separa afirmaciones, evidencia, comparaciones, señales, limitaciones y respuesta final.

3. Un prototipo funcional que persiste artefactos auditables y valida el soporte en evidencia de las respuestas.

4. Un diseño de evaluación que compara el agente contra un comparador no estructurado.

5. Una interpretación explícita de resultados: la contribución fuerte está en trazabilidad y control de afirmaciones no soportadas; la clasificación permanece como métrica secundaria con resultados mixtos.

1.9. Profundización del problema y alcance

El planteamiento anterior fijó el problema, los objetivos y las preguntas en términos operativos. Antes de pasar al marco teórico conviene desarrollar con más detalle por qué la desinformación digital se aborda aquí como un problema de trazabilidad y qué clase de contribución puede sostener un trabajo aplicado de esta naturaleza. Las cuatro subsecciones siguientes precisan ese argumento desde ángulos complementarios: la naturaleza del problema, el papel de la interfaz conversacional, la justificación del alcance multimodal y la forma de la contribución.

1.9.1. Desinformación como problema de trazabilidad

La desinformación digital no se manifiesta sólo como falsedad explícita. También aparece como recontextualización, omisión, ambigüedad, mezcla de señales visuales y uso parcial de fuentes. Por ello, el problema de investigación no puede reducirse a asignar una etiqueta final: requiere explicar cómo conservar el texto original, la afirmación normalizada, las fuentes consultadas, las comparaciones y las limitaciones que condicionan cualquier respuesta [6], [7], [8].

La formulación del problema distingue tres niveles: la afirmación que se analiza, las operaciones que producen evidencia y la redacción pública que comunica el resultado. Si esos niveles se mezclan, una coincidencia textual puede parecer soporte, una fuente reputada puede parecer decisiva aunque no cubra el caso, o una ausencia de resultados puede interpretarse erróneamente como refutación. El estudio evita esas sobreafirmaciones al presentar el prototipo como infraestructura de análisis revisable, no como clasificador universal.

1.9.2. Agente conversacional como interfaz de análisis

Un agente conversacional aporta valor cuando convierte operaciones técnicas en una explicación revisable por una persona. La interfaz conversacional no reemplaza la

auditoría; debe resumir un expediente estructurado y permitir que el lector cuestione la evidencia, las fuentes y los límites de la conclusión [1], [9], [3].

El modelo de lenguaje se presenta, por tanto, como componente restringido de extracción, organización y redacción. Su fluidez no equivale a evidencia recuperada. La utilidad del sistema se evalúa por la posibilidad de revisar la ruta de análisis y no por la seguridad retórica de la respuesta final.

1.9.3. Justificación del alcance multimodal

Incluir texto, enlace e imagen es necesario porque muchas piezas de desinformación combinan afirmaciones lingüísticas con contexto visual. El alcance multimodal de este estudio no promete análisis forense avanzado: registra OCR, metadatos, C2PA, descripciones visuales y limitaciones como señales separadas del veredicto [10], [11], [12].

Los estudios de verificación imagen-texto refuerzan esta cautela porque muestran que una señal visual puede ser útil sin ser decisiva para resolver la afirmación completa [4], [13].

La ausencia de imagen original o de metadatos completos reduce la fuerza probatoria de la señal visual. Por eso la multimodalidad se entiende como preservación de señales y restricciones, no como resolución automática de todo contenido visual.

1.9.4. Preguntas de investigación como controles

Las preguntas de investigación funcionan como controles contra el desplazamiento del objetivo. No parten de una arquitectura predeterminada ni de una métrica elegida de antemano; parten de una preocupación más básica: cuándo una respuesta sobre desinformación puede ser revisada críticamente por una persona. Esa formulación permite que el contrato de evidencia, la arquitectura del agente y la evaluación aparezcan después como decisiones metodológicas para responder a las preguntas, no como supuestos iniciales.

Las preguntas evitan prometer verdad final o resolución total de la desinformación. En cambio, preguntan por revisabilidad, información necesaria, organización del agente, aporte del rastro de evidencia y límites del enfoque. Esa secuencia sostiene una tesis aplicada sin convertir el planteamiento en una lista de módulos técnicos.

Una pregunta demasiado amplia llevaría a promesas que el prototipo no puede sostener, mientras que una pregunta demasiado técnica haría parecer que la investigación se redactó después de conocer la solución. Por eso el capítulo conserva preguntas generales pero evaluables, y reserva los detalles técnicos para metodología, diseño y evaluación.

1.9.5. Contribución como infraestructura de revisión

La contribución central es una infraestructura de revisión, no un nuevo modelo predictivo. El aporte se organiza alrededor del contrato de evidencia, la arquitectura del agente, el prototipo funcional y la evaluación de auditabilidad. Esa formulación mantiene el foco en la posibilidad de revisar el análisis, no en una promesa de clasificación universal [5], [14].

Accuracy y F1 se mantienen como diagnósticos secundarios. Presentar el prototipo como clasificador competitivo distorsionaría el alcance del trabajo: el estudio sitúa su contribución en la trazabilidad, la recuperación de evidencia y el control de afirmaciones no soportadas, y no en una promesa de superioridad predictiva global.

Con este planteamiento, el marco teórico no funciona como una lista de antecedentes, sino como el soporte que permite explicar por qué el análisis de desinformación exige recuperar, comparar, atribuir y comunicar evidencia con límites explícitos.

1.10. Organización

El Capítulo 2 revisa el marco teórico. El Capítulo 3 describe la metodología por etapas. El Capítulo 4 presenta el diseño del agente y el contrato de evidencia. El Capítulo 5 documenta la implementación. El Capítulo 6 reporta evaluación y resultados. El Capítulo 7 discute hipótesis, alcances y límites. El Capítulo 8 concluye y propone trabajo futuro.

Con esta organización, la introducción deja planteada la tensión central: una respuesta conversacional es útil sólo si permite revisar la evidencia que la sostiene.

2. Marco teórico y trabajo relacionado

La Figura 2.1 ubica la tesis dentro de la tradición de fact-checking explicable: el proceso no se limita a una etiqueta, sino que organiza subtareas observables y una explicación revisable.

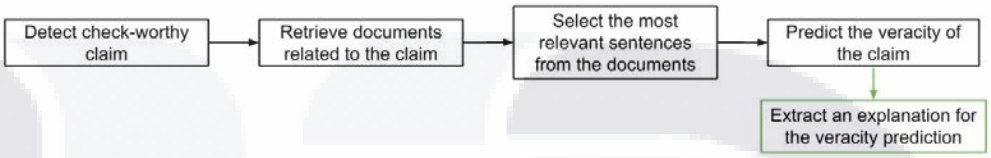

Figura 2.1: Pipeline de fact-checking automático con subtarea explicativa. [1].
Fuente: Explainable Automated Fact-Checking: A Survey, figura 1, p. 2.

2.1. Verificación automática y detección de afirmaciones

La verificación automática de hechos puede entenderse como una secuencia de tareas: identificar una afirmación verificable, recuperar evidencia relevante, comparar la afirmación con la evidencia y presentar una conclusión. La literatura revisada apoya separar estas tareas en lugar de delegarlas a una sola generación textual. La detección de afirmaciones es especialmente importante porque no todo fragmento de una publicación digital es verificable: puede haber opiniones, contexto narrativo, sarcasmo, declaraciones vagas o afirmaciones dependientes de fecha y lugar [7], [6].

En este estudio, *claim* extraction no es sólo una etapa de preprocesamiento. Es la entrada formal al contrato de evidencia. Si el sistema no conserva el texto original, el *claim* normalizado, el idioma, las entidades, fechas y limitaciones de extracción, entonces la evaluación posterior pierde trazabilidad. Por eso el diseño exige un registro explícito de la afirmación antes de cualquier integración técnica.

2.2. Recuperación de evidencia

La recuperación de evidencia determina buena parte de la calidad de la verificación. Un modelo puede redactar con fluidez, pero si la evidencia recuperada es irrelevante o incompleta, la respuesta será débil. Las fuentes revisadas sostienen que los sistemas de fact-checking requieren recuperar documentos, fragmentos o fuentes que puedan contrastarse contra el *claim* [15], [16]. Esa misma literatura señala que los recursos estructurados, como las verificaciones publicadas en el esquema ClaimReview o

el catálogo de la Fact Check API de Google, ofrecen evidencia de alta calidad pero con cobertura limitada: resultan útiles cuando una afirmación ya fue revisada por un verificador profesional, aunque no bastan para los muchos casos que carecen de antecedente publicado [17], [18].

De esa tensión se deriva una estrategia de recuperación por capas: usar primero fact-checks publicados cuando existan, recurrir a la web abierta cuando no haya antecedente, y registrar la ausencia de evidencia como una condición de insuficiencia en lugar de forzar una conclusión. El propósito es evitar que la recuperación opere como una caja negra. Para ello, cada evidencia debe conservar su fuente, URL, método de recuperación, fragmento revisable, fecha y limitaciones, de modo que un revisor pueda reconstruir por qué un documento se consideró pertinente o por qué se descartó.

2.3. Credibilidad, atribución y conflicto

La credibilidad de fuente ayuda a interpretar evidencia, pero no sustituye la comparación directa entre *claim* y evidencia. Una fuente reputada puede no cubrir todos los requisitos del *claim*; una fuente secundaria puede aportar contexto sin ser decisiva; dos fuentes pueden entrar en conflicto. Por ello, conviene distinguir credibilidad, pertinencia y suficiencia probatoria como dimensiones separadas.

Esta separación es central para la auditabilidad. La autoridad de una fuente sólo fortalece una conclusión cuando la fuente cubre los elementos factuales relevantes: fecha, ubicación, entidad, acción o magnitud. Si esos elementos no están cubiertos, la fuente puede ser contextual, parcial o insuficiente. La evaluación posterior puede medir si las citas públicas apuntan a fuentes decisivas y si las respuestas no sobreinterpretan fuentes revisadas.

2.4. Stance y NLI

Recuperar evidencia no basta. Un fragmento puede mencionar los mismos términos que el *claim* y, aun así, no apoyarlo ni contradecirlo. El estudio adopta una capa explícita de comparación *claim*-evidence que incorpora *stance*, NLI, requisitos cubiertos y requisitos faltantes. La literatura sobre similaridad semántica y NLI respalda tratar la inferencia textual como señal separada del ranking de *retrieval* [19], [20], [21].

El diseño conserva una cautela importante: NLI no es un juez final. Puede fallar en textos largos, evidencia ruidosa, traducciones o afirmaciones con condiciones temporales. Por eso el prototipo registra NLI como comparación auditada y conserva los spans seleccionados, en lugar de ocultar el razonamiento detrás de una etiqueta.

2.5. Desinformación multimodal

Las imágenes pueden contener texto, escenas, metadatos, marcas de procedencia o señales de reutilización. Sin embargo, ninguna de estas señales prueba por sí sola que una noticia sea falsa. El enfoque sintetiza esta idea al tratar OCR, descripción visual, búsqueda inversa, metadatos, provenance y watermarks como registros separados y no decisivos [12], [4], [13].

Esta decisión responde a un problema frecuente: confundir autenticidad de imagen con veracidad del *claim*. Una imagen puede ser real y usarse fuera de contexto; también puede ser generada y acompañar un *claim* verdadero por accidente. En consecuencia, las señales visuales deben integrarse como contexto probatorio, y las respuestas públicas deben explicar de forma explícita cuándo la evidencia visual no alcanza para resolver la afirmación.

2.6. Factualidad, citas y grounding

La literatura sobre respuestas con citas y evaluación de factualidad respalda medir si una respuesta generada está apoyada por evidencia. Este estudio incorpora esta línea mediante métricas de trazabilidad, calidad de atribución, consistencia afirmación-evidencia y tasa de afirmaciones no soportadas [5], [3], [14]. Estas métricas son más adecuadas que accuracy aislada para un estudio cuyo objetivo es analizar desinformación con evidencia revisable.

De esta literatura se desprende un requisito de diseño exigente: toda afirmación sustantiva de la respuesta final debe estar vinculada a evidencia o a una limitación explícita. Si el sistema no puede encontrar evidencia decisiva, debe decirlo; si una herramienta falla, debe registrarlo; si el resultado es inconcluso, debe conservar la razón de insuficiencia. La factualidad deja así de ser una propiedad difusa del texto y se vuelve una condición verificable enunciado por enunciado.

2.7. Brecha abordada

La revisión no identifica una solución única que cubra texto, enlaces, imágenes, *retrieval*, credibilidad, *stance*, NLI, provenance y *grounding* bajo un mismo contrato auditable en el entorno local del proyecto. La brecha abordada es aplicada: integrar técnicas existentes en un flujo modular, persistente y evaluable, donde la respuesta final pueda ser auditada. Esta brecha no requiere proponer un nuevo algoritmo de clasificación, sino articular una arquitectura y una evaluación coherentes con el objetivo de evidencia trazable.

2.8. Ejes transversales del trabajo relacionado

Cada decisión central del diseño del prototipo se apoya en un cuerpo específico de trabajos previos. Las ocho subsecciones siguientes recorren esas decisiones y enuncian, para cada una, el antecedente bibliográfico que la respalda y su traducción concreta hacia la arquitectura propuesta, de modo que ninguna elección del prototipo queda sin justificación en la literatura.

2.8.1. Pipelines de fact-checking frente a respuestas monolíticas

La primera decisión de diseño es descomponer la verificación en tareas observables (detección de afirmaciones, recuperación de evidencia, selección de fragmentos, comparación y comunicación) en lugar de delegarla a una sola generación textual. Esta arquitectura modular es la que recorre buena parte de la literatura de fact-checking [6], [7], [22]. Una respuesta monolítica de un modelo de lenguaje resulta insuficiente para un objetivo auditable porque no muestra por sí misma qué evidencia se recuperó, cómo se comparó ni qué limitaciones quedaron abiertas [15]. La descomposición no elimina los errores, pero los vuelve localizables, y separa con claridad dos familias de sistemas: los orientados a clasificación, que entregan una etiqueta, y los orientados a evidencia, que conservan una ruta de revisión.

2.8.2. Retrieval como condición necesaria pero no suficiente

La segunda decisión es tratar la recuperación de evidencia como una etapa propia y medible. Recuperar documentos relevantes es condición central para verificar una afirmación, pero la relevancia temática no garantiza suficiencia: un buscador puede devolver textos relacionados que no resuelven la fecha, el lugar, la entidad, la cantidad o la relación causal en disputa [15], [23]. De ahí surge la distinción, central para este trabajo, entre cobertura de recuperación y fuente decisiva [2], [16]. Esa diferencia es la que justifica que el contrato registre por separado los requisitos del *claim* que una fuente cubre y los que deja sin cubrir: una fuente relacionada puede orientar la revisión, pero sólo una fuente decisiva permite sostener o contradecir el núcleo de la afirmación.

2.8.3. Datasets evidence-rich y datasets label-only

La tercera decisión es elegir instrumentos de diagnóstico que conserven evidencia o justificaciones, y no colecciones que ofrezcan únicamente etiquetas [22], [24]. Por eso el estudio recurre a AVeriTeC, FEVER, MultiFC y AVerImaTeC como tensiones complementarias: combinan evidencia web abierta, fuentes controladas, diversidad de dominios y afirmaciones imagen-texto [23], [2], [4]. Cada conjunto tensiona una capa-

cidad distinta del sistema, pero ninguno agota la variabilidad del mundo abierto; por ello se asumen como sondas parciales y no como prueba de generalización. Esta postura condiciona la lectura posterior de los resultados, que deberán interpretarse como hallazgos acotados y no como demostración universal.

2.8.4. Credibilidad y conflicto de fuentes

La cuarta decisión separa tres dimensiones que suelen confundirse: credibilidad, pertinencia y suficiencia. La reputación de una fuente informa la interpretación, pero no reemplaza la comparación directa entre afirmación y evidencia, porque una fuente confiable puede referirse a otro momento, otra entidad o sólo una parte del caso [25], [26], [18]. El diseño hereda de esta literatura el requisito de registrar calidad de fuente, campos faltantes y contradicciones como atributos independientes, de modo que un conflicto no se resuelva por autoridad superficial. Cuando dos fuentes reputadas cubren aspectos distintos de una misma afirmación y el resultado queda abierto, la respuesta responsable debe conservar esa incertidumbre en lugar de cerrarla.

2.8.5. NLI, stance y sensibilidad a diferencias pequeñas

La quinta decisión incorpora una capa explícita de comparación semántica entre afirmación y evidencia. *Stance* y NLI son necesarios porque un documento puede mencionar una afirmación sin apoyarla ni contradecirla, y porque pequeñas diferencias de negación, sujeto o tiempo verbal alteran la relación lógica [19], [20], [21]. La literatura también advierte que esta capa es frágil: la inferencia puede degradarse en traducciones, textos largos o afirmaciones condicionadas temporalmente. De esa advertencia se sigue la decisión de tratar la inferencia como una señal auditada que se conserva junto a la fuente, el fragmento y los requisitos faltantes, y no como un veredicto autónomo que sustituya al juicio sobre la evidencia.

2.8.6. Multimodalidad sin atajos unimodales

La sexta decisión gobierna el tratamiento de la imagen. La literatura multimodal advierte que una señal visual puede parecer decisiva aunque la afirmación dependa de contexto temporal, textual o geográfico [27], [10], [11]. Por eso el diseño registra las señales visuales como contexto y mantiene separadas OCR, C2PA, metadatos, búsqueda contextual y limitaciones visuales [12], [4], [13]. La distinción operativa que se adopta es entre clasificación multimodal y verificación de afirmaciones imagen-texto: una imagen aislada no prueba verdad ni falsedad, y sólo aporta evidencia cuando se conecta con los requisitos factuales concretos del caso.

2.8.7. Grounding, atribución y evaluación de respuestas

La séptima decisión convierte la respuesta final en un objeto evaluable. La calidad de una salida generada no se mide por su fluidez, sino por su soporte en evidencia y por la atribución de cada afirmación a una fuente [5], [9]. Esta línea fundamenta el control de afirmaciones no soportadas que el prototipo aplica mediante la validación del contrato y la revisión de que la respuesta pública no introduzca enunciados sin respaldo [21], [3], [14]. La consecuencia práctica es que una cita visible no basta si no sostiene exactamente la afirmación redactada: la atribución debe verificarse a nivel de enunciado y no como adorno de referencias.

2.8.8. Herramientas aplicadas y periodismo asistido

La octava decisión define el rol del sistema frente a la persona. El ecosistema de herramientas de fact-checking asistido por inteligencia artificial insiste en el soporte humano, la trazabilidad y la revisión editorial [8], [17], [28]. El prototipo se inscribe en esa tradición: ofrece respuestas cautelosas, citas revisables y pasos de verificación, pero no sustituye al verificador humano. Esta posición fija además un límite metodológico del estudio: no incluye un experimento con usuarios ni una evaluación en sala de redacción, por lo que la utilidad práctica del expediente para un revisor real permanece como una afirmación que requiere validación posterior.

2.9. Mapa ampliado de literatura verificada

El corpus revisado admite además un ordenamiento por la función que cada familia de trabajos cumple dentro del problema. Esta cartografía documenta la amplitud de la base bibliográfica y precisa qué aporta cada grupo, de modo que la integración propuesta queda situada respecto del campo y no como una acumulación de referencias.

2.9.1. Panorama de verificación automática y datasets de referencia

El primer grupo de trabajos delimita el problema y proporciona sus instrumentos de medición. Las revisiones del área distinguen entre detección de desinformación, verificación de afirmaciones y evaluación de evidencia, y coinciden en describir la verificación automática como una tarea compuesta más que como una predicción aislada; los conjuntos de referencia, por su parte, muestran que la disponibilidad de evidencia cambia el tipo de inferencia que puede exigirse a un sistema [29], [30], [6]. Sobre esa base, otros trabajos separan con nitidez la detección de afirmaciones, la verificación con evidencia y la evaluación mediante datasets, y advierten que un benchmark no sustituye a una explicación revisable [7], [22], [31], [15]. De esta familia el estudio toma

una decisión de evaluación concreta: reportar etiquetas, pero siempre acompañadas de evidencia, justificación y límites, usando los datasets para ubicar el problema y no para prometer una solución total de la desinformación.

2.9.2. Evidencia, recuperación y selección de fragmentos

El segundo grupo aborda cómo se obtiene y se selecciona la evidencia. La recuperación aparece como condición necesaria de cualquier respuesta revisable, pero la misma literatura advierte que recuperar documentos relacionados no equivale a obtener evidencia suficiente; la selección de fragmentos, el soporte explicable y la intervención humana son los mecanismos que convierten resultados de búsqueda en insumos auditables [24], [32], [33]. Trabajos adicionales del grupo profundizan en la recuperación explicable, el ordenamiento de evidencia y las condiciones de uso humano del material recuperado, y sostienen que conviene conservar fuentes y fragmentos, no sólo el resultado de una consulta [19], [1], [34], [8]. La lectura que el estudio extrae es registrar fuentes y *spans* como objetos auditables: una búsqueda sólo se considera útil si permite inspeccionar qué parte de la fuente sostiene, contradice o deja abierta la afirmación.

2.9.3. Factualidad, atribución y respuestas generadas

El tercer grupo evalúa la respuesta misma. Los trabajos sobre factualidad y atribución muestran que una salida generada debe juzgarse por la relación entre sus afirmaciones y sus fuentes; esta línea es central para el estudio porque desplaza el interés desde la fluidez del texto hacia la posibilidad de inspeccionar soporte, contradicción, conflicto de fuentes y límites de cobertura [35], [36], [5]. Varios de estos trabajos insisten en que la atribución debe revisarse al nivel de cada afirmación y no como presencia decorativa de enlaces, lo que obliga a separar conflicto de fuentes, credibilidad y soporte textual efectivo [9], [25], [26], [37]. De aquí proviene la decisión de medir la respuesta final y no sólo el *retrieval*, e incorporar un control de afirmaciones no soportadas, porque una respuesta puede citar fuentes y aun así introducir conclusiones que esas fuentes no autorizan.

2.9.4. Multimodalidad, procedencia y verificación visual

El cuarto grupo introduce la imagen como fuente de ambigüedad. Una imagen puede ser auténtica y estar fuera de contexto, ser sintética y acompañar una afirmación verificable por otros medios, o carecer de metadatos suficientes; por eso la literatura visual se incorpora como contexto probatorio y no como sustituto de la comparación textual [38], [39], [27]. Trabajos de esta familia ubican la desinformación multimodal, los datasets imagen-texto y la evaluación visual como problemas cercanos pero no equi-

valentes, y muestran cuándo conviene registrar la imagen como señal contextual ante la falta de evidencia para un veredicto visual fuerte [40], [10], [11], [41]. La aplicación directa es separar señal visual de veredicto: OCR, búsqueda inversa, metadatos o procedencia pueden fortalecer una hipótesis, pero no sustituyen la comparación del *claim* con evidencia suficiente.

