Diseño de agentes cognitivos con arquitectura BDI jerárquica y recuperación aumentada por generación para la representación de narrativas del conflicto armado colombiano

Sergio Antonio Sosa Valero

Director: Ing. Luis Gabriel Moreno Sandoval, PhD.

Pontificia Universidad Javeriana

Facultad de Ciencias

Pregrado en Ciencia de Datos

Bogotá D.C

2026

Resumen

El conflicto armado colombiano ha durado más de seis décadas, dejando más de nueve millones de víctimas registradas, el cual generó también un legado documental inmenso: la Comisión para el Esclarecimiento de la Verdad (CEV) custodia más de 14.000 entrevistas y testimonios, de las cuales 2.240 fueron trascritas y anonimizadas. Estas constituyen un patrimonio fundamental para la memoria histórica, pero que al mismo tiempo representan un reto enorme para quien quiera analizarlos computacionalmente. Este trabajo propone un sistema de agentes cognitivos que combina una arquitectura BDI jerárquica con generación aumentada por recuperación (RAG) para para simular las narrativas de actores del conflicto armado de la manera más fiel a el registro de cada rol.

Para construirlo, se siguió el enfoque de Design Science Research en cuatro etapas. Primero, se organizó una base de conocimiento con 50.849 fragmentos en base del corpus de la CEV, clasificados según el rol del hablante y almacenados en MongoDB Atlas como base vectorial. Luego se diseñó e implementó el agente a través de un grafo de estados que articula un LLM (GPT-4o mini), el módulo RAG y una arquitectura cognitiva BDI con dos niveles. El experimento se estructuró como un diseño factorial 2×2 con cuatro condiciones: BASE, RAG, BDI y BDI+RAG. Finalmente, la evaluación combinó tres jueces LLM automatizados que calificaron diez métricas sobre el desempeño individual de cada condición, Y por otro lado tres jueces LLM de alto nivel que reciben las respuestas de una manera ciega y evalúan cual es la más autentica en el contexto del conflicto armado, todo respaldado por pruebas no paramétricas de Friedman y Wilcoxon con corrección de Bonferroni.

Los resultados muestran que el componente RAG es el que genera más impacto individual sobre todos los roles, su presencia marcó una diferencia real frente a la condición base. El BDI de manera aislada, no logró mejoras estadísticamente significativas por su cuenta; pero cuando se integra con RAG, el sistema alcanza sus mejores puntajes. Esa ventaja es consistente a lo largo de los experimentos, aunque no llega a ser estadísticamente significativa frente a RAG de manera aislada, lo que sugiere que el aporte del BDI es complementario pero moderado en términos de calidad percibida. Asimismo, el comportamiento del sistema varía entre roles, siendo el victimario el más sensible al desbalanceo presente en el corpus de la CEV.

Tabla de contenido

Resumen 2
1. Planteamiento y Justificación Del Problema 6
2. Objetivos 9
2.1 Objetivo General 9
2.2 Objetivos Específicos 9
3. Marco teórico 10
3.1 Agentes inteligentes 10
3.1.1 Agente inteligente 10
3.1.2 Agente conversacional 11
3.1.3 Agente cognitivo 11
3.2 Arquitecturas Cognitivas. 12
3.2.1 Definición de arquitectura cognitiva 12
3.2.2 Clasificación de arquitecturas cognitivas 12
3.2.3 Arquitectura BDI 13
3.3 Modelado Computacional de la Personalidad 14
3.3.1 Personalidad como conjunto de parámetros 14
3.3.2 MBTI como modelo de rasgos cognitivos 15
3.3.3 Eneagrama como modelo de motivación 15
3.4 Procesamiento de Lenguaje Natural 16
3.4.1 Tokenización y N-gramas 16
3.4.2 TF-IDF 17
3.4.3 Reconocimiento de entidades nombradas (NER) 18
3.4.4 Word Embeddings 18
3.4.5 Transformers y mecanismo de atención 19
3.5 Modelos de Lenguaje de Gran Escala 20
3.5.1 Modelos de lenguaje como distribución de probabilidad. 20
3.5.2 De los Transformers a los LLMs 21
3.5.3 Mecanismos de generación de texto 21
3.5.4 Aprendizaje en contexto (In-context learning) 22
3.6 Generación Aumentada por Recuperación (RAG) 23
3.6.1 Limitaciones de la memoria paramétrica 23
3.6.2 Definición y arquitectura RAG 24
3.6.3 Embedding semántico de fragmentos 24
3.6.4 Similitud coseno y recuperación top-k 25

3.7 Inferencia Estadística No Paramétrica 26
3.7.1 Pruebas no paramétricas para muestras relacionadas 26
3.7.2 Prueba de Friedman 26
3.7.3 Prueba de Wilcoxon de rangos con signo 27
3.7.4 Correlación biserial por rangos 28
4. Estado del arte 29
4.1 Agentes Conversacionales y modelos de lenguaje 29
4.2 Arquitecturas RAG 30
4.3 Arquitecturas BDI 32
4.4 Simulación de roles y personajes en sistemas conversacionales 33
4.5 Evaluación de agentes conversacionales 34
4.6 Vacíos identificados en la literatura 36
5. Metodología 37
5.1 Recolección de datos y preprocesamiento de texto 37
5.2 Diseño e implementación del artefacto 40
5.2.1 Interfaz de usuario y capa de API 40
5.2.2 Flujo LangGraph 41
5.2.3 Capa BDI de personalidad jerárquica 43
5.2.4 Capa RAG y Memoria a largo plazo 46
5.3 Evaluación del Artefacto 47
5.3.1 Diseño Experimental 47
5.3.2 Enfoque de evaluación: LLM-as-a-judge 48
5.3.3 Métricas de evaluación 48
5.3.4 Diseño estadístico: Evaluación con múltiples jueces. 53
6. Resultados 56
6.1 Resultados generales del artefacto 56
6.2 Desempeño componente RAG por rol 57
6.3 Desempeño componente BDI 59
6.4 Evaluación mediante múltiples jueces LLM 60
6.5 Análisis estadístico 62
6.5.1 Test de Friedman 62
6.5.2 Prueba de Wilcoxon de rangos con signo para muestras relacionadas 63
7. Discusión y conclusiones 64
7.1 Efecto de los componentes en la calidad del sistema 64
7.2 Comportamiento diferenciado por rol 66
7.4 Trabajo futuro. 67

Anexos 74

Tabla 1 Estado del arte de LLMs usados en agentes conversacionales 29
Tabla 2 Estado del arte arquitecturas RAG 31
Tabla 3 Estado del arte Arquitecturas cognitivas 32
Tabla 4 Estado del arte simulación de roles en sistemas conversacionales 33
Tabla 5 Estado del Arte: Evaluación de agentes 35
Tabla 6 Análisis descriptivo y prueba de Shapiro