2.9.5. Procedencia, marcas de agua, metadatos y estándares abiertos

El quinto grupo aporta vocabularios estructurados para registrar el origen de un contenido. Las propuestas de procedencia y los esquemas de fact-checking permiten anotar origen, acciones, actores y verificaciones previas, y el estudio los integra como señales que pueden fortalecer o limitar la respuesta pública en lugar de como veredictos automáticos [39], [42], [12]. Estándares y catálogos como ClaimReview, los metadatos de plataforma y las marcas de procedencia funcionan bien como estructura de registro, pero siguen siendo insuficientes si no se conectan con la afirmación factual evaluada [43], [17], [18], [44]. El diseño emplea estos estándares como vocabularios de trazabilidad: la procedencia se registra para explicar origen, transformaciones y disponibilidad de credenciales, sin convertirla en garantía de veracidad.

2.9.6. Inferencia textual, contradicción y razonamiento multi-salto

El sexto grupo se ocupa de la relación lógica entre afirmación y evidencia. La comparación exige inferencia textual, detección de contradicción y, en ocasiones, razonamiento multi-salto, porque dos textos pueden compartir vocabulario y aun así no resolver la afirmación analizada; de ahí la utilidad de registrar requisitos cubiertos y faltantes [20], [45], [46]. Trabajos adicionales incorporan contradicción, fact-checking abierto y razonamiento sobre evidencia encadenada, y refuerzan la decisión metodológica de tratar la inferencia como señal revisable y no como decisión final automática [47], [48], [23], [2]. Por eso la comparación semántica se mantiene separada del ordenamiento de búsqueda: permite revisar si una fuente realmente implica el *claim*, si lo contradice, si sólo comparte vocabulario o si exige un razonamiento que el sistema no puede sostener.

2.9.7. Agregación de evidencia, recuperación compleja y datos tabulares

El séptimo grupo plantea un problema epistémico: combinar fuentes débiles no produce automáticamente una conclusión fuerte. Los trabajos sobre recuperación compleja, datos tabulares y razonamiento multifuente ayudan a delimitar cuándo una evidencia cubre el *claim* y cuándo sólo aporta contexto [49], [28], [50]. Otros amplían el problema hacia la búsqueda compleja, los metadatos de plataformas y el razonamiento

guiado por lógica natural [16], [51], [52], [53]. De esta familia el estudio deriva su política de agregación conservadora: varias fuentes relevantes pero no decisivas no se transforman en una conclusión fuerte, y el sistema registra requisitos cubiertos, faltantes y conflictos antes de comprometer una etiqueta.

2.9.8. Verificación imagen-texto y evaluación multimodal

El octavo grupo conecta la multimodalidad con el diseño de la evaluación. Los benchmarks imagen-texto y los estudios de factualidad de modelos generativos muestran que las métricas deben separar disponibilidad de evidencia, compatibilidad de etiqueta, *grounding* y atribución, para que la evaluación no se reduzca a una sola etiqueta cuando el contenido combina texto, imagen, procedencia y contexto temporal [54], [4], [26]. La literatura asociada vincula los benchmarks multimodales con la factualidad generativa y la recuperación aumentada, y permite tratar disponibilidad visual, soporte de cita y *grounding* como dimensiones distintas [55], [21], [56], [3]. El estudio adopta la idea de que una evaluación multimodal responsable requiere evidencia externa y justificación, de modo que la imagen se analiza junto con texto, fuente, contexto temporal y disponibilidad de señales, nunca como atajo decisivo.

2.9.9. Citas, alucinaciones y control de afirmaciones no soportadas

El noveno grupo trata el riesgo más específico de los modelos generativos: la afirmación plausible sin respaldo. La literatura sobre citas automáticas, alucinación y evaluación de respuestas exige comprobar si la cita sostiene exactamente lo que el sistema redacta, y fundamenta tanto el control de afirmaciones no soportadas como la decisión de tratar la respuesta final como objeto evaluable [14], [57], [58]. Estudios complementarios sobre detección de afirmaciones, *check-worthiness* y evaluación de citas en respuestas generadas refuerzan que una cita sólo cuenta cuando respalda la afirmación concreta redactada [59], [60], [61], [62]. La consecuencia para la respuesta pública es directa: el objetivo no es redactar una explicación más convincente, sino una cuyos enunciados puedan rastrearse a fuentes o marcarse como limitaciones.

2.9.10. Herramientas aplicadas, revisión humana y adopción responsable

El décimo grupo sitúa el sistema en su contexto de uso. Los trabajos aplicados sobre herramientas de verificación y revisión humana muestran que el valor de un sistema conversacional no se mide sólo por su grado de automatización, sino por si ayuda a periodistas, analistas o usuarios a reconstruir la ruta de evidencia, reconocer incertidumbre y decidir qué debe revisarse manualmente [63], [64], [65]. La literatura asociada incorpora intervención humana, NLI adversarial, conocimiento de sentido común y con-

tradicción en conocimiento abierto, y sostiene que el sistema debe asistir la revisión sin reemplazar el juicio humano ni pretender resolver todo el razonamiento semántico [66], [67], [68], [69]. La lectura final del mapa es, por tanto, una orientación de adopción: la salida debe ayudar a decidir qué revisar, qué creer con cautela y qué queda fuera del alcance de la evidencia recuperada.

2.10. Síntesis del marco teórico ampliado

El marco teórico ampliado sostiene cuatro decisiones centrales. Primero, la verificación automática debe modelarse como una cadena de operaciones observables. Segundo, la evidencia debe conservarse con metadatos suficientes para que una persona pueda inspeccionarla. Tercero, las señales multimodales y de procedencia deben tratarse como contexto verificable, no como prueba automática de verdad o falsedad. Cuarto, la respuesta generada debe evaluarse por su soporte, atribución y límites, no sólo por su fluidez.

Estas decisiones explican por qué el prototipo se concibe para evaluarse con dos familias de métricas. Las métricas de auditabilidad responden si el sistema produce un expediente revisable; las métricas clasificatorias responden si la etiqueta coincide con una referencia esperada. Ambas son necesarias, pero responden a preguntas distintas y no deben sustituirse entre sí: una mide la calidad del rastro de evidencia y la otra la compatibilidad con una etiqueta de referencia. Al permanecer separadas, un buen desempeño en una dimensión no se confunde con un resultado en la otra.

2.11. Discusiones transversales del marco teórico

Más allá del inventario bibliográfico, en cada línea de trabajo permanecen tensiones sin resolver. Los apartados siguientes discuten esos debates de fondo y la postura que el estudio adopta frente a ellos, pues son los que justifican una arquitectura de evidencia y no un clasificador adicional.

2.11.1. De la etiqueta al expediente de evidencia

Una tensión común en la literatura es la diferencia entre producir una etiqueta y producir un expediente de evidencia. Los datasets clásicos permiten entrenar o evaluar compatibilidad con etiquetas esperadas; el problema de esta investigación exige además conservar por qué una respuesta fue emitida y qué evidencia quedó fuera del alcance. FEVER, MultiFC y AVeriTeC ayudan a entender esa tensión porque combinan, en distinto grado, afirmaciones, evidencia, explicaciones o trazas de verificación [22], [24], [2]. La etiqueta es útil para medir desempeño, pero el expediente es necesario para

revisar la decisión. Por eso el marco teórico no trata accuracy, F1 y trazabilidad como métricas intercambiables.

El expediente de evidencia también cambia el papel del modelo de lenguaje. En un diseño centrado sólo en etiqueta, el modelo puede aparecer como clasificador o generador de respuesta. En un diseño auditable, el modelo queda rodeado por operaciones que producen y limitan su salida: extracción de afirmaciones, recuperación, selección de spans, comparación, registro de incertidumbre y *grounding*. Los trabajos de explicación, atribución y evaluación de factualidad muestran que una respuesta generada puede sonar plausible aunque no esté suficientemente respaldada por sus fuentes [1], [5], [14].

2.11.2. Recuperación aumentada y verificación de hechos

La recuperación aumentada por generación y los sistemas de búsqueda para fact-checking comparten un problema: traer más texto no garantiza traer mejor evidencia. Los trabajos de *retrieval*, RAG y selección de evidencia muestran que la calidad depende de formular consultas adecuadas, recuperar documentos pertinentes, seleccionar fragmentos comparables y reconocer cuándo la fuente no cubre todos los requisitos del *claim* [15], [16], [56]. El diseño adopta esa cautela al separar cobertura de evidencia, fuente decisiva y respuesta pública.

La recuperación también tiene una dimensión temporal y contextual. Muchas afirmaciones de desinformación dependen de fecha, lugar, entidad específica o causalidad. Una fuente puede ser correcta y aun así no resolver el caso si habla de otro momento, otro país o una relación causal distinta. Por ello, el marco teórico respalda almacenar requisitos cubiertos y faltantes. La ausencia de una fuente decisiva no debe convertirse automáticamente en falsedad; debe conservarse como insuficiencia o no verificabilidad cuando la evidencia no permite una conclusión fuerte.

2.11.3. Semántica, contradicción y límites de NLI

NLI y *stance* detection aportan una capa de comparación que el *retrieval* por sí solo no resuelve. La literatura muestra que dos textos pueden compartir vocabulario y sostener relaciones distintas: apoyo, contradicción, neutralidad, contexto parcial o ambigüedad. Los trabajos sobre inferencia, contradicción y evaluación de consistencia justifican registrar la comparación *claim*-evidence como objeto propio [19], [20], [21]. En el prototipo, esa comparación no se oculta detrás de la respuesta final; queda disponible para revisión.

Al mismo tiempo, el marco teórico advierte que NLI no debe convertirse en juez

final. Los modelos de inferencia pueden fallar por traducción, dominio, longitud del texto, ironía, negación compleja o condiciones temporales. Por esa razón, NLI se usa como señal auditada y no como verdad autónoma. La respuesta pública debe integrar NLI con calidad de fuente, suficiencia del *span*, requisitos cubiertos y limitaciones explícitas. Esta decisión explica por qué el sistema puede ser conservador aun cuando una señal semántica parezca favorable.

2.11.4. Multimodalidad, procedencia y confianza

La literatura multimodal aporta una advertencia adicional: verificar una imagen no equivale a verificar el *claim* que la acompaña. Benchmarks y estudios sobre imagen-texto muestran que las señales visuales son útiles, pero rara vez son suficientes por sí solas [10], [11], [4]. Los trabajos sobre procedencia y credenciales de contenido refuerzan la misma cautela: una señal técnica puede registrar contexto sin decidir la veracidad factual [12].

Esta distinción es central para el análisis de noticias falsas en entornos digitales. El sistema debe registrar OCR, metadatos, procedencia, marcas y disponibilidad de imagen, pero debe evitar convertir esas señales en veredicto automático. En el diseño, la imagen se integra al contrato como evidencia contextual; en la evaluación, la falta de evidencia visual completa se interpreta como límite metodológico; en la comunicación pública, la respuesta debe explicar si la imagen permite sostener algo, si sólo aporta contexto o si no alcanza para resolver el *claim*.

2.11.5. Herramientas asistidas y responsabilidad humana

Los trabajos sobre verificación asistida por IA y herramientas para periodistas coinciden en que la automatización debe apoyar, no sustituir, la revisión humana [8], [17], [63]. Esta perspectiva cambia el criterio de utilidad: un sistema puede ser valioso si reduce trabajo de búsqueda, ordena evidencia, expone conflictos y señala incertidumbre, incluso si no siempre mejora la etiqueta final. El prototipo se ubica en esa línea de apoyo revisable.

La responsabilidad humana también exige claridad documental. Si el sistema presenta una respuesta sin mostrar evidencia, el usuario queda obligado a confiar en la autoridad del modelo. Si presenta evidencia, limitaciones y rutas de verificación, el usuario puede cuestionar la respuesta y completar la revisión. Por ello, el marco teórico conecta literatura de herramientas, atribución y *grounding* con el diseño del prototipo. La meta no es producir una respuesta más contundente, sino una respuesta más inspeccionable.

2.11.6. Implicaciones para la contribución del estudio

Al integrar las líneas anteriores, el marco teórico justifica una contribución aplicada: un agente que organiza evidencia y límites bajo un contrato persistente. La literatura respalda cada componente, y el prototipo los reúne en una arquitectura local orientada a la auditabilidad: *claim* extraction, *retrieval*, comparación semántica, señales multimodales, *grounding* y evaluación de afirmaciones soportadas [1], [5], [14]. Esta integración explica por qué la contribución no se formula como liderazgo predictivo ni como reemplazo del verificador humano.

De esta integración se desprende también el criterio con que se valoran los resultados. Un desempeño alto en trazabilidad y un nivel bajo de afirmaciones no soportadas respaldan la contribución de auditabilidad; un desempeño clasificatorio por debajo de la línea base delimita el alcance predictivo y orienta el trabajo futuro. Ambas lecturas son compatibles porque evalúan dimensiones distintas. El valor académico del estudio reside en hacer explícita esa separación, implementarla en un prototipo y verificarla con artefactos reproducibles.

2.11.7. Cierre de brecha teórica

La brecha teórica que queda después de revisar estas fuentes no es la ausencia de modelos, datasets o métricas; es la falta de una integración que mantenga juntos el análisis conversacional, la evidencia recuperada, la comparación semántica, las señales multimodales y la explicación pública. Los trabajos revisados resuelven partes del problema: algunos se concentran en detección de afirmaciones, otros en datasets, otros en procedencia o atribución [7], [5], [12]. Esa dispersión justifica una arquitectura que no reemplaza esas líneas, sino que las coordina bajo un registro común.

La amplitud del marco teórico responde a esa naturaleza compuesta del problema. Un enfoque sólo textual dejaría fuera imagen, procedencia y metadatos; uno centrado sólo en modelos generativos dejaría fuera selección de evidencia y *grounding*; uno limitado a clasificación dejaría sin explicar al usuario qué se revisó y qué permaneció abierto. Analizar desinformación con un agente auditable exige sostener simultáneamente decisiones de recuperación, comparación, atribución, comunicación y evaluación. Esa base permite leer el prototipo como una infraestructura de revisión documentada, no como un árbitro automático de verdad, y delimita sus responsabilidades técnicas y comunicativas.

2.12. Lectura visual de la literatura incorporada

Las figuras incorporadas sintetizan antecedentes visuales relevantes para el marco teórico. La primera, tomada de Kotonya y Toni, representa el *pipeline* de fact-checking automático con una subtarea explícita de explicación. Esa organización confirma que la explicación no aparece después de la predicción como añadido retórico, sino como parte de la cadena de verificación. Es la misma exigencia del análisis auditable: recuperación, comparación y comunicación permanecen encadenadas y no se tratan como pasos independientes.

La figura de AVeriTeC aporta otro matiz. En lugar de mostrar sólo una etiqueta, organiza *claim*, metadatos, preguntas de evidencia, respuestas y veredicto. Esta estructura ayuda a entender por qué la salida pública no se evalúa como una frase aislada. El *claim* necesita descomponerse en requisitos verificables; la evidencia debe responder preguntas concretas; el veredicto debe depender de esa relación y no de similitud superficial. El contrato de evidencia del prototipo cumple una función análoga: hace que cada respuesta pueda inspeccionarse por piezas.

La figura de ALCE conecta el problema con generación de respuestas citadas. Su valor visual está en mostrar que una respuesta generada puede incluir pasajes recuperados y citas explícitas, pero aun así debe evaluarse si esas citas sostienen cada afirmación. Esa precaución se adopta como criterio de *grounding*: una respuesta con enlaces o referencias sólo es útil si cada enunciado sustantivo está apoyado por la fuente citada. Por ello, la evaluación conserva *unsupported-statement rate* como métrica primaria de auditabilidad.

La figura de AVerImaTeC introduce la dimensión imagen-texto de forma más rica que un diagrama propio. Presenta una afirmación visual, preguntas relacionadas con la imagen, evidencia auxiliar y una justificación que explica por qué el contenido queda refutado. Su aporte al marco teórico es mostrar que la multimodalidad no se reduce a clasificar una imagen: exige contexto, temporalidad, identificación de personas, búsqueda de imágenes relacionadas y razonamiento sobre la correspondencia entre texto e imagen. De ahí que el prototipo registre las señales visuales como elementos separados y no como veredicto automático.

La figura de FActScore aporta la perspectiva de la evaluación de factualidad. Al descomponer una respuesta en afirmaciones atómicas y calcular cuántas están soportadas, muestra por qué una evaluación binaria puede ocultar diferencias importantes. Dos respuestas pueden fallar como respuesta perfecta, pero una puede contener más piezas verificables que otra. El estudio aplica una idea cercana al evaluar respuesta pública,

grounding y afirmaciones no soportadas: la calidad de una salida conversacional se mide por el soporte de sus partes, no sólo por la impresión global que produce.

En conjunto, estas figuras muestran cinco formas en que la literatura ha representado problemas cercanos: *pipeline* explicable, evidencia descompuesta, generación con citas, razonamiento imagen-texto y evaluación atómica de factualidad. Esa lectura visual complementa el análisis textual y delimita la integración propuesta: un agente conversacional que conserva una ruta de evidencia revisable desde la entrada hasta la respuesta fundamentada.

3. Metodología

Este capítulo describe cómo se produjo la evidencia de la tesis y cuál es su criterio de interpretación. La metodología no se reduce a la construcción del prototipo: define el camino que conecta problema, literatura, hipótesis, diseño experimental, implementación y evaluación. Esa conexión es necesaria porque el objetivo del trabajo no es emitir una etiqueta aislada sobre noticias falsas, sino demostrar si un agente puede conservar un expediente revisable. Las figuras iniciales ubican esa lógica. La Figura 3.1 muestra cómo AVeriTeC organiza una afirmación mediante preguntas, evidencias y veredicto; por eso se usa aquí como referencia para explicar que el estudio evalúa expedientes de evidencia, no frases sueltas. La Figura 3.2 resume después las etapas propias del proyecto y el tipo de decisión que aporta cada una.

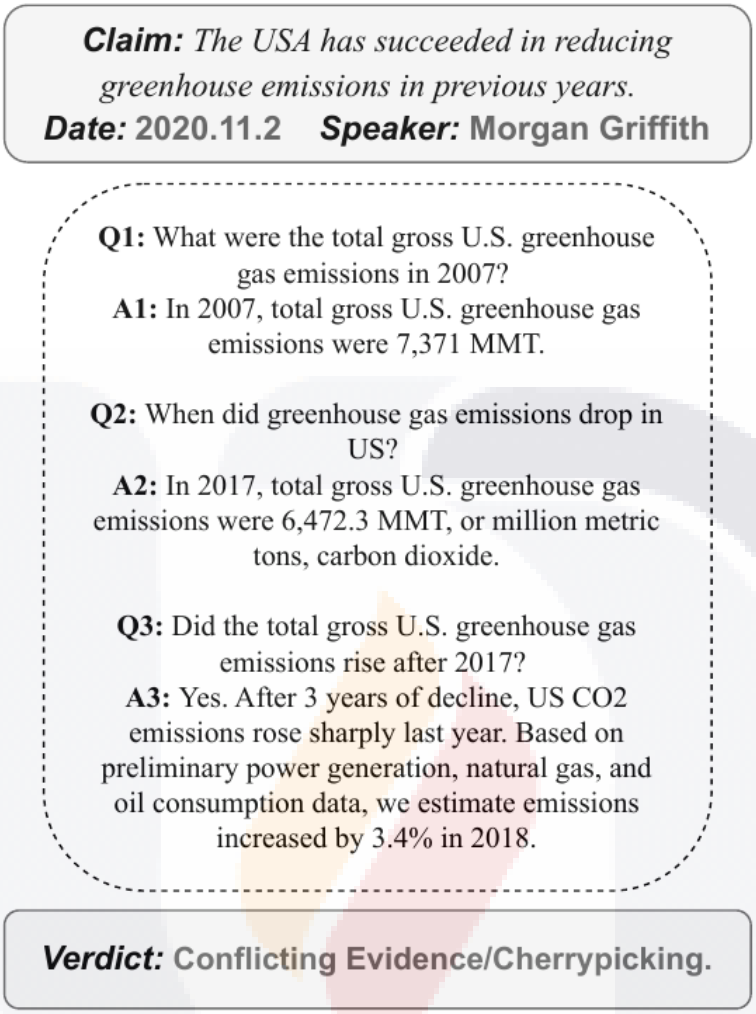

Figura 3.1: Instancia AVeriTeC con afirmación, preguntas de evidencia y veredicto. [2].
Fuente: The Automated Verification of Textual Claims (AVeriTeC) Shared Task, figura 1, p. 1.

Figura 3.2: Mapa temporal de etapas metodológicas, decisiones y artefactos principales.
Fuente: elaboración propia.

3.1. Enfoque general

La investigación se organizó como un proceso incremental con etapas verificables. El propósito de esa organización fue mantener una relación explícita entre problema, literatura, hipótesis, diseño, implementación, evaluación y conclusiones. Cada etapa debía producir evidencia suficiente para justificar la siguiente, sin confundir la planeación metodológica con el sistema evaluado.

El flujo local se dividió en etapas: la etapa de alcance definió el problema y sus restricciones; la revisión bibliográfica construyó la base conceptual; la síntesis conceptual convirtió la literatura en hipótesis; el diseño experimental fijó el contrato de evidencia; la implementación materializó y validó el prototipo; la evaluación local analizó si los artefactos permitían avanzar; la evaluación pública extendida amplió la evidencia empírica; y el cierre metodológico integró revisión y refinamiento.

3.2. Etapa de alcance y problema

La etapa de alcance transformó el tema general de noticias falsas en un objetivo evaluable. El resultado fue una investigación centrada en diseño y desarrollo de un agente de análisis auditable, no en entrenamiento de un modelo ni en superación de benchmarks. Esta etapa fijó restricciones: no video, no forense avanzado de deepfakes, no conjuntos de datos propios, no estudios con usuarios y no afirmación de liderazgo en benchmarks.

El árbol de problemas organizó dependencias: primero el contrato de evidencia, después extracción de afirmaciones, recuperación de evidencia, credibilidad, *stance*/NLI,

imágenes, orquestación, experiencia de usuario, controles anti-alucinación y evaluación contra comparador. Esta secuencia se mantuvo en etapas posteriores.

3.3. Revisión bibliográfica

La revisión bibliográfica recopiló y cribó literatura sobre verificación de hechos, recuperación de evidencia, NLI, multimodalidad, atribución y evaluación de factualidad. Las tarjetas de lectura y el cribado distinguieron fuentes centrales, fuentes con advertencia y fuentes diferidas. El resultado no fue una revisión exhaustiva de todo el campo, sino una base suficiente para justificar decisiones de diseño y métricas.

3.4. Síntesis conceptual e hipótesis

La síntesis conceptual integró la literatura en decisiones de diseño. La síntesis definió el agente como sistema de orquestación de evidencia. Las hipótesis H1-H5 se enfocan en trazabilidad, cobertura de evidencia, *stance*/NLI, señales multimodales no decisivas y métricas de auditabilidad. La síntesis también dejó visibles dos riesgos metodológicos: la búsqueda inversa de imágenes y la evaluación multimodal dependen de señales incompletas y de disponibilidad externa.

3.5. Diseño experimental

El diseño experimental convirtió las hipótesis en un plan de comparación medible y operacionalizó el cuarto objetivo específico: la comparación entre respuestas con y sin rastro explícito de evidencia se materializó en dos condiciones. La condición A fue el agente estructurado con contrato de evidencia, es decir, la respuesta con rastro explícito. La condición B fue una línea base no estructurada: un modelo de lenguaje abierto que recibía la misma entrada y una búsqueda web simple, pero sin acceso al registro estructurado del agente, es decir, la respuesta sin rastro explícito. Como modelo comparador se empleó Gemma 4 E4B, la variante de cuatro mil millones de parámetros efectivos de la familia de modelos abiertos Gemma de Google, ejecutada de forma local mediante Ollama, un entorno que descarga y corre modelos de lenguaje cuantizados en el equipo del investigador. Las métricas primarias fueron traceability rate, cobertura de evidencia, calidad de atribución, consistencia *claim*-evidence, unsupported-statement rate y manejo de incertidumbre. Las métricas secundarias fueron compatibilidad de etiqueta, latencia y comportamiento en modo degradado.

Para evitar interpretaciones ambiguas, cada métrica se definió de forma operativa antes de la evaluación. La Tabla 3.1 reúne la definición, el rango y el criterio de interpretación de las métricas primarias de auditabilidad y de las secundarias de

clasificación.

Tabla 3.1: Operacionalización de las métricas de evaluación

Fuente: elaboración propia.

Métrica	Definición operativa	Rango	Interpretación
Primarias (auditabilidad)			
Retrieval hit rate	Proporción de casos con al menos una evidencia relevante recuperada.	[0,1]	Mayor es mejor.
Decisive-source rate	Proporción de casos con una fuente decisiva para el *claim*.	[0,1]	Mayor es mejor.
Traceability rate	Proporción de afirmaciones de la respuesta con *grounding* registrado.	[0,1]	Mayor es mejor.
Unsupported-statement rate	Proporción de afirmaciones de la respuesta sin soporte.	[0,1]	Menor es mejor.
Secundarias (clasificación)			
Compatibilidad de etiqueta (accuracy)	Proporción de veredictos que coinciden con la etiqueta esperada.	[0,1]	Mayor es mejor.
Macro-F1	Media no ponderada de F1 por clase.	[0,1]	Mayor es mejor; sensible a clases minoritarias.
Weighted-F1	F1 por clase ponderado por su soporte.	[0,1]	Mayor es mejor.

La separación entre familias se reflejó también en la relación con los objetivos específicos. La Tabla 3.2 hace explícita esa correspondencia para evitar que el estudio se interprete como una propuesta de clasificación tradicional: la variable principal de éxito es la auditabilidad, y la clasificación es un diagnóstico secundario.

Tabla 3.2: Correspondencia entre objetivos específicos y criterios de evaluación

Fuente: elaboración propia.

Objetivo específico	Criterio o métrica principal
OE1. Analizar el problema de revisión de respuestas conversacionales.	Diagnóstico conceptual y diseño del contrato de evidencia.
OE2. Definir los elementos mínimos de un análisis revisable.	Completitud del contrato: *claim*, fuentes, evidencia, comparación y límites.
OE3. Diseñar y desarrollar el prototipo trazable.	Validación de contrato y traceability rate.
OE4. Evaluar respuestas con y sin rastro explícito de evidencia.	Comparación agente frente al comparador en métricas primarias y secundarias.
OE5. Identificar límites técnicos y metodológicos.	Diagnósticos de error y deltas con intervalos de confianza.

Esta etapa también definió el esquema del contrato de evidencia. El contrato funcionó como eje técnico de la investigación: todo módulo debía escribir en él y toda respuesta final debía poder validarse contra él.

3.6. Implementación del prototipo

La etapa de implementación materializó el prototipo. El modo controlado usó casos deterministas para validar el contrato sin depender de APIs externas. El modo abierto usó integraciones reales cuando estuvieron configuradas: búsqueda, Google Fact Check, extracción de URLs, señales de imagen, NLI local y respaldo de modelo de lenguaje. La salida se persistió como contrato de evidencia, manifiesto de ejecución, reporte de validación y registro de refinamiento.

3.7. Evaluación local y evaluación pública extendida

La evaluación local analizó artefactos guardados y concluyó que el prototipo podía avanzar a evaluación comparativa: los contratos fueron válidos y el comparador fue ejecutable.

La evaluación pública extendida analizó 60 muestras de familias públicas y comparó el agente contra un comparador basado en Gemma 4 E4B ejecutado mediante Ollama. Esta evaluación no reemplaza el criterio de avance local; lo amplía con evidencia empírica más rica.

El tamaño de la muestra se fijó en 60 casos distribuidos de forma balanceada

en 15 muestras por cada una de las cuatro familias de datos (AVeriTeC, FEVER Gold, MultiFC y AVerImaTeC), de modo que cada familia tuviera igual peso y ninguna dominara el promedio. Esta decisión responde a una restricción práctica: cada muestra exige ejecutar el agente completo y el comparador con búsqueda y extracción en vivo, lo que limita el volumen factible bajo recursos locales. El conjunto se concibe, por tanto, como una sonda exploratoria con cobertura diversa y no como una muestra dimensionada para inferencia poblacional. Para acotar la incertidumbre derivada de este tamaño, las diferencias frente al comparador se acompañan de intervalos de confianza al 95 % obtenidos por *bootstrap*, reportados junto con cada métrica de clasificación.

3.8. Criterios de validez

La metodología separó evidencia de diseño, evidencia de ejecución y evidencia de evaluación. Una afirmación de diseño se sostuvo con documentos de etapa; una afirmación de implementación se sostuvo con código y artefactos de ejecución; una afirmación de desempeño se sostuvo con métricas y reportes de evaluación persistidos. Cuando un número no estuvo respaldado por el conjunto de artefactos revisado, se marcó como límite y no se usó como resultado final.

3.9. Desarrollo metodológico por etapas

La metodología se presenta por etapas para mostrar qué evidencia produce cada decisión y cómo condiciona la siguiente.

3.9.1. Modelo metodológico de referencia

El estudio usa un modelo de trabajo por etapas para dar orden metodológico al proceso. La idea central es que cada decisión importante debe dejar evidencia revisable antes de avanzar a la siguiente. Esto evita que el prototipo aparezca como una solución construida de manera aislada: primero se delimita qué problema puede atenderse, después se justifica con literatura, luego se formulan hipótesis, se diseña una forma de medirlas, se implementa el sistema y finalmente se interpretan los resultados.

La referencia metodológica no sustituye el criterio de investigación ni funciona como una herramienta externa que produzca conclusiones. Su papel es más limitado: sirve para ordenar preguntas, productos intermedios y criterios de avance. En este estudio, lo relevante no es la ruta técnica de los archivos, sino la relación entre cada etapa y la afirmación que permite sostener. El alcance define límites; la revisión bibliográfica aporta conceptos; la síntesis formula hipótesis; el diseño experimental fija métricas; la implementación demuestra viabilidad técnica; y la evaluación muestra resultados y

restricciones.

3.9.2. Criterios de avance

Un criterio de avance funciona como control de honestidad. La etapa de alcance no termina porque exista una idea general, sino cuando el problema queda delimitado. La revisión bibliográfica no termina por acumular referencias, sino cuando cubre las decisiones de diseño que el prototipo necesita tomar. La implementación no termina por producir una respuesta fluida, sino cuando conserva contratos válidos, manifiestos, reportes de validación y limitaciones visibles.

Este enfoque evita sobreafirmaciones más allá de lo que los artefactos permiten. Si una afirmación depende de literatura, debe tener una referencia clara. Si depende de ejecución, debe tener una ejecución documentada o reporte. Si depende de desempeño, debe tener una métrica documentada. Cuando alguno de esos apoyos falta, la afirmación se debilita, se marca como límite o se elimina.

3.9.3. Evaluación local y evaluación extendida

La evaluación local y la evaluación pública extendida cumplen funciones distintas. La primera comprueba que el prototipo produce objetos válidos: contratos, trazabilidad, *grounding* y comparación con un comparador inicial. La segunda introduce 60 muestras públicas y un comparador basado en Gemma 4 E4B ejecutado mediante Ollama para observar comportamiento en casos más variados. Mezclar ambas lecturas produciría una conclusión confusa: la evaluación local prueba estructura y la extendida tensiona desempeño.

La consecuencia metodológica es directa: la evaluación extendida amplía la evidencia empírica sin convertir al sistema en un clasificador general ni sustituir los criterios locales de validez. Sus resultados se reportan en dos registros complementarios, como evidencia de auditabilidad y como diagnóstico de los límites predictivos del prototipo.

3.9.4. Criterios de evidencia

Cada afirmación sustantiva de la investigación se ubica en uno de tres planos. El plano conceptual proviene de literatura sobre fact-checking, *retrieval*, NLI, multimodalidad, procedencia y *grounding*. El plano técnico proviene del contrato, el código y los artefactos persistidos. El plano experimental proviene de reportes de evaluación y métricas. Esta separación impide que una cita teórica se use como si fuera resultado empírico, o que una ejecución local se use como si demostrara generalización amplia.

También aclara el papel del comparador. No se usa para declarar una victoria

total del agente; se usa para contextualizar costos y beneficios de introducir un contrato estructurado. Si el comparador supera al agente en F1, esa información se reporta. Si el agente conserva trazabilidad y afirmaciones soportadas, esa información también se reporta. La contribución depende de leer ambos planos sin esconder ninguno.

3.9.5. Trazabilidad de afirmaciones

La trazabilidad también guía la forma de presentar los resultados. Cada afirmación sustantiva conserva relación con una fuente de evidencia: el planteamiento del problema con el alcance, el marco teórico con la revisión bibliográfica, las hipótesis con la síntesis conceptual, el contrato de evidencia con el diseño experimental, el prototipo con la implementación y los resultados con las evaluaciones realizadas. Esta relación permite revisar de dónde surge una afirmación y qué tan fuerte es su soporte.

Este punto es importante porque los resultados técnicos pueden exagerarse al convertirse en argumento. Una evaluación que reporta unsupported-statement rate 0.0 sólo respalda esa condición bajo las muestras evaluadas; no garantiza ausencia universal de alucinaciones. Del mismo modo, si el comparador obtiene mejor macro-F1, ese resultado se reporta como límite del enfoque y como criterio para trabajo futuro. La interpretación conserva la diferencia entre contribución de auditabilidad y desempeño clasificatorio.

3.9.6. Decisiones de alcance que permanecen visibles

El estudio conserva varias decisiones de alcance porque ayudan a leer los resultados. No se entrena un modelo nuevo; se evalúa una arquitectura de orquestación y evidencia. No se crea un *dataset* propio; se usan muestras públicas para sostener comparabilidad. No se afirma resolución completa de desinformación multimodal; se registran señales visuales disponibles y sus límites. No se presenta el prototipo como verificador autónomo; se presenta como infraestructura de revisión.

Estas decisiones reducen la promesa inicial, pero hacen más claro el alcance del trabajo. Permiten distinguir contribución técnica, contribución metodológica y trabajo futuro. La contribución técnica está en el contrato y el prototipo. La contribución metodológica está en separar evidencia, comparación, limitación y respuesta. El trabajo futuro queda en mejorar clasificación, cobertura visual, calibración semántica, latencia y validación con usuarios.

3.10. Operacionalización de hipótesis

Las hipótesis de trabajo, presentadas en el Capítulo 1, se derivan de la relación entre el problema de investigación, la literatura revisada y el alcance del estudio. Aquí se operacionalizan para el diseño experimental: cada hipótesis se traduce en señales observables, métricas y condiciones de comparación.

- H1. Evidencia estructurada y trazabilidad. Un agente que registra afirmaciones normalizadas, evidencias, metadatos de fuente, señales *stance*/NLI y limitaciones en un contrato estructurado debe producir respuestas más trazables que un comparador basado en Gemma 4 E4B ejecutado mediante Ollama con búsqueda simple. Las señales de evaluación son traceability rate, unsupported-statement rate, atribución de fuentes y presencia de limitaciones explícitas [1], [5], [14].

- H2. Recuperación separada y cobertura de evidencia. Tratar la recuperación de fact-checks y web abierta como etapa auditable debe mejorar cobertura y relevancia frente a una búsqueda integrada sólo en la instrucción del comparador. Las señales de evaluación son *retrieval* hit rate, número de evidencias relevantes, fuente decisiva y manejo explícito de evidencia insuficiente [15], [17], [2].

- H3. *Stance*, NLI y etiquetas sobreconfiadas. Comparar afirmación y evidencia antes de agregar el resultado debe reducir etiquetas verdaderas o falsas no soportadas. Las señales de evaluación son consistencia *claim*-evidence, registros NLI, diagnósticos de desajuste y conservación de incertidumbre cuando la evidencia es neutral o débil [19], [20], [21].

- H4. Señales visuales no decisivas y explicación multimodal. En entradas de imagen o imagen-texto, registrar OCR, contexto visual, metadatos, procedencia y fallos de disponibilidad debe mejorar la explicación sin convertir esas señales en prueba automática de autenticidad o falsedad. Las señales de evaluación son cobertura de campos visuales, distinción entre procedencia y veredicto, y respuestas que evitan sobreinterpretar imagen o metadatos [12], [4], [13].

- H5. Métricas de auditabilidad frente a accuracy aislada. Una evaluación mixta centrada en trazabilidad, soporte de evidencia, atribución e incertidumbre debe describir la aportación del prototipo mejor que la compatibilidad de etiqueta por sí sola. Las señales de evaluación son cobertura de evidencia, citation *grounding*, consistencia *claim*-evidence, compatibilidad de etiqueta, latencia y modo degradado [22], [2], [3].

La Tabla 3.3 resume esta operacionalización y vincula cada hipótesis con la variable observable que la representa y con la métrica o señal que permite contrastarla.

Tabla 3.3: Operacionalización de las hipótesis de trabajo

Fuente: elaboración propia.

H	Variable observable	Métrica o señal de validación
H1	Trazabilidad y soporte de la respuesta	Traceability rate, unsupported-statement rate, atribución y presencia de limitaciones.
H2	Cobertura y suficiencia de evidencia	*Retrieval* hit rate, decisive-source rate y manejo explícito de evidencia insuficiente.
H3	Consistencia *claim*-evidence	Señales *stance*/NLI registradas y diagnósticos de desajuste (`nli_stance_mismatch`).
H4	Señales visuales no decisivas	Cobertura de campos visuales y separación entre procedencia y veredicto.
H5	Auditabilidad frente a clasificación aislada	Familia de métricas primarias contrastada con accuracy y F1.

Estas hipótesis establecen el criterio de lectura de los capítulos posteriores. Si una métrica clasificatoria mejora pero el expediente pierde soporte o trazabilidad, el resultado no satisface el centro del estudio. Si la trazabilidad mejora pero la clasificación queda por debajo del comparador, la contribución queda respaldada sólo en el plano de auditabilidad y debe reportarse con ese límite.

3.11. Trazabilidad metodológica del estudio

La metodología se interpreta como una cadena de decisiones verificables y no como una secuencia de tareas administrativas. Cada etapa produce un artefacto que condiciona la siguiente: el alcance define qué problema se acepta resolver, la revisión bibliográfica delimita qué técnicas se consideran pertinentes, la síntesis convierte esa literatura en hipótesis evaluables, el diseño experimental fija el contrato de evidencia, la implementación materializa el contrato y la evaluación decide qué afirmaciones pueden sostenerse. Esta organización permite sostener las conclusiones con evidencia acumulativa: una conclusión no aparece aislada, sino conectada con una decisión previa y con

un resultado posterior.

La primera consecuencia de esta lectura es que el estudio no descansa en una promesa general sobre inteligencia artificial. El objeto empírico es un prototipo local que analiza afirmaciones mediante evidencia registrable. Por ello, la metodología evita formular el sistema como detector universal de noticias falsas. La pregunta metodológica central es más acotada: si un agente conversacional puede convertir una entrada potencialmente desinformativa en un expediente que una persona pueda revisar. El expediente incluye *claim* normalizado, fuentes recuperadas, fragmentos, comparación semántica, señales visuales, limitaciones y respuesta pública. Esta estructura desplaza la atención desde una etiqueta aislada hacia la inspeccionabilidad del proceso.

La segunda consecuencia es operativa: cada afirmación sustantiva debe ubicarse en el plano que la sostiene —conceptual, técnico o experimental— antes de darse por válida. Así, sostener que la procedencia visual no equivale a veracidad se apoya en literatura y diseño; sostener que el sistema guardó contratos válidos se apoya en las validaciones locales; y sostener que no hubo superioridad clasificatoria global se apoya en la comparación de métricas. Esta disciplina impide que una cita teórica se presente como hallazgo empírico o que una ejecución controlada se lea como prueba de generalización.

La tercera consecuencia es el uso de criterios de avance. Una etapa se considera suficiente cuando produce artefactos capaces de sostener el siguiente tipo de afirmación. La revisión bibliográfica es suficiente si cubre las decisiones de diseño que el prototipo necesita tomar. El diseño experimental es suficiente si define una estructura que pueda validarse y medirse. La implementación es suficiente si produce salidas persistidas y contratos válidos. La evaluación es suficiente si reporta tanto fortalezas como límites. Esta lógica evita avanzar por inercia textual: si una afirmación carece de fuente, artefacto o métrica, se debilita o se elimina.

La metodología también conserva una distinción crítica entre evaluación local y evaluación pública extendida. La evaluación local prueba que el contrato, la persistencia y las validaciones funcionan en condiciones controladas. La evaluación pública extendida introduce 60 muestras y un comparador estricto basado en Gemma 4 E4B para observar comportamiento comparativo. Estas dos evaluaciones no responden exactamente la misma pregunta. La primera responde si el prototipo produce objetos válidos y trazables. La segunda responde cómo se comporta ese patrón frente a casos más variados y frente a una alternativa no estructurada. Mezclarlas generaría una lectura equivocada del alcance.

Finalmente, la metodología define cómo se interpreta la reproducibilidad. El estu-

dio no asume que la web, los resultados de búsqueda o las APIs externas permanecerán idénticas. La reproducibilidad se entiende como capacidad de auditar qué se ejecutó, qué se registró y qué números se reportaron. Esto exige conservar artefactos, rutas, métricas y referencias suficientes para reconstruir el argumento. Bajo este criterio, una conclusión es verificable cuando su rastro de evidencia puede revisarse.

4. Diseño del agente

La Figura 4.1 introduce la relación entre pasajes recuperados, generación y citas, una precaución central para el diseño de grounding de la respuesta pública.

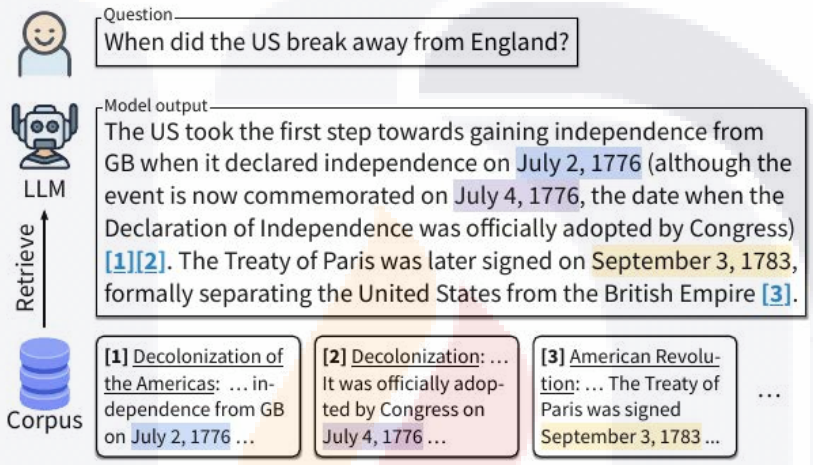

Figura 4.1: Configuración ALCE para generar respuestas con pasajes citados. [3].
Fuente: Enabling Large Language Models to Generate Text with Citations, figura 1, p. 1.

La Figura 4.2 resume en capas cómo el contrato de evidencia limita la respuesta y evita que una sola señal se convierta en veredicto.

Figura 4.2: Diagrama por capas del contrato de evidencia y componentes del agente.
Fuente: elaboración propia.

La Figura 4.3 conserva el detalle de objetos persistidos que sostienen la auditoría.

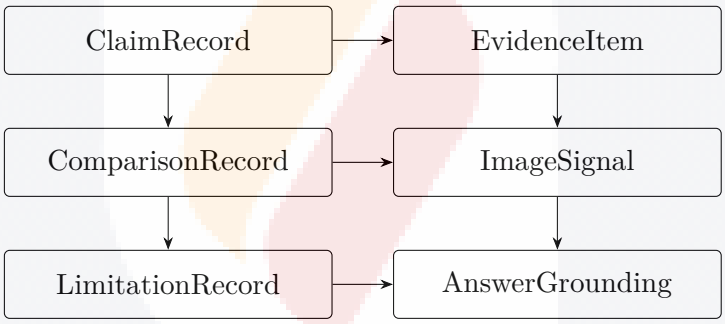

Figura 4.3: Estructura propia del contrato de evidencia.
Fuente: elaboración propia.

4.1. Principio de auditabilidad

El diseño se guía por un principio: la respuesta final debe poder auditarse. Esto significa que el usuario o evaluador debe poder reconstruir qué afirmación se analizó, qué evidencia se recuperó, qué relación se estableció entre afirmación y evidencia, qué fuentes se citaron, qué limitaciones afectaron el resultado y qué afirmaciones de la respuesta pública están respaldadas.

Este principio evita dos errores. El primero es convertir el modelo de lenguaje en juez autónomo de verdad. El segundo es esconder la incertidumbre detrás de una etiqueta aparentemente definitiva. El agente puede emitir una etiqueta cautelosa, pero esa etiqueta debe derivarse de evidencia estructurada.

4.2. Arquitectura conceptual

La arquitectura separa módulos:

1. Entrada y análisis inicial.
2. Extracción y normalización de afirmaciones.
3. Planeación de evidencia.
4. Recuperación de fact-checks y web abierta.
5. Extracción de contenido de URLs.
6. Señales de imagen.
7. Selección de spans y NLI.
8. Agregación conservadora.
9. Redacción de respuesta pública.
10. Validación de *grounding*.

La separación no es sólo organizacional. Permite que cada decisión deje artefactos y que las fallas se registren como limitaciones. Si una herramienta no está disponible, el contrato no desaparece; se degrada explícitamente.

4.3. Contrato de evidencia

El contrato de evidencia definido en diseño experimental organiza la información en registros. `ClaimRecord` preserva la afirmación original y normalizada. `EvidenceRecord` describe fuente, URL, método de recuperación, texto, tipo de evidencia y limitaciones. `ComparisonRecord` une *claim* y evidencia con *stance*, NLI, confianza y requisitos cubiertos. `ImageSignalRecord` separa OCR, descripción, metadatos, procedencia y marcas. `LimitationRecord` registra fallas y degradaciones. `FinalAnswer` contiene veredicto, citas, soporte en evidencia y reporte público.

La ventaja de este contrato es que convierte una respuesta conversacional en un objeto verificable. La respuesta ya no es sólo texto; es una vista pública de una estructura interna más completa.

4.4. Agregación conservadora

La agregación es conservadora por diseño. Una fuente contextual no debe volverse decisiva sólo por ser relevante. Una evidencia que menciona un tema no necesariamente apoya la afirmación. Un fact-check puede cubrir una afirmación parecida pero no idéntica. Por eso el sistema registra requisitos cubiertos y faltantes, y mantiene labels como `insufficient_evidence` o `not_verifiable` cuando la evidencia no alcanza.

La ausencia de evidencia no se interpreta como falsedad. Esta regla es epistémica y ética: en desinformación, no encontrar evidencia suficiente puede deberse a limitaciones de búsqueda, acceso, idioma, fecha o cobertura, no necesariamente a que la afirmación sea falsa.

4.5. Señales de imagen

El diseño multimodal separa señales visuales de decisiones de verdad. OCR, etiquetas, contexto de búsqueda inversa, metadatos, C2PA y marcas se registran cuando están disponibles, pero no prueban por sí mismos la veracidad de una afirmación. Esta política evita sobreinterpretar indicios técnicos. Una fotografía real puede emplearse fuera de su contexto original; una imagen generada puede acompañar una afirmación que se analiza con fuentes textuales.

4.6. Respuesta pública

La respuesta pública se redacta en un formato compacto: veredicto, confianza, explicación breve, evidencia, fuentes y limitaciones. La decisión de mantener una respuesta legible no elimina el contrato interno. El usuario ve una síntesis; el evaluador puede revisar el contrato completo.

4.7. Validación

El diseño incluye validación de esquema y *grounding*. La validación revisa que los contratos cumplan el esquema y que la respuesta pública no introduzca afirmaciones sustantivas sin respaldo. Este control es una de las aportaciones centrales: pedirle al modelo de lenguaje que sea cuidadoso no es suficiente; el sistema debe verificar la estructura que produce.

Con este diseño, el contrato de evidencia funciona como puente entre la intención metodológica y la implementación. El siguiente capítulo muestra cómo esas decisiones se materializaron en módulos, modos de ejecución, artefactos persistidos y validaciones que hacen posible evaluar si la respuesta final permanece ligada a evidencia revisable.

4.8. Detalle del contrato y flujo de decisión

Las secciones anteriores describieron los componentes del agente por separado. Esta los recorre en el orden en que actúan sobre una entrada concreta, para mostrar cómo el contrato de evidencia encadena las decisiones que después se implementan y evalúan. El hilo conductor es el principio de auditabilidad: cada salida pública debe poder rastrearse hacia las afirmaciones, la evidencia y las limitaciones que la sostienen, de modo que el esquema del contrato define objetos persistentes para afirmaciones, evidencias, comparaciones, señales y *grounding* [1], [5], [14]. El contrato opera, en este recorrido, como la frontera entre las operaciones técnicas y las afirmaciones que el sistema puede comunicar como justificadas [3].

4.8.1. Del claim al plan de verificación

El flujo comienza antes de recuperar nada. Un *claim* complejo se descompone en preguntas de verificación, requisitos factuales y criterios de suficiencia, porque sin esa descomposición la recuperación carecería de objetivo. El objeto `claim_plan` registra esas preguntas, las fuentes preferidas y los campos que faltan por cubrir, y fija de antemano qué tendría que aportar una fuente para considerarse decisiva [7], [16], [26]. Esta planificación temprana es lo que permite, más adelante, distinguir entre una búsqueda que devolvió texto relacionado y una que efectivamente resolvió los requisitos del caso.

4.8.2. De la recuperación a la comparación

Con el plan fijado, el agente recupera evidencia combinando fact-checks publicados, búsqueda abierta, extracción de URLs y conjuntos controlados según el contexto; cada elemento se inscribe en el contrato con su método de recuperación, fuente, fragmento, calidad y limitaciones [17], [18], [2]. La etapa siguiente es la que impide confundir mención con soporte: la comparación explícita entre afirmación y evidencia, que registra el *stance* efectivo, la cobertura de requisitos y la razón por la que un fragmento resulta suficiente o insuficiente [19], [20], [21]. Sólo lo que sobrevive a esta comparación entra como soporte real al paso de agregación.

4.8.3. De la agregación a la respuesta con grounding

La agregación cierra el flujo del lado de la decisión. Prioriza no sobreafirmar cuando la evidencia es parcial, conflictiva o contextual, y deja registro de esa cautela mediante diagnósticos como `overly_conservative_aggregation`, que funciona a la vez como límite reconocido y como agenda técnica [25], [26], [23]. La respuesta pública es entonces una síntesis controlada por *grounding*, no una nueva ronda libre de razona-

miento del modelo: el validador comprueba que cada enunciado sustantivo tenga su registro de *grounding* y marca como afirmación no soportada cualquier conclusión que el expediente no respalde [5], [3], [14]. Al final del recorrido, la salida comunicada es el último eslabón de una cadena cuyos pasos quedaron todos inscritos en el contrato.

4.9. Decisiones epistémicas del contrato

El diseño del agente se organiza alrededor de una política epistémica conservadora. El sistema no transforma automáticamente evidencia relacionada en soporte fuerte, ni convierte ausencia de evidencia en falsedad. Esta decisión es central porque la desinformación digital suele combinar hechos verdaderos, contexto incompleto, imágenes reales fuera de lugar y afirmaciones causalmente ambiguas. Una arquitectura que sólo busque una etiqueta final puede ocultar esos matices; una arquitectura orientada a la auditabilidad los conserva como parte de la salida.

El contrato de evidencia separa niveles que con frecuencia se mezclan en respuestas generadas. El primer nivel es la entrada original: texto, enlace o imagen. El segundo nivel es el *claim* normalizado, que conserva entidad, tiempo, lugar, magnitud y relación factual cuando están presentes. El tercer nivel es la evidencia recuperada, con fuente, URL, fragmento y método de obtención. El cuarto nivel es la comparación entre *claim* y evidencia. El quinto nivel registra señales visuales y de procedencia. El sexto nivel contiene limitaciones y respuesta pública. Separar estos niveles permite ubicar errores: extracción incorrecta, evidencia débil, fuente no decisiva, comparación semántica dudosa o redacción demasiado fuerte.

La agregación conservadora se deriva de esa separación. Varias fuentes pertinentes no bastan si ninguna cubre el requisito central del *claim*. Una fuente confiable no es decisiva si habla de otra fecha, otra entidad o un contexto distinto. Un resultado visual no resuelve por sí mismo una afirmación textual. Por ello, la respuesta pública se formula desde suficiencia y no desde simple acumulación. El sistema puede decir que hay evidencia relacionada, que hay señales contextuales o que falta una fuente decisiva. Esa cautela no es una debilidad accidental, sino una condición de auditabilidad.

El papel del modelo de lenguaje también queda restringido. El modelo de lenguaje puede ayudar a extraer, normalizar, resumir y redactar, pero no se le concede autoridad autónoma para decidir verdad. Su salida se vuelve aceptable sólo cuando puede vincularse con objetos del contrato. Esta restricción reduce el riesgo de explicaciones plausibles sin soporte. Además, facilita revisión humana: el evaluador puede revisar el *claim*, abrir la fuente, inspeccionar el *span*, comparar la etiqueta y decidir si la respuesta pública respeta los límites.

La contribución de diseño, por tanto, no es un módulo aislado. Es una forma de coordinar componentes conocidos bajo una interfaz común de evidencia. *Retrieval*, NLI, señales visuales, metadatos y *grounding* no se presentan como soluciones independientes, sino como campos de un expediente. El valor de la arquitectura aparece cuando esos campos permanecen disponibles para auditoría.

4.10. Criterios técnicos de auditabilidad

Las decisiones técnicas relevantes son las que modifican una propiedad evaluable del sistema: trazabilidad, separación de responsabilidades, control de afirmaciones no soportadas o posibilidad de revisión humana. Los detalles operativos importan sólo cuando permiten reproducir una métrica, entender una limitación o explicar por qué un módulo afecta la evidencia registrada. Por eso, el diseño se describe por funciones, contratos y efectos evaluables.

La decisión central es separar entrada, normalización, recuperación, comparación, agregación y respuesta. Esa separación permite evaluar cada etapa con criterios distintos. La recuperación se juzga por cobertura y fuente decisiva; la comparación se juzga por relación semántica; la respuesta se juzga por *grounding* y ausencia de afirmaciones no soportadas. La arquitectura no se define por empaquetado local, sino por la estructura de evidencia que conserva.

También es relevante explicar por qué el contrato antecede a la respuesta. Una arquitectura conversacional puede producir texto convincente sin conservar el razonamiento que lo originó. El contrato invierte esa prioridad: primero se registran objetos verificables y después se redacta la salida pública. Esta decisión reduce el espacio para alucinación no detectada y permite que una persona revise el expediente. En términos de diseño, el lenguaje natural es una capa de comunicación; la evidencia estructurada es la capa de justificación.

Otra decisión importante es mantener límites como objetos explícitos. Muchos sistemas tratan errores, ausencia de fuente o falta de imagen como fallas internas. En este estudio, esas condiciones se vuelven parte del resultado. Si no hay evidencia visual suficiente, se registra. Si una fuente sólo aporta contexto, se marca. Si NLI no resuelve la relación, se conserva el diagnóstico. Esta política mejora la honestidad de la respuesta, aunque pueda hacerla menos contundente.

Por último, el diseño distingue entre utilidad técnica y utilidad académica. Una categoría como fuente decisiva ayuda porque expresa una condición epistémica; un detalle operativo sólo importa si permite reproducir una métrica o entender una limitación.

Con este criterio, el capítulo siguiente muestra cómo el diseño se materializó en artefactos ejecutables, contratos persistidos y validaciones que permiten evaluar auditabilidad.

5. Implementación del prototipo

La Figura 5.1 muestra por qué la implementación multimodal debe registrar preguntas, evidencia web y justificación, no sólo una señal visual aislada.

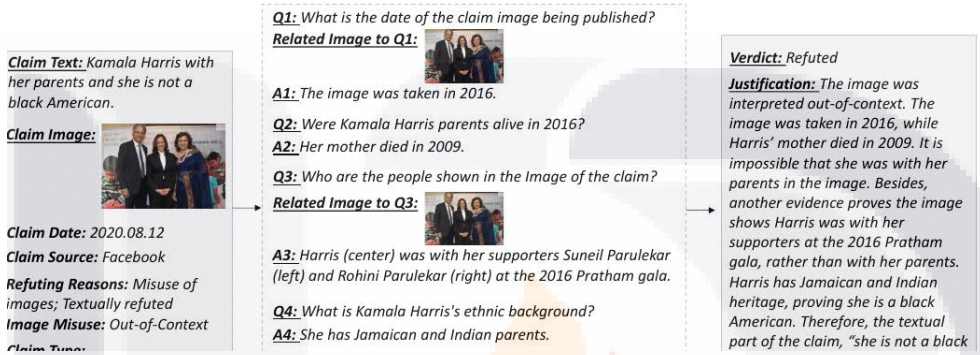

Figura 5.1: Claim imagen-texto con preguntas de evidencia multimodal. [4].
Fuente: AVerImaTeC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web, figura 1, p. 2.

La Figura 5.2 sintetiza el flujo operativo que conecta entrada, recuperación, comparación y respuesta revisable.

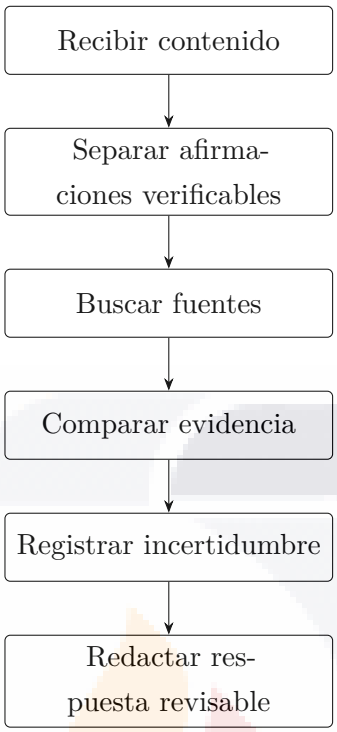

Figura 5.2: Flujo propio de verificación y respuesta pública.
Fuente: elaboración propia.

5.1. Entorno técnico

El prototipo se organiza como una cadena operativa de análisis auditable. La ejecución se expone mediante una interfaz de línea de comandos, un servicio web y un flujo de agente. El flujo usa Agno como capa de orquestación; Agno es un marco en Python para construir agentes que coordinan llamadas a modelos de lenguaje, herramientas externas y pasos intermedios bajo un control de flujo explícito. Sobre esa capa, funciones deterministas ensamblan contratos, aplican reglas de agregación, escriben artefactos y ejecutan validación. Esta división importa para la auditabilidad: la orquestación decide qué paso se ejecuta, pero el contenido del expediente lo producen funciones verificables y no la improvisación del modelo.

El entorno de ejecución contempla dos modos. El modo controlado usa evidencia controlada y permite validar el contrato sin dependencias externas. El modo abierto usa herramientas configuradas: búsqueda web, Fact Check API, extracción de URLs, OCR/visión/metadatos, NLI local y respaldo de modelo de lenguaje cuando está disponible.

La Tabla 5.1 consolida las tecnologías y versiones principales que sostienen la reproducibilidad del prototipo. Las versiones corresponden a las restricciones declaradas

en el archivo de dependencias del proyecto.

Tabla 5.1: Tecnologías y dependencias principales del prototipo

Fuente: elaboración propia a partir del archivo de dependencias del proyecto.

Componente	Tecnología	Versión
Lenguaje base	Python	≥ 3.11
Orquestación de agentes	Agno	2.6.12
Ejecución local del comparador	Ollama	$\geq 0.5.0$
Modelo comparador	Gemma 4 E4B, `hf.co/google/gemma-4-E4B-it-qat-q4_0-gguf:Q4_0`	QAT, Q4_0, GGUF
Validación de contratos	Pydantic, jsonschema	≥ 2.12, ≥ 4.25
Extracción de contenido web	trafilatura, Newspaper4k, Crawl4AI, BeautifulSoup	< 2.1, $\geq 0.9.5$, $\geq 0.8.9$, ≥ 4.13
Búsqueda web	ddgs	≥ 9.0
Servicio y API	FastAPI, uvicorn	≥ 0.124, ≥ 0.38
Inferencia local de modelos	transformers, torch	≥ 4.52, ≥ 2.7

5.2. Entradas y salidas

El sistema acepta texto, URL e imagen. Cada ejecución produce cuatro tipos de artefactos:

- contrato de evidencia por muestra;
- manifiesto de configuración, herramientas disponibles, tiempos y estado de muestras;
- reporte de validación de esquema y *grounding*;
- registro de validación, degradación y refinamiento.

Estos artefactos son tan importantes como la respuesta pública. El estudio evalúa el sistema por la calidad del rastro persistido, no sólo por el texto visible.

5.3. Recuperación de evidencia y extracción de URL

El modo abierto usa búsqueda abierta y la Fact Check API de Google cuando hay credenciales. Para extraer el contenido de las URLs, el protocolo documenta una cadena de bibliotecas especializadas en recuperar el texto principal de una página web: `trafilatura` como extractor primario, seguido de Newspaper4k, Crawl4AI y BeautifulSoup cuando el contenido principal no está disponible. Estas bibliotecas se intentan en orden hasta obtener un texto utilizable, y el agente registra calidad de fuente, calidad de extracción, hechos extraídos y campos faltantes como parte del expediente.

Esta decisión responde a un problema práctico: una fuente relevante no siempre expone el contenido necesario de forma limpia. El sistema no debe tratar una URL de alta reputación como evidencia suficiente si no contiene los campos requeridos para el *claim*.

5.4. NLI y selección de spans

La implementación separa selección de fragmentos de inferencia NLI. En modo abierto, el NLI local usa un modelo multilingüe configurable y se carga de forma perezosa. Si local NLI no está disponible o no es concluyente, puede registrarse una ruta de respaldo. El resultado se almacena como señal auditada en el contrato, junto con limitaciones cuando aplica.

5.5. Imagen y degradación

Las señales de imagen se ejecutan como tareas acotadas. OCR, etiquetas, detección web, metadatos, C2PA y búsqueda de marcas se registran como evidencia o limitación. Si una herramienta falla, excede tiempo o no está configurada, el contrato debe mostrarlo. Esta política hace que la degradación sea parte del resultado y no un detalle oculto.

5.6. Validación controlada

La ejecución controlada inicial valida tres contratos contra el esquema actual. El reporte de validación confirma que los objetos producidos cumplen la estructura esperada. Esta evidencia demuestra que la implementación puede producir artefactos compatibles con el esquema del contrato de evidencia. No debe interpretarse como evaluación de desempeño general, porque usa casos deterministas.

5.7. Entorno de ejecución abierto

La implementación conserva rutas de ejecución para análisis individuales y para interacción abierta. Esa ruta abierta permitió verificar texto, URL e imagen con degradación explícita. Estas ejecuciones aportan evidencia de funcionalidad técnica, pero no sustituyen la evaluación pública extendida.

En conjunto, la implementación dejó los objetos necesarios para pasar de una arquitectura propuesta a una evaluación verificable. Los contratos, manifiestos y reportes de validación permiten observar si el agente conserva evidencia, reconoce límites y evita afirmaciones públicas no soportadas. Por eso el siguiente capítulo puede evaluar el prototipo desde métricas de trazabilidad, soporte y clasificación secundaria, no sólo desde la impresión textual de sus respuestas.

5.8. Implementación y soporte de auditabilidad

El capítulo de implementación cumple una función concreta dentro del estudio: mostrar cómo la arquitectura propuesta se volvió un sistema ejecutable y revisable. No basta con indicar que existe un agente; la exposición técnica muestra qué módulos producen el contrato, qué herramientas pueden fallar, qué evidencia queda persistida y qué parte de la respuesta pública depende de esos registros.

5.8.1. Estructura de módulos

La implementación activa distribuye responsabilidades entre modelos de datos, orquestación, clientes externos, validación, datasets y comparación con el comparador. Esa división permite revisar una ejecución sin confundir la afirmación analizada, las operaciones que producen evidencia y el texto que finalmente llega al usuario.

La separación modular no elimina deuda técnica ni convierte al sistema en producto terminado. Su aporte es más acotado: cuando una métrica cambia, una herramienta se degrada o una fuente no cubre el *claim*, el expediente permite ubicar qué parte del proceso originó esa lectura. Por eso la implementación se evalúa como infraestructura auditable, no como clasificador universal.

5.8.2. Interfaces de ejecución

El prototipo expone rutas de ejecución para experimentos reproducibles y para uso interactivo. La interfaz de línea de comandos, el servicio web y el flujo de agente comparten el mismo ensamblaje de evidencia, de modo que las ejecuciones de modo abierto y evaluación pueden compararse sin cambiar el contrato conceptual del análisis [50].

Esta decisión también fija un límite. La presencia de interfaces no significa que el estudio entregue un producto de usuario final; significa que el flujo puede activarse desde más de una entrada y aun así conservar artefactos revisables. La evaluación, por tanto, se concentra en trazabilidad, soporte y degradación documentada.

5.8.3. Extracción, comparación y señales multimodales

La extracción de URL se trata como una cadena controlada porque las páginas reales varían en estructura, limpieza y accesibilidad. `trafilatura`, Newspaper4k, Crawl4AI y BeautifulSoup se usan como alternativas de extracción, pero una fuente sólo cuenta como evidencia suficiente si expone contenido útil para el *claim*. Una página reputada que no contiene los campos necesarios queda registrada como evidencia limitada, no como soporte decisivo [16], [50].

La comparación semántica sigue la misma lógica. El NLI local conserva una señal auditable de relación entre afirmación y evidencia; cuando no está disponible o no resulta concluyente, el sistema registra la ruta de respaldo como condición del expediente [19], [20], [21].

Las señales de imagen se manejan por separado, como OCR, etiquetas, metadatos, procedencia y contexto, para evitar que una señal débil se convierta en conclusión fuerte [10], [4]. La revisión visual se entiende como evidencia contextual y limitada, no como prueba automática de veracidad [12], [13].

5.8.4. Persistencia, validación y degradación

Cada ejecución deja artefactos que permiten reconstruir el análisis: contrato de evidencia, manifiesto de configuración, reporte de validación y registro de refinamiento. Estos materiales registran configuración, herramientas disponibles, tiempos, contratos y fallos. La persistencia es parte del argumento metodológico porque vuelve verificable una respuesta que, de otro modo, quedaría reducida a prosa conversacional [5], [14].

La validación revisa esquema, *grounding* y respuesta pública antes de usar resultados como evidencia de tesis. Si una herramienta falla o no está configurada, el contrato debe mostrarlo; si una fuente no decide el caso, la respuesta debe conservar esa limitación. Esta política explica por qué el prototipo puede ser útil aunque sus métricas secundarias de clasificación no superen globalmente al comparador: la contribución principal está en la ruta de revisión que deja disponible [9], [3].

5.9. Flujo de datos y degradación observable

La implementación traduce el contrato de evidencia en un flujo de datos persistente. La entrada se recibe como texto, enlace o referencia de imagen; después se normaliza para identificar afirmaciones verificables y requisitos factuales. La recuperación genera candidatos de evidencia y conserva metadatos mínimos de fuente. La comparación semántica opera sobre fragmentos seleccionados y registra soporte, contradicción, neutralidad o insuficiencia. La respuesta pública se produce al final, cuando ya existen objetos que permiten justificarla o limitarla.

La persistencia es parte del diseño técnico. Cada ejecución conserva archivos que permiten reconstruir el análisis sin depender de la memoria del proceso. El manifiesto registra configuración, modo de ejecución, muestra analizada y rutas de salida. El contrato conserva afirmaciones, evidencias, comparaciones, señales y limitaciones. El reporte de validación indica si los objetos cumplen el esquema esperado. Esta separación facilita depuración: un fallo de extracción, un fallo de *retrieval*, un fallo de NLI o una afirmación no soportada pueden ubicarse en campos diferentes.

La degradación observable se implementa como una política explícita. Cuando una fuente no puede extraerse, cuando la imagen no está disponible, cuando una señal de procedencia no existe o cuando la comparación semántica no alcanza suficiencia, el sistema registra la limitación. La respuesta no oculta esas fallas ni las convierte en conclusiones fuertes. Esta decisión es importante porque las dependencias del sistema son variables: resultados de búsqueda, páginas web, formatos de imagen, disponibilidad de metadatos y latencia del modelo pueden cambiar entre ejecuciones.

El prototipo conserva además una frontera clara entre modo controlado y modo abierto. El modo controlado permite verificar el contrato con muestras conocidas y resultados reproducibles. El modo abierto permite analizar entradas reales con búsqueda, extracción y servicios disponibles en el momento de ejecución. El estudio no mezcla estos modos como si tuvieran el mismo peso experimental. El modo controlado demuestra estructura y validación; el modo abierto demuestra viabilidad operativa y revela límites de dependencia externa.

La implementación de NLI y selección de spans sigue la misma lógica. La comparación semántica no decide sola el veredicto; produce una señal que se combina con cobertura, calidad de fuente, requisitos faltantes y limitaciones. Un *span* relacionado puede no cubrir el *claim* completo. Un resultado neutral puede indicar que la fuente habla del tema sin resolverlo. Una contradicción requiere incompatibilidad sustantiva y no sólo diferencia de redacción. Por eso el sistema registra la comparación como objeto

revisable.

El manejo de imágenes se mantiene deliberadamente acotado. OCR, metadatos, procedencia, marcas y búsqueda contextual aparecen como señales, no como prueba absoluta. La implementación puede registrar que una imagen no está disponible, que no contiene texto útil, que no tiene credenciales verificables o que su procedencia es insuficiente. Ese registro evita que la respuesta pública sugiera una verificación visual que no ocurrió. El estudio conserva así una frontera honesta: multimodalidad significa integrar señales disponibles, no resolver forense avanzado.

Finalmente, el entorno de ejecución expone la misma estructura a distintas superficies: línea de comandos, API y flujo de agente. Esta variedad no cambia el contrato; sólo cambia la forma de invocar el análisis. La estabilidad metodológica depende de esa decisión. Si cada interfaz produjera una salida distinta, la evaluación perdería comparabilidad. Al conservar un contrato común, el prototipo permite discutir arquitectura, implementación y resultados como partes de un mismo sistema.

5.10. Implementación como soporte de investigación

No toda decisión de implementación tiene el mismo peso para la investigación. Las relevantes son aquellas que transforman una entrada en evidencia revisable y que, por tanto, sostienen una métrica, una hipótesis o una limitación. Por eso la exposición se organiza alrededor de funciones —extracción de afirmaciones, recuperación de fuentes, selección de fragmentos, comparación semántica, registro de señales visuales, validación y respuesta pública—: es en ese nivel funcional, y no en el de los detalles de empaquetado, donde la implementación afecta lo que el estudio puede afirmar.

El nivel de detalle se ajusta al argumento: se explica qué se implementó y por qué importa para los resultados, y se omiten rutas, comandos y detalles de empaquetado que envejecen rápido y no modifican la interpretación. Así la arquitectura no queda como propuesta abstracta ni el capítulo se reduce a documentación de software.

La persistencia de artefactos es un ejemplo de decisión técnica con valor académico. Guardar contratos, métricas y diagnósticos no es una comodidad operativa; es la condición que permite auditar el experimento. Sin persistencia, la respuesta del agente sería efímera y la evaluación dependería de confianza en una ejecución no inspeccionable. Con persistencia, una afirmación de la investigación puede rastrearse a un registro concreto, una métrica documentada o una limitación observada.

La validación de esquema cumple una función similar. No basta con que el agente produzca una respuesta lingüística; debe producir objetos con campos esperados y

relaciones coherentes. La validación confirma que el expediente tiene estructura suficiente para evaluación. Cuando falla, el problema no es sólo técnico: el estudio pierde la base para afirmar auditabilidad. Por esa razón, validación y contrato se presentan como parte del método, no como detalle secundario.

La implementación también materializa la distinción entre modo controlado y modo abierto. El modo controlado permite comprobar consistencia interna; el modo abierto permite observar dependencias reales y variabilidad. Ambos son necesarios porque una investigación aplicada debe mostrar que el sistema funciona bajo condiciones verificables y que reconoce límites cuando se acerca a escenarios abiertos. Esta separación impide usar resultados controlados como prueba de generalización y evita interpretar fallos de web abierta como errores exclusivamente algorítmicos.

Finalmente, la implementación demuestra que auditabilidad requiere disciplina de degradación. Cuando una fuente no se puede extraer, una imagen no está disponible o un *span* no cubre el *claim* completo, el sistema debe comunicar esa condición. La respuesta fundamentada se vuelve más útil cuando conserva incertidumbre en lugar de ocultarla. Esta política conecta código, método y ética de comunicación: un agente para desinformación no debe maximizar contundencia, sino soporte revisable. A partir de esos artefactos persistidos, la evaluación puede medir si el prototipo realmente conserva trazabilidad, soporte de afirmaciones y límites visibles, en lugar de juzgarlo sólo por la fluidez de su respuesta final.

6. Evaluación y resultados

La Figura 6.1 ilustra por qué la evaluación de factualidad debe observar afirmaciones atómicas y soporte, no sólo la apariencia global de una respuesta.

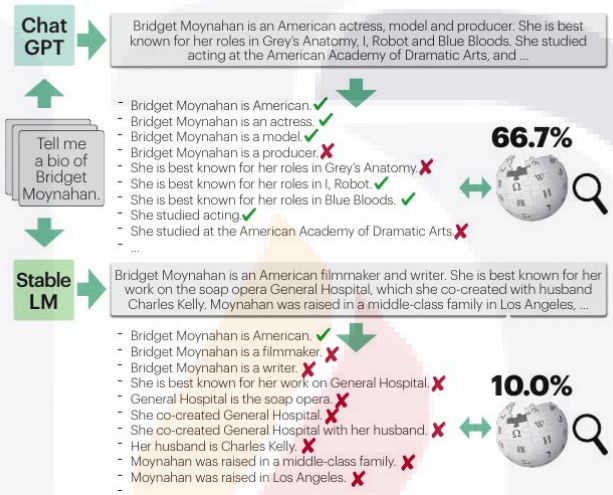

Figura 6.1: Evaluación de factualidad mediante afirmaciones atómicas. [5].
Fuente: FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation, figura 1, p. 1.

La Figura 6.2 presenta las métricas clasificatorias secundarias; la Figura 6.3 separa el comportamiento por dataset y fuente decisiva.

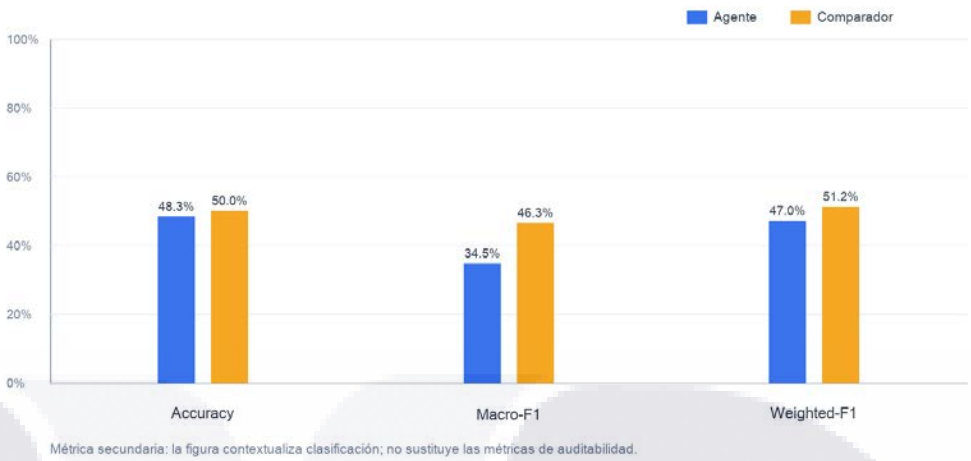

Figura 6.2: Comparación de accuracy, macro-F1 y weighted-F1 entre agente y comparador.

Fuente: elaboración propia.

Figura 6.3: Matriz por dataset de compatibilidad de etiqueta y fuente decisiva.

Fuente: elaboración propia.

La Figura 6.4 distingue métricas primarias de auditabilidad y métricas secundarias de clasificación para evitar interpretar todos los resultados como la misma evidencia.

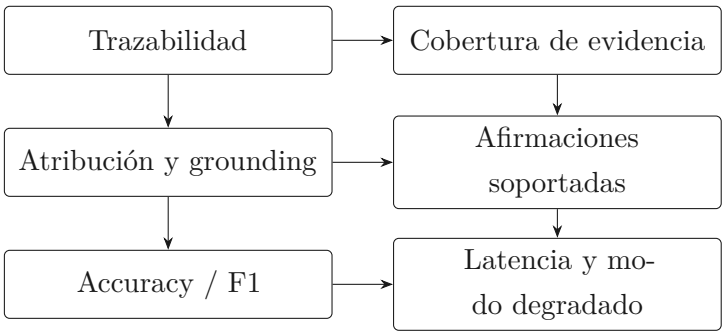

Figura 6.4: Matriz propia de métricas primarias y secundarias.
Fuente: elaboración propia.

6.1. Diseño de evaluación

La evaluación siguió el diseño experimental: comparar un agente estructurado contra un comparador basado en Gemma 4 E4B ejecutado mediante Ollama, con búsqueda simple. La comparación se interpretó por familias de métricas. Las métricas primarias midieron auditabilidad; las secundarias midieron compatibilidad de etiqueta y desempeño clasificatorio. Esta separación evitó reducir el estudio a accuracy.

6.2. Evaluación local

La evaluación local analizó una ejecución controlada del agente y un comparador comparable. El reporte de decisión local indicó avance metodológico: los contratos fueron válidos, el soporte en evidencia estuvo completo y hubo evidencia comparable del comparador. El resumen reportó validación de contrato 1.0, traceability rate 1.0, cobertura de evidencia 1.0 y unsupported-statement rate 0.0.

La compatibilidad de etiqueta local fue 0.3333 sobre tres muestras. Por el tamaño de la muestra, esta cifra no tiene validez estadística y no debe leerse como resultado clasificatorio; su única función fue confirmar que existían artefactos comparables antes de escalar a la evaluación pública extendida.

6.3. Evaluación pública extendida

La evaluación pública extendida se ejecutó sobre 60 muestras. Metodológicamente, este bloque se interpreta como evaluación comparativa del prototipo frente a un comparador estricto basado en Gemma 4 E4B. El reporte de evaluación pública extendida registró ejecución completa del comparador en las 60 muestras.

6.4. Métricas primarias

En la evaluación extendida, el agente obtuvo:

- *Retrieval* hit rate: 1.0.
- Decisive-source rate: 0.6.
- Traceability rate: 1.0.
- Unsupported-statement rate: 0.0.

Estas métricas respaldaron la contribución principal. El agente no sólo produjo un veredicto; conservó evidencia, limitaciones y soporte en evidencia. La tasa 0.0 de afirmaciones no soportadas fue especialmente relevante porque conectó el diseño anti-alucinación con un resultado medible.

6.5. Métricas secundarias de clasificación

Las métricas de clasificación son secundarias por diseño: el comparador no recibe el contrato de evidencia del agente y la política de agregación del agente es deliberadamente conservadora, de modo que estas cifras describen el costo clasificatorio de priorizar auditabilidad y no el objetivo central del estudio. Con esa lectura, la clasificación no mostró superioridad global frente al comparador. Los resultados verificados, con intervalos de confianza al 95 % por *bootstrap*, fueron:

Tabla 6.1: Comparación secundaria de clasificación entre agente y comparador (IC 95 % por *bootstrap*)

Fuente: elaboración propia.

Métrica	Agente [IC 95 %]	Comparador [IC 95 %]	Delta
Accuracy	0.4833 [0.350, 0.617]	0.5 [0.383, 0.633]	-0.0167
Macro-F1	0.3454 [0.252, 0.509]	0.4633 [0.356, 0.564]	-0.1179
Weighted-F1	0.47 [0.347, 0.602]	0.5121 [0.402, 0.638]	-0.0421

Los intervalos de confianza de agente y comparador se solapan ampliamente en las tres métricas. El intervalo *bootstrap* de la diferencia de macro-F1, que es la métrica con mayor delta aparente, va de $-0{,}2505$ a $0{,}0829$ e incluye el cero; por tanto, la inferioridad clasificatoria observada no alcanza significancia estadística con esta muestra. Esta tabla fija un límite honesto: el trabajo no demostró que el agente fuera mejor clasificador global, pero tampoco hay evidencia estadística de que el comparador lo supere de forma sistemática. Lo que sí demostró es un rastro auditable robusto, y la clasificación queda como línea de refinamiento.

6.6. Resultados por dataset

El resumen por *dataset* mostró variabilidad:

- `averimatec`: n=15, compatibilidad del agente 0.4667, compatibilidad del comparador 0.5333.
- `averitec`: n=15, compatibilidad del agente 0.2667, compatibilidad del comparador 0.1333.
- `fever_gold`: n=15, compatibilidad del agente 0.8667, compatibilidad del comparador 0.9333.
- `multi_fc`: n=15, compatibilidad del agente 0.3333, compatibilidad del comparador 0.4.

La mejora relativa apareció en AVeriTeC, pero no en la mayoría de familias. FEVER Gold mostró el mejor valor absoluto del agente, aunque el comparador siguió arriba. Esto confirmó que el sistema debe presentarse por sus métricas de auditabilidad y no por desempeño clasificatorio promedio.

6.7. Diagnósticos

Los diagnósticos principales fueron:

- `nli_stance_mismatch`: 10.
- `image_evidence_unavailable`: 8.
- `overly_conservative_aggregation`: 8.
- `relevant_evidence_not_decisive`: 5.

Estas categorías convirtieron errores en agenda técnica. El sistema falló principalmente donde se esperaba mayor dificultad: calibración de *stance*/NLI, disponibilidad de evidencia visual completa, agregación conservadora y selección de evidencia decisiva.

6.8. Interpretación

Los resultados respaldaron H1 y H5 en términos de auditabilidad. H2 se respaldó por *retrieval* hit rate 1.0, aunque decisive-source rate 0.6 mostró que recuperar evidencia no equivalía a resolver todas las afirmaciones. H3 quedó parcialmente respaldada: el sistema conservó señales NLI, pero los diagnósticos indicaron desajustes. H4 quedó parcialmente respaldada porque las señales visuales se registraron como limitaciones o contexto, pero la disponibilidad de evidencia visual completa sigue siendo débil.

6.9. Amenazas a la validez

La evaluación usó un conjunto fijo de 60 muestras y mapeos de etiqueta conservadores. La dependencia de búsqueda web y APIs introdujo variabilidad. La comparación con el comparador fue metodológicamente limpia bajo una configuración estricta, pero no agotó todas las configuraciones de instrucción posibles. La clasificación secundaria puede verse afectada por la política conservadora de no convertir ausencia de evidencia en falsedad.

6.10. Análisis ampliado de la evaluación

El análisis ampliado interpreta los resultados por familias métricas, límites y relación con los objetivos de investigación.

6.10.1. Separación de familias métricas

Auditabilidad y clasificación se reportan como familias métricas separadas porque F1, trazabilidad y tasa de afirmaciones no soportadas responden preguntas diferentes; por eso la evaluación pública extendida presenta *retrieval* hit rate 1.0, traceability rate 1.0 y unsupported-statement rate 0.0 junto a F1 menor al comparador [5], [3], [14].

6.10.2. Comparador estricto

El comparador estricto basado en Gemma 4 E4B es metodológicamente útil porque no recibe el contrato del agente ni etiquetas esperadas. La comparación fue completa en las 60 muestras y permite contrastar una respuesta conversacional no estructurada contra una arquitectura con expediente de evidencia [9], [50].

La elección de Gemma 4 E4B ejecutado mediante Ollama en modo estricto responde a tres criterios. Primero, reproducibilidad: es un modelo abierto que corre de forma local, sin depender de una API comercial cuyas versiones cambian sin aviso. Segundo, comparabilidad de recursos: opera bajo las mismas condiciones de cómputo local que el agente, de modo que la diferencia observada no se atribuye a un modelo de mayor escala. Tercero, control: el modo estricto exige una respuesta válida del modelo sin heurísticas de respaldo, lo que aísla el efecto de añadir un contrato de evidencia. Por estas razones, la comparación es deliberadamente exploratoria y no exhaustiva: el comparador prueba una ruta no estructurada de búsqueda simple y generación, y ofrece una referencia estricta para contrastar el costo y el beneficio de introducir un expediente de evidencia, pero no representa todas las configuraciones de instrucción, *retrieval* aumentado o modelos especializados que podrían usarse en trabajo futuro.

6.10.3. Interpretación por dataset

Los resultados por *dataset* muestran que el comportamiento del agente depende del tipo de evidencia disponible. El reporte por *dataset* muestra mejoría relativa en AVeriTeC y debilidad relativa en otras familias [22], [24], [2].

6.10.4. Diagnósticos como agenda técnica

Los diagnósticos transforman errores en rutas de mejora: NLI, evidencia visual, agregación y fuente decisiva. Los conteos principales son `nli_stance_mismatch` 10, `image_evidence_unavailable` 8, `overly_conservative_aggregation` 8 y `relevant_evidence_not_decisive` 5 [19], [25], [4].

6.10.5. Bootstrap e incertidumbre

Los intervalos *bootstrap* al 95 % acotan la incertidumbre de las diferencias observadas en el conjunto de 60 muestras. Los deltas negativos frente al comparador son pequeños y están rodeados de variabilidad muestral, especialmente el de macro-F1: su intervalo va de $-0{,}2505$ a $0{,}0829$ e incluye el cero, lo que impide afirmar una diferencia clasificatoria estadísticamente significativa. La evidencia sostiene una lectura directa: el prototipo conserva trazabilidad y soporte de evidencia, y la inferioridad clasificatoria aparente no es concluyente bajo esta muestra [22], [2].

La muestra permite observar patrones, familias de error y la tensión entre auditabilidad y clasificación, pero no autoriza generalizar a todo el espacio de la desinformación digital. Por eso cada cifra se reporta junto con su intervalo y se interpreta como evidencia acotada sobre el prototipo, no como una medida de desempeño universal.

6.10.6. Condiciones de producción de cada métrica

Las amenazas principales son tamaño de muestra, dependencia de web viva, mapeo de etiquetas, configuración del comparador y cobertura visual. Estos límites no invalidan la evaluación, pero sí delimitan lo que puede afirmarse con rigor. Los reportes conservados y sus identificadores de verificación reducen ambigüedad sobre qué ejecución sostiene los números, mientras que las limitaciones registradas evitan presentar al agente como juez autónomo de verdad [2], [4], [14].

Cada métrica se interpreta junto con su condición de producción: qué muestra se evaluó, qué evidencia estaba disponible, qué etiquetas se mapearon y qué dependencias externas pudieron afectarla. Reportar esas condiciones es lo que permite distinguir un límite del diseño de un artefacto de la medición.

6.10.7. Resultado sustentado

El resultado sustentado es que el agente produce rastros auditables sólidos, no que clasifica mejor de forma global. La evaluación local confirma que el contrato puede generarse y validarse; la evaluación pública extendida muestra que esa estructura conserva trazabilidad y control de afirmaciones no soportadas en 60 muestras. Esta evidencia sostiene la contribución de auditabilidad, pero no autoriza una afirmación de superioridad predictiva global [1], [5], [14].

La conclusión evaluativa conecta métricas primarias con objetivos específicos: trazabilidad para revisar el proceso, tasa de afirmaciones no soportadas para controlar la respuesta pública, fuente decisiva para medir suficiencia de evidencia y métricas clasificatorias para ubicar compatibilidad con etiquetas. La mejora predictiva queda como trabajo futuro; no debe presentarse como resultado cumplido.

6.10.8. Validez de la lectura evaluativa

El criterio de validez de esta evaluación no consiste en volver perfectos los resultados, sino en explicar con precisión qué evidencia existe y qué límites permanecen. Una evaluación aceptable para una tesis aplicada puede mostrar fortalezas de trazabilidad y, al mismo tiempo, reconocer límites de clasificación, tamaño muestral, cobertura visual y dependencia de fuentes externas [8], [3].

La revisión de atribución y recuperación ayuda a leer esos límites como condiciones de mejora, no como permiso para ocultar resultados desfavorables [14], [50].

La validez de esta evaluación se apoya en reproducibilidad documental, consistencia entre métricas y ausencia de resultados inventados. La aprobación local no sustituye la revisión de asesoría, los requisitos institucionales ni la validación con más muestras; su alcance es más acotado: demostrar que los números reportados provienen de artefactos persistidos y no se transforman en promesas no respaldadas.

6.10.9. Matriz de lectura entre objetivos y resultados

La evaluación puede leerse como una matriz entre las preguntas de investigación, los objetivos específicos y los resultados observados. El primer objetivo, analizar por qué una respuesta conversacional puede ser difícil de revisar, se responde mostrando que una etiqueta o explicación final no basta si no conserva evidencia, alcance e incertidumbre. El segundo objetivo, definir qué información debe preservarse, se responde con el expediente de análisis que registra afirmación, fuentes, evidencia, relación semántica, límites y respuesta pública. El tercer objetivo, diseñar y desarrollar un agente que no se presente como autoridad final, se responde con la separación entre recuperación, com-

paración, agregación conservadora y redacción condicionada por evidencia. El cuarto objetivo, evaluar el aporte del rastro explícito, se responde comparando trazabilidad, soporte de afirmaciones, claridad de límites y compatibilidad de etiqueta frente a una referencia conversacional no estructurada. El quinto objetivo, identificar límites, se responde con diagnósticos sobre evidencia parcial, ambigüedad semántica, señales visuales incompletas y costos de cautela.

La evaluación local cumple una función de validez interna y la evaluación pública extendida una de validez externa. La primera comprueba que el sistema puede producir el expediente que el estudio necesita para contestar sus preguntas; la segunda comprueba si ese expediente conserva utilidad al aplicarse a muestras públicas heterogéneas. Ninguna de las dos convierte al prototipo en árbitro de verdad ni en *benchmark* definitivo; juntas precisan qué aporta un rastro explícito de evidencia y qué límites aparecen cuando la evidencia no es decisiva.

La comparación con el comparador también se interpreta por capas. En clasificación, el comparador conserva ventaja promedio. En auditabilidad, el agente conserva estructura que el comparador no produce como objeto validable. Esta diferencia no permite declarar una victoria total de un sistema sobre otro; permite explicar qué se gana y qué se pierde al pasar de una respuesta generativa simple a una arquitectura con contrato. Se gana trazabilidad, control de límites y posibilidad de inspección. Se pierde simplicidad y, en esta versión del prototipo, no se logra una mejora clasificatoria global. La evaluación es valiosa precisamente porque muestra esa tensión de forma explícita.

6.10.10. Profundización de resultados por familia de datos

AVeriTeC tensiona la recuperación y comparación en entorno cercano al fact-checking web. La mejora relativa del agente en esa familia sugiere que el contrato ayuda cuando el problema exige organizar evidencia abierta y conservar limitaciones. Sin embargo, el resultado no debe extrapolarse de forma amplia porque son 15 muestras y porque la disponibilidad web puede cambiar. En esta familia, el estudio puede decir que la arquitectura muestra una señal favorable de utilidad de revisión, no que resuelve el *dataset*.

FEVER Gold tensiona una situación distinta: evidencia controlada y afirmaciones diseñadas para comparación textual. El desempeño alto del agente en esta familia confirma que el flujo puede funcionar cuando existe evidencia relevante y relativamente directa. Al mismo tiempo, el comparador queda por arriba, lo que muestra que la estructura auditable no garantiza ventaja de etiqueta incluso en condiciones favorables. Este resultado es útil porque separa competencia predictiva de inspeccionabilidad: el

agente puede producir un expediente más revisable aunque no sea el mejor asignador de etiquetas.

MultiFC tensiona diversidad de dominios, estilos de *claim* y fuentes. El desempeño menor en esa familia es consistente con un sistema que todavía depende de selección de evidencia y reglas conservadoras. AVerImaTeC tensiona la parte multimodal: la falta de imagen original o evidencia visual suficiente aparece como límite metodológico. Estas dos familias son importantes porque impiden que el estudio quede apoyado sólo en casos favorables. Muestran que la propuesta necesita mejores estrategias de recuperación, normalización y comparación en escenarios heterogéneos.

6.10.11. Criterio de lectura de los deltas negativos

Los deltas negativos frente al comparador se reportan como parte central del resultado. De ellos se desprende una conclusión precisa: el prototipo mejora la forma en que se documenta el razonamiento, pero no logra una mejora promedio en compatibilidad de etiquetas. Reportarlos de forma explícita es lo que sostiene esa conclusión y la separa de una lectura sesgada por casos favorables.

También permite formular trabajo futuro con prioridades claras. Si el objetivo fuera sólo subir F1, podría bastar con relajar agregación o copiar estrategias del comparador. Pero ese camino podría aumentar afirmaciones no soportadas. La mejora deseable debe preservar unsupported-statement rate bajo, trazabilidad completa y limitaciones explícitas. Por eso los deltas negativos no implican abandonar la arquitectura; implican calibrarla. La pregunta futura no es simplemente cómo hacer que el agente prediga más etiquetas esperadas, sino cómo hacerlo sin perder el contrato que justifica el estudio.

6.11. Entorno experimental y comparador

La evaluación se ejecutó bajo un entorno local acotado. El hardware objetivo documentado fue un MacBook Pro con Apple M1 y 16 GB de RAM. Esta condición importa porque el prototipo se diseñó para una tesis aplicada reproducible, no para una infraestructura distribuida ni para entrenamiento de modelos. La latencia, la disponibilidad de herramientas y la selección de muestras deben leerse dentro de esa restricción.

El comparador principal fue Gemma 4 E4B vía Ollama, identificado como `hf.co/google/gemma-4-E4B-it-qat-q4_0-gguf:Q4_0`. Este identificador describe el artefacto exacto que se ejecutó: la variante instruida (*it*) del modelo, distribuida en formato GGUF con cuantización consciente del entrenamiento (*qat*) a cuatro bits (*q4_0*), lo que permite correrla en un equipo de escritorio sin GPU dedicada. El comparador se ejecutó como comparación estricta con modelo de lenguaje y registró 60/60 muestras

con *success rate* 1.0. Esto evita una comparación artificial contra una salida vacía o degradada: la línea base usada en la evaluación extendida fue un comparador real y estricto bajo el mismo problema de etiquetado.

La latencia agregada de la evaluación pública extendida fue min 3574.85 ms, max 59103.17 ms y media 24052.88 ms. Estos valores explican por qué el trabajo futuro incluye robustez operativa: un expediente auditable puede ser metodológicamente útil y, al mismo tiempo, requerir optimización para uso interactivo. El estudio no interpreta la latencia como criterio de veracidad, sino como límite de adopción práctica del *pipeline*.

La validación controlada de la etapa de implementación se interpreta aparte. En modo controlado valida estructura, contratos y respuesta fundamentada, pero no invoca Gemma ni mide desempeño real del modelo. Por ello, esos resultados no se usan para afirmar calidad predictiva. La evidencia empírica principal sobre comportamiento con modelo local proviene de la evaluación pública extendida y de sus reportes comparativos.

6.12. Lectura integrada de métricas y límites

Los resultados se leen en dos planos. En el plano de auditabilidad, el agente muestra un comportamiento consistente: *retrieval* hit rate 1.0, traceability rate 1.0 y unsupported-statement rate 0.0 en la evaluación pública extendida. Estos números indican que el sistema recupera evidencia, conserva trazas y evita introducir afirmaciones sustantivas sin soporte registrado bajo las condiciones evaluadas. El resultado es importante porque corresponde al objetivo principal del estudio: producir expedientes revisables.

En el plano clasificatorio, la lectura es distinta. El agente obtiene accuracy 0.4833 frente a 0.5 del comparador, macro-F1 0.3454 frente a 0.4633 y weighted-F1 0.47 frente a 0.5121. Estos resultados no sostienen superioridad global sobre el comparador. La interpretación correcta no consiste en esconder esa diferencia, sino en ubicarla: la arquitectura mejora inspeccionabilidad, pero todavía no resuelve por completo la asignación de etiquetas. El estudio conserva ambas lecturas porque pertenecen a preguntas diferentes.

El decisive-source rate 0.6 es una métrica intermedia especialmente relevante. Un *retrieval* hit rate alto muestra que el sistema encuentra material relacionado, pero una fuente decisiva exige cubrir los requisitos centrales del *claim*. La diferencia entre ambas métricas confirma una de las premisas del diseño: relevancia no equivale a suficiencia. En desinformación real, muchos documentos mencionan el tema sin resolver la afirmación. El contrato obliga a registrar esa diferencia en lugar de convertirla en una etiqueta

fuerte.

Los diagnósticos agregados completan la lectura. Las categorías `nli_stance_mismatch`, `image_evidence_unavailable`, `overly_conservative_aggregation` y `relevant_evidence_not_decisive` muestran dónde se concentran los límites. No son sólo errores; son pistas sobre el tipo de mejora necesaria. El desajuste NLI/*stance* indica que la comparación semántica necesita más robustez ante matices temporales, negación, ambigüedad y formulaciones largas. La falta de evidencia visual indica que la multimodalidad depende de disponibilidad real de imagen, metadatos y búsqueda inversa. La agregación conservadora muestra una tensión entre cautela y compatibilidad con etiqueta esperada. La evidencia relacionada no decisiva señala que el sistema debe distinguir mejor entre contexto y prueba.

La evaluación también permite responder a las hipótesis sin forzar conclusiones. H1 queda respaldada porque la evidencia estructurada mejora trazabilidad y control de afirmaciones no soportadas. H2 queda respaldada parcialmente: *retrieval* separado mejora cobertura, pero la fuente decisiva no siempre aparece. H3 queda parcial o no concluyente porque el registro NLI existe, pero los desajustes siguen siendo frecuentes. H4 queda parcial porque las señales visuales aportan límites y contexto, aunque la cobertura visual es incompleta. H5 queda respaldada en términos metodológicos: las métricas de auditabilidad capturan una contribución que accuracy y F1 por sí solos no describen.

La lectura integrada evita dos errores. El primer error sería declarar éxito total porque las métricas de trazabilidad son fuertes. El segundo sería declarar fracaso total porque las métricas clasificatorias quedan por debajo del comparador. Ninguna lectura es suficiente. La evidencia muestra un prototipo útil para producir expedientes auditables y limitado como clasificador global. Esa tensión es precisamente el resultado sustentado: la auditabilidad puede mejorar la revisión sin resolver todavía todo el problema predictivo.

6.13. Lectura cualitativa de errores y aciertos

La interpretación de resultados mejora cuando se observan aciertos y errores como familias de comportamiento, no sólo como conteos agregados. En los aciertos más claros, el sistema recupera una fuente pertinente, selecciona un fragmento comparable y formula una respuesta cautelosa. Estos casos muestran el valor del contrato: la etiqueta no aparece sola, sino acompañada por evidencia y límites. Una persona puede revisar si el *span* cubre el requisito central del *claim*, si la fuente es adecuada y si la respuesta pública mantiene el mismo nivel de certeza que la evidencia.

En los errores por evidencia no decisiva, el sistema suele encontrar documentos relacionados, pero no una prueba suficiente. Esta categoría es especialmente importante porque representa un fallo común en verificación real. Un documento puede mencionar la entidad correcta, el evento correcto o un tema cercano, pero no resolver causalidad, temporalidad, cantidad o atribución. Si el sistema convirtiera esa relación temática en soporte fuerte, aumentaría riesgo de sobreafirmación. El comportamiento conservador reduce ese riesgo, aunque también puede bajar compatibilidad con la etiqueta esperada.

Los errores de NLI y *stance* tienen otro perfil. En estos casos, el problema no es sólo encontrar evidencia, sino interpretar la relación semántica entre el *claim* y el fragmento. Una negación, una condición temporal o un cambio de sujeto pueden alterar el veredicto. El diagnóstico `nli_stance_mismatch` indica que la comparación automática todavía necesita calibración. Sin embargo, registrar el desajuste ya es útil: permite localizar el punto de revisión y evita que la respuesta final oculte una inferencia débil.

Los casos con evidencia visual no disponible muestran el límite multimodal más visible. Esta categoría no se interpreta como falla accidental de implementación, sino como recordatorio metodológico: verificar imagen-texto depende de acceso a imagen original, contexto, búsqueda inversa, metadatos y fuentes externas. Cuando esas señales no aparecen, la respuesta responsable conserva insuficiencia. Esta decisión protege al sistema de convertir ausencia de metadatos en sospecha injustificada o de aceptar una imagen por mera apariencia.

La evaluación, por tanto, no sólo mide desempeño; también enseña qué tipo de sistema se está construyendo. Un clasificador optimizado para etiqueta puede ocultar por qué acertó o falló. Un agente auditable deja expuesta la ruta, incluso cuando no coincide con la etiqueta de referencia. Esa exposición convierte los errores en material de mejora: ajustar *retrieval*, calibrar NLI, fortalecer evidencia visual y refinar agregación sin perder trazabilidad.

7. Discusión

El alcance real del prototipo se interpreta a partir de los resultados ya reportados, distinguiendo aportación y límite: el sistema demuestra una arquitectura auditable con evidencia persistida, pero no demuestra superioridad clasificatoria global ni preparación para despliegue operativo. Ambas lecturas se sostienen a la vez, sin reducir el balance a una conclusión optimista ni a una interpretación limitada a accuracy y F1. La discusión aborda tres preguntas: qué sostiene el diseño, qué queda abierto por las métricas y los diagnósticos, y cómo se cierran las hipótesis H1-H5. Las figuras siguientes ubican los patrones de error y las rutas de mejora. La Figura 7.1 muestra la frecuencia de diagnósticos principales y permite leer los límites como líneas de mejora concretas.

Figura 7.1: Frecuencia de diagnósticos principales en la evaluación pública extendida.
Fuente: elaboración propia.

La Figura 7.2 organiza esos diagnósticos en una taxonomía interpretativa.

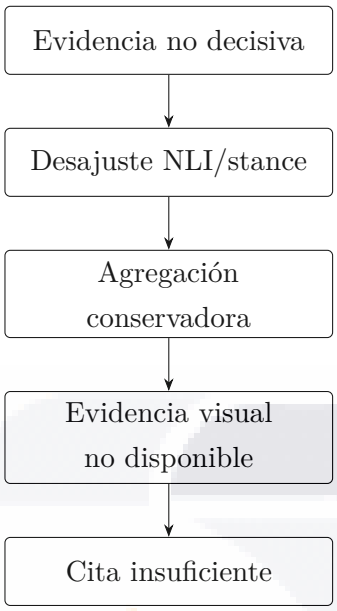

Figura 7.2: Taxonomía propia de fallos observados.
Fuente: elaboración propia.

7.1. Alcances demostrados

El prototipo demuestra que es posible implementar un agente que produce evidencia auditable antes de redactar una respuesta. La prueba más fuerte no es una etiqueta más acertada, sino la persistencia de contratos válidos, *grounding* completo y ausencia de afirmaciones públicas sin soporte. En términos de tesis, esto responde al objetivo central: diseñar y desarrollar una arquitectura que haga visible el proceso de análisis.

7.2. Limitaciones demostradas

El prototipo no demuestra superioridad clasificatoria global. Tampoco demuestra que el sistema sea un verificador autónomo de verdad, ni que pueda resolver cualquier *claim* multimodal. No procesa video, no realiza forense avanzado de imágenes, no entrena modelos y no crea datasets propios. Estas ausencias deben leerse como límites de alcance, no como fallas ocultas.

7.3. Respuesta a hipótesis

H1 se respalda porque el sistema alcanza traceability rate 1.0 y unsupported-statement rate 0.0 en la evaluación pública extendida. H2 se respalda de forma parcial: *retrieval* hit rate 1.0 indica cobertura, pero decisive-source rate 0.6 muestra que no toda evidencia recuperada es suficiente. H3 queda abierta: NLI y *stance* existen como

señales auditadas, pero `nli_stance_mismatch` aparece como diagnóstico principal. H4 se respalda sólo como política de explicación multimodal: las señales de imagen se registran sin sobreinterpretarse, aunque la cobertura visual completa queda limitada. H5 se respalda en la medida en que las métricas de auditabilidad capturan la contribución mejor que accuracy aislada.

7.4. Comparación con el comparador

El comparador estricto basado en Gemma 4 E4B es deliberadamente simple: recibe la misma entrada y búsqueda simple, pero no el contrato estructurado del agente. La comparación muestra que el contrato mejora la auditabilidad, no necesariamente la clasificación. Esto es consistente con el diseño: la política conservadora puede reducir falsos respaldos, pero también puede producir etiquetas de insuficiencia en casos donde el comparador arriesga una respuesta.

7.5. Multimodalidad

El tratamiento multimodal es útil como estructura de señales, pero limitado como evidencia final. Registrar OCR, metadatos o contexto visual ayuda a explicar el análisis, pero no resuelve por sí solo la veracidad de una noticia. Por ello, los metadatos y las marcas de procedencia se tratan como señales contextuales, no como veredictos. El trabajo futuro debe mejorar acceso a imágenes completas y contexto visual, no endurecer señales débiles como si fueran prueba definitiva.

7.6. Implicaciones

La contribución del estudio es una forma de diseñar agentes conversacionales más responsables para dominios donde la confianza importa. En vez de pedir al modelo de lenguaje que ”sea veraz”, el sistema le exige trabajar sobre evidencia persistida y limita su respuesta con validación. Este patrón puede trasladarse a otros dominios: salud, derecho, educación o análisis documental, siempre que se adapte el contrato de evidencia.

El aporte del estudio se ordena en tres planos: lo que el prototipo demostró como arquitectura auditable, lo que no demostró como clasificador general y lo que debe mejorarse en trabajo futuro. Estos tres planos sintetizan la contribución sin ocultar límites ni convertir la trazabilidad en una promesa de verdad automática.

7.7. Discusión de hipótesis y alcance

La discusión conecta las hipótesis con los resultados sin convertir métricas favorables en conclusiones más amplias que la evidencia.

7.7.1. Hipótesis H1-H5 como cierre evaluativo

Las hipótesis no se leen como aprobadas o rechazadas de forma mecánica. Cada una se interpreta según la familia de evidencia que la evalúa. H1 y H5 reciben el respaldo más claro porque las métricas de trazabilidad y afirmaciones no soportadas corresponden directamente al objetivo de auditabilidad. H2 queda respaldada con cautela: recuperar evidencia no significa encontrar una fuente decisiva. H3 y H4 permanecen parciales porque los diagnósticos de NLI y disponibilidad visual muestran límites todavía importantes.

Las métricas positivas no se convierten en promesas generales: el expediente mejora la inspeccionabilidad bajo las condiciones evaluadas, pero el agente no resuelve de manera autónoma todos los casos de desinformación. Esa diferencia delimita el alcance del trabajo sin sobreafirmarlo.

7.7.2. Auditabilidad aunque F1 no mejore

El resultado clasificatorio secundario no debe ocultarse. El comparador conserva mejores valores agregados de accuracy y F1. Sin embargo, esos valores no describen por sí solos si una respuesta puede revisarse. La aportación del agente está en conservar *claim*, fuente, *span*, comparación, limitación y *grounding* como objetos verificables. Esa estructura permite revisar una decisión incluso cuando la etiqueta final no coincide con la referencia.

Por eso la discusión no presenta una victoria total del agente sobre el comparador. Presenta una tensión: se gana trazabilidad, control de afirmaciones no soportadas y transparencia de límites; no se gana superioridad clasificatoria global. Esa tensión es más útil académicamente que una conclusión optimista, porque señala con precisión dónde madura el prototipo y dónde necesita trabajo futuro.

7.7.3. Conservadurismo y costo predictivo

La agregación conservadora protege contra sobreafirmación, pero tiene costo. Cuando el sistema no promueve evidencia relacionada a evidencia decisiva, puede emitir etiquetas cautelosas que bajan compatibilidad con datasets. Ese comportamiento no debe celebrarse como éxito completo ni descartarse como falla simple. Debe entenderse como un punto de calibración: el prototipo necesita distinguir mejor entre evidencia

insuficiente, contradicción explícita y soporte parcial.

La mejora futura no consiste sólo en subir F1. Si relajar la agregación aumenta etiquetas esperadas pero también aumenta afirmaciones no soportadas, se perdería la contribución central. La mejora deseable debe subir desempeño sin romper trazabilidad, *grounding* ni declaración de límites.

7.7.4. Multimodalidad como límite abierto

El estudio incorpora señales de imagen, pero no demuestra verificación visual completa. OCR, metadatos, procedencia y marcas pueden aportar contexto, aunque no prueban por sí solos la verdad de un *claim*. La categoría `image_evidence_unavailable` muestra que el límite no es sólo técnico; también es metodológico. Una imagen puede estar fuera de contexto, no estar disponible, no tener metadatos útiles o no resolver la afirmación textual que la acompaña.

El valor del prototipo está en registrar esa insuficiencia sin convertirla en veredicto. Esta cautela evita conclusiones injustificadas y reconoce que la multimodalidad es uno de los puntos donde una respuesta generativa puede sonar más segura de lo que la evidencia permite.

7.7.5. Transferibilidad del patrón

El patrón de contrato de evidencia puede transferirse a otros dominios informativos sensibles, pero no de forma automática. Salud, derecho, educación o análisis institucional requerirían fuentes, vocabularios, criterios de suficiencia y validaciones externas propias. Lo transferible no es una etiqueta ni un conjunto fijo de herramientas, sino la disciplina de separar evidencia, comparación, limitación y respuesta.

Esta aportación metodológica es relevante porque muchos sistemas conversacionales ocultan el proceso detrás de una respuesta fluida. El estudio muestra una alternativa: antes de comunicar, registrar; antes de concluir, comparar; antes de afirmar, revisar soporte. Esa lógica puede adaptarse, siempre que cada dominio defina qué cuenta como evidencia suficiente.

7.7.6. Alcance de la contribución

La interpretación integrada separa tres preguntas. La primera es si el problema existe y está delimitado: la literatura de fact-checking, atribución, *grounding* y evaluación de respuestas con citas justifica la necesidad de respuestas inspeccionables. La segunda es si la solución corresponde al problema: el contrato de evidencia, la persistencia de artefactos y la validación de campos responden directamente a la falta de

trazabilidad. La tercera es si los resultados sostienen la contribución: las métricas de auditabilidad la sostienen, mientras que las métricas de clasificación delimitan lo que el prototipo todavía no resuelve.

Esa separación evita dos lecturas incompletas. Una lectura demasiado optimista convertiría la trazabilidad en garantía de veracidad; eso sería incorrecto. Una lectura demasiado negativa reduciría el trabajo a los deltas de accuracy y F1; eso también sería incompleto. La interpretación más consistente presenta el estudio como prueba aplicada de una arquitectura auditable: suficiente para demostrar diseño, implementación y evaluación de un expediente revisable; insuficiente para presentarse como clasificador autónomo listo para despliegue operativo.

7.7.7. Criterios de suficiencia epistémica

Una fuente recuperada no aporta soporte sólo por contener palabras similares al *claim*. Debe cubrir entidad, tiempo, lugar, cantidad, relación causal o condición factual relevante. Una comparación NLI tampoco decide por sí sola; debe revisarse junto con el *span*, la calidad de la fuente y las limitaciones. Una respuesta pública clara sigue siendo insuficiente si no puede rastrearse a evidencia concreta y declarar lo que permanece inconcluso.

Estos criterios explican por qué el contrato funciona como mecanismo de prudencia. El sistema no se evalúa sólo por cuántas veces acierta una etiqueta, sino por cómo llega a ella y por qué decide no afirmarla cuando falta soporte. La clasificación secundaria funciona como sensor de costo; la trazabilidad funciona como criterio de revisión.

7.7.8. Implicaciones para sistemas conversacionales

Un asistente usado en temas informativos sensibles necesita mostrar más estructura que una respuesta común. El usuario necesita saber qué *claim* se analizó, qué fuentes se consultaron, qué evidencia se descartó y qué incertidumbres permanecen. El texto generado se vuelve una capa de comunicación sobre artefactos auditables, no el lugar donde se decide la verdad del caso.

Esta implicación importa porque la fluidez lingüística produce autoridad aparente. Una respuesta convincente puede ocultar evidencia débil. Por eso la arquitectura introduce fricción deliberada: etiquetas cautelosas, limitaciones explícitas, diagnóstico de degradación y registro reproducible. Esa fricción puede hacer la respuesta menos contundente, pero aumenta su revisabilidad.

7.8. Casos conceptuales de interpretación

La discusión se aclara al imaginar cuatro tipos de caso. El primero es una afirmación textual con evidencia suficiente. En ese escenario, una fuente identificable cubre entidad, fecha, acción y contexto. El agente puede producir una respuesta con etiqueta fuerte siempre que cite el fragmento correspondiente y no agregue información externa. Este caso muestra el funcionamiento ideal del contrato: la respuesta pública resume una ruta de evidencia que puede inspeccionarse.

El segundo caso es una afirmación con evidencia relacionada pero no decisiva. Aquí las fuentes recuperadas hablan del tema, pero no cubren causalidad, magnitud, temporalidad o alcance. Una arquitectura orientada sólo a etiqueta podría forzar una respuesta. En cambio, el diseño orientado a la auditabilidad conserva insuficiencia. Esta decisión puede reducir compatibilidad con una etiqueta esperada, pero protege contra sobreafirmación. En términos de tesis, ese comportamiento es coherente con el objetivo de inspeccionabilidad.

El tercer caso es una imagen auténtica fuera de contexto. La pregunta no es sólo si la imagen fue manipulada, sino si corresponde al lugar, fecha y evento afirmados por el texto. Metadatos, procedencia, OCR o búsqueda inversa pueden aportar señales, pero la conclusión depende de cómo esas señales se relacionan con el *claim*. Si la imagen aparece previamente en otro contexto, la contradicción se dirige a la atribución contextual. Si no hay metadatos, la ausencia se registra como límite y no como refutación automática.

El cuarto caso combina texto verificable e imagen no concluyente. La afirmación textual puede evaluarse con fuentes documentales, mientras la imagen permanece como señal débil o irrelevante. Este caso justifica que el contrato separe *claim* textual, evidencia web y señales visuales. Una respuesta responsable puede sostener o debilitar el *claim* textual y, al mismo tiempo, explicar que la imagen no aporta prueba decisiva.

Estos casos muestran por qué el estudio no se sostiene como sistema de verdad automática. Su alcance es más específico: el agente organiza evidencia para que una persona pueda revisar dónde surge la certeza, dónde aparece la duda y dónde faltan datos. Esa contribución es compatible con resultados clasificatorios imperfectos. De hecho, los resultados imperfectos refuerzan la importancia de conservar el expediente. Cuando la etiqueta falla, el rastro permite diagnosticar por qué falló; cuando la etiqueta acierta, el rastro permite revisar si acertó por la razón correcta.

7.9. Criterios para extensiones futuras

Las extensiones futuras se evalúan por su capacidad de mejorar el expediente, no sólo por elevar una métrica aislada. Una mejora de *retrieval* se considera valiosa si aumenta fuentes decisivas y conserva spans revisables. Una mejora de NLI se considera valiosa si reduce desajustes sin convertir la comparación semántica en juez final. Una mejora visual se considera valiosa si registra disponibilidad, procedencia y límites, no si promete forense total. Una mejora de generación se considera valiosa si produce respuestas más claras sin introducir afirmaciones no soportadas.

También se requiere ampliar evaluación con cuidado. Más muestras, más dominios y más comparadores pueden fortalecer la validez externa, pero sólo si se mantiene la separación entre métricas de auditabilidad y métricas de clasificación. Incluir revisores humanos sería especialmente útil porque permitiría medir si el expediente realmente facilita inspección: tiempo de revisión, tipos de corrección, acuerdo entre evaluadores y utilidad percibida. Esa línea conectaría el prototipo con herramientas asistidas para periodistas, analistas y usuarios críticos.

La discusión final, por tanto, no propone abandonar la etiqueta ni absolutizar la trazabilidad. Propone leerlas juntas. La etiqueta comunica una decisión resumida; la trazabilidad permite cuestionarla. En sistemas conversacionales para desinformación, esa segunda función es indispensable porque una respuesta fluida puede generar confianza indebida. El estudio aporta una forma de hacer visible lo que normalmente queda oculto detrás de la fluidez del modelo.

7.10. Adopción responsable y límites de comunicación

Una herramienta auditable puede fallar si comunica sus resultados como certeza automática. Por ello, la adopción responsable exige diseñar tanto el expediente como la forma de presentarlo. La respuesta pública debe mostrar el nivel de soporte, distinguir evidencia decisiva de contexto y señalar cuándo una conclusión no puede sostenerse. El usuario no debe interpretar una fuente recuperada como prueba suficiente ni una etiqueta cautelosa como indecisión arbitraria. La comunicación debe enseñar qué se sabe, qué se infiere y qué permanece abierto.

Este punto es especialmente relevante en entornos de alta circulación, donde una respuesta breve puede compartirse sin el expediente que la sostiene. Si el sistema produce una etiqueta sin contexto, puede convertirse en autoridad opaca. Si produce una explicación demasiado extensa, puede volverse impracticable para usuarios no expertos. La tensión de diseño está en ofrecer una respuesta legible y, al mismo tiempo, conservar

acceso a evidencia, citas, comparaciones y limitaciones. El estudio resuelve esta tensión de manera inicial al separar respuesta pública y contrato de evidencia.

La adopción en periodismo requeriría controles adicionales. Un periodista necesita saber qué fuentes fueron consultadas, cuáles quedaron fuera, qué fragmentos sostienen cada afirmación y qué incertidumbre permanece. También necesita detectar cuándo el sistema se volvió demasiado conservador o cuándo aceptó evidencia débil. En este escenario, el agente funcionaría como asistente de preparación: acelera la organización inicial, pero la decisión editorial sigue siendo humana.

La adopción en alfabetización mediática tiene otro énfasis. El valor no está sólo en resolver un caso, sino en enseñar una práctica de verificación. Mostrar que una imagen puede estar fuera de contexto, que una cita puede no sostener una afirmación o que una fuente relacionada no es necesariamente decisiva ayuda al usuario a desarrollar criterio. En este escenario, el agente debe evitar tono de autoridad total y favorecer explicaciones que inviten a revisar evidencia.

La adopción institucional exige trazabilidad y consistencia. Equipos de monitoreo necesitan comparar casos, conservar decisiones y revisar cambios en el tiempo. El contrato de evidencia puede apoyar ese trabajo porque normaliza campos y diagnósticos. No obstante, el estudio no cubre gobernanza, privacidad, escalamiento, seguridad ni integración con flujos profesionales. Esos límites deben quedar claros antes de cualquier despliegue fuera del contexto académico.

En síntesis, la contribución no termina en el algoritmo. Un agente para desinformación opera en un campo social sensible: una conclusión equivocada puede amplificar sospechas, desacreditar contenido legítimo o dar falsa tranquilidad. Por eso, la auditabilidad es también una postura de comunicación responsable. Hacer visible la ruta de evidencia reduce la dependencia en la autoridad del modelo y devuelve al lector la posibilidad de cuestionar. Con esa lectura, las conclusiones ya no necesitan presentar una promesa nueva: deben sintetizar qué quedó demostrado, qué quedó limitado y qué trabajo futuro puede fortalecer la misma cadena de evidencia.

8. Conclusiones y trabajo futuro

8.1. Conclusión general

El estudio diseñó, implementó y evaluó un agente conversacional para analizar desinformación digital desde un principio de auditabilidad. La aportación central no consiste en afirmar que el sistema decide la verdad de manera autónoma, ni en sostener superioridad clasificatoria frente a todos los comparadores. El resultado sustentado es más específico: el prototipo convierte una entrada potencialmente desinformativa en un expediente revisable, con afirmación normalizada, evidencia recuperada, comparación semántica, señales visuales, limitaciones y respuesta fundamentada.

Esa cadena de verificación se entiende como una secuencia de etapas observables —detección de la afirmación, recuperación de evidencia, comparación semántica, registro de señales visuales y redacción condicionada— y no como una caja única que emite un veredicto. Esa organización es la que vuelve evaluable al prototipo: permite localizar en qué etapa se gana evidencia, en cuál se pierde suficiencia y en cuál se introduce incertidumbre, de modo que la contribución pueda juzgarse por operaciones concretas y no por la impresión global que produce la respuesta.

El problema de investigación partió de una tensión conocida en sistemas conversacionales: una respuesta fluida puede parecer confiable aunque no muestre la ruta que la sostiene. En desinformación, esa opacidad es particularmente riesgosa. Una noticia falsa, una imagen fuera de contexto o una afirmación incompleta no se resuelven sólo con una etiqueta. Requieren identificar qué se afirma, qué fuente cubre la afirmación, si el fragmento recuperado es decisivo, si existen señales visuales pertinentes y si la respuesta final evita añadir hechos no respaldados. La arquitectura propuesta responde a esa tensión desde la auditabilidad.

8.2. Respuesta a preguntas y objetivos

La primera pregunta abordó las condiciones que hacen difícil revisar una respuesta conversacional sobre desinformación. El análisis mostró que la dificultad aparece cuando la salida oculta la afirmación revisada, la evidencia consultada, la relación entre fuente y conclusión, y la incertidumbre que permanece. El objetivo correspondiente se cumplió al analizar ese problema de revisión y al desplazar el centro del estudio desde la etiqueta final hacia la inspeccionabilidad del análisis.

La segunda pregunta abordó la información que necesita conservarse para que

una afirmación digital pueda revisarse críticamente. La respuesta metodológica fue un expediente que conserva afirmación, fuentes, evidencia, comparación, señales contextuales, limitaciones y respuesta pública. El objetivo correspondiente se cumplió al definir esos elementos mínimos y al mostrar por qué cada uno reduce una forma distinta de opacidad.

La tercera pregunta abordó cómo puede organizarse un agente conversacional para producir respuestas útiles sin presentarse como autoridad final de verdad. El prototipo respondió separando recuperación, comparación semántica, señales multimodales, agregación conservadora y redacción condicionada por evidencia. En la evaluación local, la validación del contrato alcanzó 1.0, la trazabilidad 1.0 y la tasa de afirmaciones no soportadas 0.0. Estos resultados muestran coherencia interna: el sistema genera objetos revisables y no sólo respuestas textuales.

La cuarta pregunta abordó qué aporta un rastro explícito de evidencia frente a una respuesta conversacional que sólo entrega una explicación final. La evaluación extendida de 60 muestras mostró *retrieval* hit rate 1.0, decisive-source rate 0.6, traceability rate 1.0 y unsupported-statement rate 0.0. Al mismo tiempo, la comparación contra el comparador no mostró superioridad clasificatoria global. La respuesta a la pregunta es, por tanto, precisa: el rastro explícito mejora la revisabilidad y el control de soporte, pero no garantiza mejor etiqueta promedio.

La quinta pregunta abordó los límites que aparecen ante texto, enlaces, imágenes o evidencia incompleta. Los diagnósticos muestran límites en evidencia visual disponible, comparación semántica, fuente decisiva y agregación conservadora. El objetivo correspondiente se cumplió al identificar esos límites sin convertirlos en excusas ni ocultarlos detrás de las métricas favorables de auditabilidad. La literatura sobre fact-checking explicable, atribución y factualidad respalda esta forma de separar soporte, comunicación y evaluación [1], [5], [14].

La evaluación también recupera condiciones experimentales que no deben omitirse. El entorno objetivo fue un MacBook Pro con Apple M1 y 16 GB de RAM. El modelo principal para el comparador fue Gemma 4 E4B vía Ollama, con identificador `hf.co/google/gemma-4-E4B-it-qat-q4_0-gguf:Q4_0`; la comparación estricta con modelo de lenguaje registró 60/60 muestras con estrategia de modelo de lenguaje y success rate 1.0. La latencia del agente en evaluación pública extendida fue min 3574.85 ms, max 59103.17 ms y media 24052.88 ms. Estos datos no son decoración técnica: delimitan el alcance operativo de la conclusión.

La delimitación del alcance se sostuvo con la misma lógica. El prototipo no se

presenta como detector universal de noticias falsas. La clasificación secundaria quedó por debajo del comparador global: accuracy 0.4833 frente a 0.5, macro-F1 0.3454 frente a 0.4633 y weighted-F1 0.47 frente a 0.5121. Estos números impiden formular una conclusión de superioridad predictiva. Sin embargo, no cancelan la contribución de auditabilidad; la precisan y ayudan a responder la quinta pregunta sobre límites.

8.3. Aportaciones

La primera aportación es conceptual. Un agente conversacional de fact-checking puede leerse como infraestructura de revisión, no como oráculo de veracidad. Esta postura desplaza el criterio de éxito: además de preguntar si la etiqueta coincide con una referencia, se pregunta si una persona puede inspeccionar la ruta que produjo la respuesta. En un dominio donde la confianza excesiva en modelos generativos puede amplificar errores, esa inspeccionabilidad es una condición de uso responsable.

La segunda aportación es arquitectónica. El contrato de evidencia funciona como frontera entre operaciones técnicas y afirmaciones comunicables. *Retrieval*, NLI, señales visuales y respuesta pública no se mezclan en un texto único; se articulan como objetos diferenciados. Esta estructura facilita auditoría, depuración y evaluación. También hace visible una diferencia que suele perderse: evidencia relacionada no equivale a evidencia suficiente, fuente confiable no equivale a fuente decisiva y procedencia visual no equivale a veracidad factual.

La tercera aportación es experimental. La evaluación separa métricas de auditabilidad y métricas de clasificación. Esta separación evita dos lecturas simplistas. La primera sería declarar éxito total porque la trazabilidad es alta. La segunda sería declarar fracaso total porque el F1 queda por debajo del comparador. Los resultados muestran una contribución intermedia y respaldada: el agente produce expedientes sólidos para revisión, pero todavía necesita mejorar compatibilidad de etiquetas y robustez semántica.

La cuarta aportación es metodológica. El documento conserva una cadena de evidencia entre literatura, diseño, implementación, evaluación y conclusiones. Las figuras de literatura incorporadas en el marco teórico refuerzan este punto: el campo ya representa la verificación como *pipeline* explicable, evidencia descompuesta, generación con citas, afirmaciones imagen-texto y evaluación atómica de factualidad. El prototipo toma esas ideas y las integra en una estructura local orientada a trazabilidad.

8.4. Limitaciones

La primera limitación es la comparación semántica. Las categorías de diagnóstico muestran que `nli_stance_mismatch` sigue siendo un punto crítico. El sistema puede recuperar evidencia pertinente y aun así no resolver adecuadamente la relación entre *claim* y fuente. La inferencia textual en desinformación exige manejar negación, temporalidad, entidades cercanas, ambigüedad causal y formulaciones parciales. Por ello, NLI debe tratarse como señal auditada, no como juez final.

La segunda limitación es la fuente decisiva. El *retrieval* hit rate alto indica cobertura temática, pero el decisive-source rate 0.6 muestra que no siempre aparece evidencia suficiente para cerrar el caso. Esta brecha es central: muchos documentos mencionan el tema, pero no cubren la condición específica del *claim*. La arquitectura lo registra como insuficiencia en lugar de convertirlo en una respuesta fuerte, pero esa cautela también reduce compatibilidad con etiquetas esperadas.

La tercera limitación es multimodal. El prototipo incorpora texto, enlace e imagen, pero no resuelve forense visual avanzado. Las señales visuales ayudan a registrar OCR, metadatos, procedencia, disponibilidad y contexto; no garantizan por sí mismas autenticidad ni veracidad. En afirmaciones imagen-texto, una imagen auténtica puede estar fuera de contexto, una imagen sintética puede acompañar una afirmación verificable por otras fuentes y la ausencia de metadatos no prueba falsedad. La literatura de AVerImaTeC y VERITE confirma que la multimodalidad exige evidencia externa y control de sesgos unimodales [11], [4].

La cuarta limitación es el tamaño y composición de la evaluación. Sesenta muestras permiten observar tendencias, comparar familias de datos y generar diagnósticos, pero no bastan para estimar desempeño general en todo el espacio de desinformación digital. La muestra tensiona el sistema con AVeriTeC, FEVER, MultiFC y AVerImaTeC, aunque cada familia conserva sólo una fracción de su complejidad. La generalización debe evaluarse con más casos, más dominios y revisión humana.

La quinta limitación es operativa. Un flujo de evidencia depende de búsqueda, extracción, disponibilidad de páginas, metadatos, latencia y comportamiento del modelo. Estas dependencias cambian con el tiempo. La conservación de artefactos y métricas documentadas mitiga el problema, pero no elimina la variabilidad de la web viva. Esta restricción es inherente a sistemas que verifican contenido digital abierto.

8.5. Implicaciones

La implicación principal para el diseño de agentes es que la respuesta final no debe ser el único producto. En tareas de desinformación, el producto importante es el expediente: una estructura que permite revisar qué se comparó, qué fuente se usó, qué faltó y qué nivel de certeza es razonable. Esta idea se alinea con trabajos de fact-checking explicable y generación con citas, donde la utilidad depende de mostrar soporte y no sólo de producir texto convincente.

La implicación para evaluación es que accuracy y F1 son necesarios, pero insuficientes. Sirven para medir compatibilidad con etiquetas, no para medir si el usuario puede auditar el proceso. Métricas como traceability rate, decisive-source rate, unsupported-statement rate y citation *grounding* capturan otra dimensión: la calidad del rastro. En sistemas conversacionales, esta dimensión es indispensable porque la fluidez puede ocultar incertidumbre.

La implicación para uso humano es igualmente clara. El sistema debe presentarse como apoyo a verificación asistida, no como sustituto de periodistas, analistas o revisores. Su valor está en ordenar señales, reducir búsqueda manual, exponer conflictos y formular límites. Cuando la evidencia es insuficiente, una respuesta cautelosa puede ser más útil que una etiqueta contundente, porque el expediente deja explícito qué información falta para concluir.

8.6. Trabajo futuro

La primera línea de trabajo futuro es mejorar calibración *stance*/NLI. Se requieren modelos o reglas de comparación que manejen mejor temporalidad, negación, equivalencias parciales, afirmaciones compuestas y contradicción contextual. Esta mejora debe conservar explicabilidad: subir una métrica pierde valor si la comparación deja de ser inspeccionable.

La segunda línea es aumentar fuentes decisivas. El *retrieval* debe pasar de cobertura temática a suficiencia probatoria. Esto implica mejores consultas, selección de spans, evaluación de autoridad, control temporal y agregación multifuente. La pregunta futura no es sólo si se recuperó un documento relacionado, sino si ese documento cubre los requisitos centrales del *claim*.

La tercera línea es fortalecer evidencia visual real. El sistema necesita integrar búsqueda inversa, procedencia, OCR, metadatos y señales de manipulación con criterios más finos. Aun así, la mejora visual debe mantener una frontera clara: una señal de imagen puede apoyar, debilitar o contextualizar un *claim*, pero no debe convertirse

automáticamente en veredicto.

La cuarta línea es incorporar evaluación humana. Revisores humanos permitirían medir utilidad práctica del expediente: tiempo de inspección, errores detectados, correcciones justificadas, acuerdo entre revisores y confianza calibrada. Esta línea conectaría el prototipo con escenarios reales de verificación asistida.

La quinta línea es mejorar latencia y robustez del *pipeline*. Un agente auditable debe producir trazas completas sin volverse impracticable. Optimizar recuperación, extracción, comparación y generación permitiría acercar el prototipo a uso interactivo, siempre que la reducción de latencia no elimine campos necesarios para auditoría.

La sexta línea es ampliar comparadores y dominios. La comparación actual con un comparador estricto es útil, pero no agota alternativas de instrucción, *retrieval* aumentado, modelos especializados o herramientas multimodales. La extensión debe mantener comparabilidad: cada nuevo comparador debe evaluarse con las mismas preguntas sobre etiqueta, evidencia, trazabilidad y afirmaciones no soportadas.

8.7. Lectura final de hipótesis

Las hipótesis H1-H5, presentadas en el Capítulo 1 y operacionalizadas en el Capítulo 3, se contrastan aquí con las métricas y los diagnósticos ya reportados. Cada una se cierra con tres elementos: evidencia observada, alcance del respaldo y límite que permanece abierto.

La primera hipótesis planteó que una estructura explícita de evidencia permitiría producir respuestas más auditables. Los resultados respaldan esta afirmación en el sentido metodológico: el sistema conserva afirmaciones, fuentes, spans, comparaciones y límites, y la evaluación pública extendida reporta traceability rate 1.0 junto con unsupported-statement rate 0.0. Esta evidencia no significa que la respuesta sea siempre correcta, sino que deja una ruta verificable para revisar por qué se llegó a ella.

La segunda hipótesis se relacionó con recuperación de evidencia. El *retrieval* hit rate 1.0 muestra que el sistema consigue material relacionado, pero el decisive-source rate 0.6 impide una lectura triunfalista. La hipótesis queda respaldada sólo parcialmente: separar *retrieval* del resto del *pipeline* ayuda a conservar evidencia, aunque todavía no garantiza que esa evidencia sea decisiva. Esta diferencia es una de las aportaciones más importantes de la evaluación, porque obliga a distinguir pertinencia de suficiencia.

La tercera hipótesis trató la comparación semántica. Los diagnósticos muestran que NLI y *stance* aportan estructura, pero también concentran errores. Por ello, el respaldo es limitado. El estudio demuestra que conviene registrar la comparación *claim-*

evidence como objeto propio, pero no demuestra que la implementación actual resuelva la inferencia textual compleja. El valor está en hacer visible el punto de fallo y en evitar que el modelo oculte la incertidumbre detrás de una respuesta fluida.

La cuarta hipótesis abordó multimodalidad. El sistema registra señales visuales, pero la evidencia de imagen no siempre está disponible ni es concluyente. La conclusión, de nuevo, es parcial: integrar imagen, OCR, metadatos o procedencia aporta contexto, aunque no basta para verificación forense. La figura de AVerImaTeC ayuda a interpretar este resultado, porque muestra que verificar afirmaciones imagen-texto requiere preguntas, evidencia web y justificación; no se resuelve mirando la imagen como objeto aislado.

La quinta hipótesis sostuvo que métricas de auditabilidad capturan una contribución que accuracy y F1 no describen por sí solas. Esta hipótesis sí queda respaldada por la separación de resultados. Si sólo se observaran métricas clasificatorias, el agente parecería poco competitivo frente al comparador. Si sólo se observaran métricas de trazabilidad, parecería más exitoso de lo que es. La lectura conjunta muestra el aporte real: una mejora en inspeccionabilidad con límites predictivos todavía abiertos.

8.8. Escenarios de uso y criterio de éxito

El primer escenario de uso es la revisión periodística asistida. En este contexto, el agente no decide qué publicar ni sustituye el criterio editorial. Su función sería preparar un expediente inicial: *claim* normalizado, fuentes candidatas, evidencia relacionada, limitaciones y una respuesta cautelosa. El éxito no se mediría sólo por etiqueta, sino por reducción de tiempo de búsqueda, facilidad para detectar fuentes insuficientes y claridad de los puntos que requieren verificación humana.

El segundo escenario es la alfabetización informacional. Un usuario no experto puede recibir una explicación estructurada que muestre por qué una afirmación no debe aceptarse de inmediato. La respuesta fundamentada puede señalar que existe evidencia relacionada, que falta una fuente decisiva o que la imagen no resuelve el *claim*. En ese caso, el valor educativo aparece en enseñar una práctica de lectura crítica: no confiar en una afirmación sólo porque parece verosímil, ni rechazarla sin revisar evidencia.

El tercer escenario es el análisis institucional de contenidos. Organizaciones que monitorean desinformación necesitan consistencia, trazabilidad y criterios repetibles. Un *pipeline* auditable permite comparar casos, revisar decisiones y detectar patrones de error. Sin embargo, este escenario exige robustez adicional: mayor cobertura de dominios, control de latencia, evaluación humana y políticas claras para manejar incer-

tidumbre. El estudio aporta una base, no un despliegue operativo completo.

En los tres escenarios, el criterio de éxito conserva dos niveles. El nivel predictivo pregunta si la etiqueta es correcta. El nivel de revisión pregunta si el usuario puede revisar la ruta de evidencia. Un sistema que acierta sin mostrar soporte puede ser útil en un benchmark, pero frágil como herramienta pública. Un sistema que muestra soporte pero falla demasiadas etiquetas necesita calibración. La meta futura es acercar ambos niveles: mejorar precisión sin sacrificar inspeccionabilidad.

Confianza y persuasión no son lo mismo. Una respuesta larga, bien redactada o segura no es necesariamente más confiable; la confianza verificable proviene de soporte, atribución, límites y posibilidad de revisión. De ahí que la respuesta pública se trate como último eslabón del *pipeline* y no como producto aislado: la calidad del texto final depende de la calidad del expediente que lo sostiene.

8.9. Criterios de continuidad

La continuidad del proyecto se evalúa con criterios más exigentes que una mejora aislada de accuracy. El primer criterio es conservación de trazabilidad. Cualquier versión futura tendría que permitir reconstruir la ruta entre entrada, *claim*, fuente, comparación y respuesta. Si una optimización hace más opaca esa ruta, se aleja de la contribución central del estudio aunque mejore una métrica secundaria.

El segundo criterio es suficiencia de evidencia. La mejora deseable no consiste en recuperar más documentos, sino en recuperar mejores fuentes para los requisitos del *claim*. Un sistema con muchos enlaces puede seguir siendo débil si ninguno resuelve la afirmación. Por ello, decisive-source rate y revisión de spans deben seguir como métricas centrales. La pregunta no es cuánta evidencia aparece, sino qué parte de la afirmación cubre.

El tercer criterio es calibración de incertidumbre. Una respuesta útil no siempre es una respuesta más fuerte. En algunos casos, la salida correcta será decir que la evidencia disponible no permite concluir. La calibración futura debe reducir conservadurismo innecesario sin caer en afirmaciones no respaldadas. Ese equilibrio es difícil, pero define la diferencia entre un clasificador agresivo y una herramienta de revisión responsable.

El cuarto criterio es integración multimodal verificable. Agregar más señales visuales sólo mejora el sistema si esas señales se conectan con el *claim*. OCR, metadatos, búsqueda inversa, procedencia o detección de manipulación deben registrarse con límites. La multimodalidad no debe convertirse en una promesa general de análisis forense; debe operar como conjunto de evidencias parciales que una persona puede revisar.

El quinto criterio es evaluación con usuarios. Un expediente auditable puede ser técnicamente completo y aun así poco útil si el usuario no entiende cómo revisarlo. Futuras evaluaciones deben observar si periodistas, analistas o estudiantes identifican mejor fuentes insuficientes, detectan contradicciones y calibran confianza con ayuda del sistema. Esa evidencia cerraría la brecha entre métrica interna y utilidad práctica.

Estos criterios permiten que la línea futura conserve coherencia con el estudio. Mejorar el sistema no significa añadir módulos de manera acumulativa, sino fortalecer la cadena de evidencia. Cada componente nuevo debe responder una pregunta: qué parte del expediente mejora, qué límite reduce y cómo se verifica que no introdujo sobreafirmación.

8.10. Balance de aportación y límites

El balance final se entiende como una contribución situada. El prototipo no compite como modelo universal de detección de desinformación, sino como arquitectura de revisión asistida. Esta diferencia cambia el estándar de evaluación. Un sistema universal tendría que demostrar robustez amplia, superioridad predictiva y cobertura multimodal avanzada. Este trabajo demuestra algo más acotado: que una cadena de análisis puede conservar evidencia y límites suficientes para que una respuesta conversacional sea discutible, corregible y auditable.

Ese balance es relevante porque muchas aplicaciones de inteligencia artificial se presentan desde la promesa de automatización. El estudio adopta una postura distinta. Automatizar parte del análisis puede ser útil, pero sólo si no oculta incertidumbre ni reemplaza el juicio humano. El expediente de evidencia funciona como mecanismo de contención: obliga al sistema a mostrar qué sabe, qué encontró, qué comparó y qué no pudo sostener. La respuesta pública gana valor cuando se apoya en ese registro y pierde valor cuando intenta sonar concluyente sin evidencia decisiva.

El trabajo también deja una advertencia metodológica. Agregar *retrieval* a un modelo de lenguaje no resuelve por sí mismo el fact-checking: la recuperación puede traer textos relacionados, contradictorios, desactualizados o incompletos. Las citas tampoco bastan si no respaldan exactamente la afirmación redactada. La literatura de ALCE y FActScore ayuda a entender este punto: la atribución debe evaluarse por afirmaciones, no por presencia decorativa de referencias. El prototipo incorpora esa advertencia al medir *grounding* y afirmaciones no soportadas.

La dimensión multimodal refuerza la misma cautela. Una imagen puede aumentar credibilidad percibida incluso cuando no prueba el *claim*. Puede ser auténtica, ge-

nerada, editada, recortada, reutilizada o simplemente irrelevante. Por eso, las señales visuales no deben integrarse como atajos de veracidad. Deben tratarse como evidencias parciales dentro de un expediente más amplio. Esta decisión limita lo que el prototipo puede afirmar, pero aumenta la honestidad de su comunicación.

El resultado final, entonces, es una base sólida para investigación aplicada. El trabajo articula literatura, diseño, implementación, evaluación y límites con una misma lógica: cada afirmación importante debe poder rastrearse. Cuando la evidencia es fuerte, el sistema puede formular una respuesta más clara. Cuando la evidencia es débil, debe preservar duda. Cuando el comparador supera al agente en clasificación, el resultado se reporta como límite y no se oculta. Esta disciplina es parte de la aportación.

El cierre no presenta el sistema como terminado. La contribución queda mejor delimitada al explicitar lo que falta: mejorar inferencia semántica, aumentar fuentes decisivas, fortalecer señales visuales, reducir latencia, ampliar muestras y validar utilidad con personas. Esas tareas futuras no parten de una idea vaga, sino de diagnósticos concretos y de un contrato que permite medir si una mejora conserva o debilita auditabilidad.

Esta coherencia atraviesa el documento completo. El marco teórico ubica los problemas que el diseño debe resolver; la metodología define qué evidencia cuenta como suficiente para avanzar; la implementación produce los objetos revisables que sostienen cada afirmación; la evaluación separa trazabilidad, evidencia, clasificación y límites; y las conclusiones explican qué tipo de confianza puede construirse cuando el sistema documenta su propio razonamiento. El hilo conductor permanece estable —evidencia suficiente, respuesta cautelosa, revisión humana y trazabilidad—, de modo que el trabajo se explica sin recurrir a promesas generales sobre inteligencia artificial y sin esconder resultados desfavorables. Esa misma disciplina mantiene distinguidos tres planos relacionados pero no equivalentes en una investigación aplicada: el desempeño experimental, la utilidad metodológica y la madurez operativa.

8.11. Cierre

El cierre sostiene una afirmación prudente: es posible convertir una respuesta conversacional sobre desinformación en un objeto revisable. Ese objeto no garantiza verdad absoluta, pero sí mejora las condiciones de inspección. Permite preguntar de dónde salió una afirmación, qué fuente la sostiene, qué comparación se hizo, qué señal visual existe y qué limitación permanece abierta.

El aporte académico está en esa disciplina de trazabilidad. La investigación mues-

tra que un agente puede organizar evidencia de manera útil sin prometer liderazgo predictivo. También muestra que las métricas fuertes de auditabilidad y las métricas débiles de clasificación pueden coexistir, y que esa tensión debe reportarse, no esconderse. La confianza no se solicita por la fluidez de la respuesta; se construye mediante soporte, atribución, límites y revisión.

Glosario

Tabla 8.1: Glosario de conceptos usados en el estudio

Fuente: elaboración propia.

Término	Uso en el estudio
Claim	Afirmación factual que puede contrastarse con evidencia.
Evidence item	Fuente, fragmento o señal registrada para revisar el *claim*.
Span	Fragmento específico usado para comparar *claim* y evidencia.
Stance	Postura de un texto frente a una afirmación: a favor, en contra o neutral.
Grounding	Relación entre una afirmación de la respuesta y su soporte registrado.
Fuente decisiva	Fuente que cubre requisitos centrales del *claim* y permite apoyar o contradecir.
Señal visual	OCR, metadatos, procedencia, marca de agua o indicio contextual asociado a imagen.
Modo degradado	Situación en la que una dependencia falla y el sistema conserva una limitación visible.
Comparador	Línea base no estructurada usada para contextualizar los resultados secundarios.
Agno	Marco de orquestación de agentes en Python que coordina herramientas, modelos y pasos del flujo.

Ollama	Entorno de ejecución local de modelos de lenguaje abiertos en formato cuantizado.
Gemma 4 E4B	Modelo de lenguaje abierto de Google, en su variante de cuatro mil millones de parámetros efectivos, usado como comparador.

Estos términos se usan de forma consistente para evitar cruces de sentido. En particular, evidencia relacionada no significa evidencia suficiente; procedencia no significa verdad; trazabilidad no significa acierto clasificatorio; y ausencia de evidencia no significa falsedad.

Referencias

[1] ACL Anthology / COLING, "Explainable automated fact-checking: A survey." https://aclanthology.org/2020.coling-main.474/, 2020.

[2] NeurIPS Datasets and Benchmarks / arXiv, "Averitec: A dataset for real-world claim verification with evidence from the web." https://arxiv.org/abs/2305.13117, 2023.

[3] arXiv / EMNLP, "Enabling large language models to generate text with citations." https://arxiv.org/abs/2305.14627, 2023.

[4] arXiv, "Averimatec: A dataset for automatic verification of image-text claims with evidence from the web." https://arxiv.org/abs/2505.17978, 2025.

[5] arXiv / EMNLP, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." https://arxiv.org/abs/2305.14251, 2023.

[6] ACL Anthology / Findings of ACL, "Scientific fact-checking: A survey of resources and approaches." https://aclanthology.org/2023.findings-acl.387/, 2023.

[7] Natural Language Processing Journal, "Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research." https://doi.org/10.1016/j.nlp.2024.100066, 2024.

[8] Information Processing and Management, "The state of human-centered nlp technology for fact-checking." https://doi.org/10.1016/j.ipm.2022.103219, 2022.

[9] arXiv / EMNLP, "Factuality of large language models: A survey." https://arxiv.org/abs/2402.02420, 2024.

[10] International Journal of Multimedia Information Retrieval, "A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities." https://doi.org/10.1007/s13735-023-00296-3, 2023.

[11] International Journal of Multimedia Information Retrieval, "Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias." https://doi.org/10.1007/s13735-023-00312-6, 2024.

[12] Coalition for Content Provenance and Authenticity, "C2pa technical specification." https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html, 2025.

[13] ACL Anthology / EMNLP, "Image, tell me your story! predicting the original meta-context of visual misinformation." https://doi.org/10.18653/v1/2024.emnlp-main.448, 2024.

[14] ACL Anthology / Findings of EMNLP, "Automatic evaluation of attribution by large language models." https://doi.org/10.18653/v1/2023.findings-emnlp.307, 2023.

[15] Springer / PMC, "Bert for evidence retrieval and claim verification." https://pmc.ncbi.nlm.nih.gov/articles/PMC7148011/, 2020.

[16] ACL Anthology / Findings of ACL, "Evidence retrieval is almost all you need for fact verification." https://doi.org/10.18653/v1/2024.findings-acl.551, 2024.

[17] Google for Developers, "Fact check tools api." https://developers.google.com/fact-check/tools/api/reference/rest/, 2025.

[18] Schema.org, "Claimreview." https://schema.org/ClaimReview, 2026.

[19] Knowledge-Based Systems, "Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference." https://doi.org/10.1016/j.knosys.2022.109265, 2022.

[20] arXiv, "Get your vitamin c! robust fact verification with contrastive evidence." https://arxiv.org/abs/2103.08541, 2021.

[21] ACL Anthology / EMNLP, "Minicheck: Efficient fact-checking of llms on grounding documents." https://doi.org/10.18653/v1/2024.emnlp-main.499, 2024.

[22] ACL Anthology / NAACL, "Fever: a large-scale dataset for fact extraction and verification." https://aclanthology.org/N18-1074/, 2018.

[23] ACL Anthology / Findings of EMNLP, "Hover: A dataset for many-hop fact extraction and claim verification." https://aclanthology.org/2020.findings-emnlp.309/, 2020.

[24] arXiv, "Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims." https://arxiv.org/abs/1909.03242, 2019.

[25] arXiv / Hugging Face Papers, "Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms." https://huggingface.co/papers/2505.17762, 2025.

[26] arXiv, "Halueval: A large-scale hallucination evaluation benchmark for large language models." https://arxiv.org/abs/2305.11747, 2023.

[27] arXiv / LREC, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection." https://arxiv.org/abs/1911.03854, 2019.

[28] IEEE Transactions on Artificial Intelligence, "Trumorgpt: Graph-based retrieval-augmented large language model for fact-checking." https://doi.org/10.1109/TAI.2025.3567369, 2025.

[29] ACL Anthology / Findings of NAACL, "A survey on stance detection for mis- and disinformation identification." https://aclanthology.org/2022.findings-naacl.94/, 2022.

[30] IEEE Transactions on Computational Social Systems, "A literature review on detecting, verifying, and mitigating online misinformation." https://doi.org/10.1109/TCSS.2023.3289031, 2023.

[31] ACL Anthology / FEVER Workshop, "The fact extraction and verification (fever) shared task." https://aclanthology.org/W18-5501/, 2018.

[32] arXiv / ACL, "Liar, liar pants on fire: A new benchmark dataset for fake news detection." https://arxiv.org/abs/1705.00648, 2017.

[33] ACL Anthology / EMNLP, "Fact or fiction: Verifying scientific claims." https://aclanthology.org/2020.emnlp-main.609/, 2020.

[34] ACL Anthology / NAACL Demonstrations, "Improving evidence retrieval for automated explainable fact-checking." https://aclanthology.org/2021.naacl-demos.10/, 2021.

[35] ACL Anthology / EMNLP, "Explainable automated fact-checking for public health claims." https://aclanthology.org/2020.emnlp-main.623/, 2020.

[36] ACL Anthology / ACL, "Dialfact: A benchmark for fact-checking in dialogue." https://aclanthology.org/2022.acl-long.263/, 2022.

[37] arXiv / SIGKDD Explorations, "Fake news detection on social media: A data mining perspective." https://arxiv.org/abs/1708.01967, 2017.

[38] IEEE / arXiv, "Spotfake: A multimodal framework for fake news detection." https://arxiv.org/abs/1907.04748, 2019.

[39] Adobe / Content Authenticity Initiative, "Content credentials." https://contentcredentials.org/, 2025.

[40] Social Network Analysis and Mining, "Multimodal fake news detection on social media: A survey of deep learning techniques." https://doi.org/10.1007/s13278-023-01104-w, 2023.

[41] arXiv / Hugging Face Papers, "Factify 2: A multimodal fake news and satire news dataset." https://huggingface.co/papers/2304.03897, 2023.

[42] arXiv, "From verification to amplification: Auditing reverse image search as algorithmic gatekeeping in visual misinformation fact-checking." https://arxiv.org/abs/2603.09130, 2026.

[43] arXiv, "Synthid-image: Image watermarking at internet scale." https://arxiv.org/abs/2510.09263, 2025.

[44] Google Search Central, "Fact check (claimreview) markup for search." https://developers.google.com/search/docs/appearance/structured-data/factcheck, 2025.

[45] arXiv, "Feverous: Fact extraction and verification over unstructured and structured information." https://arxiv.org/abs/2106.05707, 2021.

[46] arXiv, "Wice: Real-world entailment for claims in wikipedia." https://arxiv.org/abs/2303.01432, 2023.

[47] GitHub / OpenFactVerification, "Loki: Open-source tool for fact verification." https://github.com/Libr-AI/OpenFactVerification, 2024.

[48] arXiv, "Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms." https://arxiv.org/abs/2406.08772, 2024.

[49] ACL Anthology / ACL, "Automated justification production for claim veracity in fact checking: A survey on architectures and approaches." https://doi.org/10.18653/v1/2024.acl-long.361, 2024.

[50] AAAI, "Mrr-fv: Unlocking complex fact verification with multi-hop retrieval and reasoning." https://doi.org/10.1609/aaai.v39i24.34802, 2025.

[51] Google Fact Check Tools, "Fact check explorer." https://toolbox.google.com/factcheck/explorer, 2025.

[52] IPTC, "Iptc photo metadata standard." https://iptc.org/standards/photo-metadata/, 2024.

[53] ACL Anthology / EMNLP, "Natural logic-guided autoregressive multi-hop document retrieval for fact verification." https://doi.org/10.18653/v1/2022.emnlp-main.411, 2022.

[54] arXiv, "Tabfact: A large-scale dataset for table-based fact verification." https://arxiv.org/abs/1909.02164, 2019.

[55] ACL Anthology / ACL, "Truthfulqa: Measuring how models mimic human falsehoods." https://aclanthology.org/2022.acl-long.229/, 2022.

[56] arXiv, "Retrieval-augmented generation for large language models: A survey." https://arxiv.org/abs/2312.10997, 2023.

[57] ACL Anthology / EMNLP, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models." https://doi.org/10.18653/v1/2023.emnlp-main.557, 2023.

[58] OpenReview / arXiv, "Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output." https://arxiv.org/abs/2311.09000, 2023.

[59] ACL Anthology / Findings of EMNLP, "Claim check-worthiness detection as positive unlabelled learning." https://aclanthology.org/2020.findings-emnlp.43/, 2020.

[60] Online Social Networks and Media, "Factrank: Developing automated claim detection for dutch-language fact-checkers." https://doi.org/10.1016/j.osnem.2020.100113, 2021.

[61] Princeton NLP / arXiv, "Alce: Automatic llms citation evaluation benchmark." https://github.com/princeton-nlp/ALCE, 2023.

[62] FEVER Workshop / ACL, "Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models." https://doi.org/10.18653/v1/2024.fever-1.29, 2024.

[63] CLEF / CEUR Workshop Proceedings, "Overview of checkthat! 2020: Automatic identification and verification of claims." https://ceur-ws.org/Vol-2696/, 2020.

[64] CLEF / CEUR Workshop Proceedings, "Overview of the clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news." https://ceur-ws.org/Vol-2936/, 2021.

[65] ACL Anthology / SemEval, "Semeval-2016 task 6: Detecting stance in tweets." https://aclanthology.org/S16-1003/, 2016.

[66] ACL Anthology / ACL, "Adversarial nli: A new benchmark for natural language understanding." https://aclanthology.org/2020.acl-main.441/, 2020.

[67] arXiv, "Creak: A dataset for commonsense reasoning over entity knowledge." https://arxiv.org/abs/2109.01653, 2021.

[68] arXiv, "Wikicontradiction: Detecting self-contradiction articles on wikipedia." https://arxiv.org/abs/2111.08543, 2021.

[69] Social Network Analysis and Mining, "Deep learning for misinformation detection on online social networks: A survey and new perspectives." https://doi.org/10.1007/s13278-020-00696-x, 2020.

A. Reproducibilidad mínima y artefactos

Este anexo reúne los artefactos que sostienen las afirmaciones centrales del estudio y que permiten auditar los resultados reportados.

Tabla A.1: Artefactos probatorios y función dentro de la investigación

Fuente: elaboración propia.

Artefacto probatorio	Función en la investigación
Contrato de evidencia	Define afirmaciones, fuentes, comparaciones, señales visuales, limitaciones y soporte en evidencia como objetos revisables.
Ejecuciones de evaluación	Conservan los casos analizados, métricas, diagnósticos y comparación con el comparador.
Reportes de validación	Muestran si las salidas cumplen esquema, trazabilidad y control de afirmaciones no soportadas.
Bibliografía verificada	Vincula el argumento con literatura de fact-checking, soporte en evidencia, multimodalidad y evaluación.
Inventario visual	Registra figuras propias y capturas de literatura con atribución y función dentro del argumento.

Los artefactos conservados permiten verificar tres correspondencias: que las métricas citadas en el texto coinciden con los reportes guardados, que las conclusiones se derivan de objetos del contrato y no de impresiones generales sobre la respuesta del modelo, y que cada figura y referencia cumple una función argumental reconocible — explicar el *pipeline*, justificar la evaluación, delimitar el alcance multimodal o mostrar límites de la evidencia.

Los números centrales del expediente son: evaluación local con validación de contrato 1.0, traceability rate 1.0 y unsupported-statement rate 0.0; evaluación pública

extendida con 60 muestras, *retrieval* hit rate 1.0, decisive-source rate 0.6, traceability rate 1.0 y unsupported-statement rate 0.0.

La reproducibilidad planteada por el estudio está orientada a auditoría: no garantiza que la web futura devuelva las mismas fuentes ni que todas las dependencias externas permanezcan estables, pero permite reconstruir qué evidencia se usó, qué métricas se reportaron y qué límites se conservaron en la interpretación.

B. Categorías de veredicto y criterios de soporte

El agente no emite una etiqueta binaria de verdad o falsedad, sino una de cinco categorías de veredicto que reflejan el estado de la evidencia para cada afirmación. El veredicto se condiciona al soporte registrado en el contrato de evidencia y no a la apariencia de la respuesta, de modo que la auditabilidad antecede a la clasificación: una respuesta puede coincidir con una clase esperada y aun así ser poco inspeccionable.

Tabla B.1: Categorías de veredicto y condición de soporte

Fuente: elaboración propia.

Categoría	Condición de soporte
`supported`	Existe evidencia suficiente para los requisitos centrales del *claim*.
`contradicted`	Hay incompatibilidad clara con una fuente decisiva.
`mixed`	Partes del *claim* quedan apoyadas y otras contradichas o abiertas.
`insufficient_evidence`	Se recupera material relacionado, pero sin suficiencia para concluir.
`not_verifiable`	El *claim* no puede evaluarse con las señales disponibles.

Cada categoría se sostiene en los mismos objetos del contrato que el resto del documento. La afirmación normalizada conserva entidad, tiempo, lugar y relación factual; la evidencia se valora por su cobertura de los requisitos centrales y no por su mera pertinencia temática; la comparación *claim*-evidence integra señales de *stance* y NLI como evidencia auditada; y las señales visuales aportan contexto sin equivaler a veracidad. El grado de certeza de la respuesta pública se debilita cuando falta una fuente decisiva, de modo que ninguna afirmación sustantiva se comunica sin soporte o sin formularse explícitamente como límite.

C. Síntesis de validez y procedencia de figuras

Las amenazas a la validez del estudio se concentran en cinco puntos. Primero, auditabilidad no equivale a veracidad: un expediente bien formado puede contener evidencia insuficiente. Segundo, la evaluación pública extendida no cubre todos los dominios de desinformación. Tercero, el comparador empleado permite una comparación controlada, pero no representa todas las alternativas posibles. Cuarto, la evidencia visual depende de la disponibilidad real de imagen, metadatos y resultados externos. Quinto, la reproducibilidad queda afectada por servicios web cambiantes, aunque los artefactos locales conserven el rastro reportado.

El trabajo futuro derivado de estos límites contempla ampliar muestras y dominios, mejorar la inferencia NLI, elevar la tasa de fuentes decisivas, integrar señales visuales más robustas y medir la utilidad con revisores humanos. Cada extensión sostiene la contribución del estudio sólo si preserva fuentes, spans, comparaciones, limitaciones y *grounding* dentro del contrato de evidencia.

Las figuras tomadas de la literatura se incluyen como capturas puntuales con atribución, número de figura, fuente y página en su pie, y ejemplifican cómo el campo representa pipelines de verificación, evidencia descompuesta, generación con citas y evaluación de factualidad. Las figuras de elaboración propia —el diseño del agente, el contrato de evidencia, las matrices de evaluación y la taxonomía de fallos— constituyen aportaciones del proyecto. La Tabla C.1 resume la procedencia de las figuras reproducidas de la literatura.

Tabla C.1: Procedencia de las figuras reproducidas de la literatura

Fuente: elaboración propia a partir de las publicaciones citadas.

Contenido reproducido	Publicación de origen	Ubicación
Pipeline de fact-checking con subtarea de explicación	Explainable Automated Fact-Checking: A Survey [1]	Fig. 1, p. 2
Instancia de afirmación, preguntas y veredicto	The Automated Verification of Textual Claims (AVeriTeC) Shared Task [2]	Fig. 1, p. 1
Generación de respuestas con citas verificables	Enabling Large Language Models to Generate Text with Citations [3]	Fig. 1, p. 1

Contenido reproducido	Publicación de origen	Ubicación
Verificación de afirmaciones imagen-texto con evidencia web	AVerImaTeC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web [4]	Fig. 1, p. 2
Evaluación de factualidad por afirmaciones atómicas	FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation [5]	Fig. 1, p. 1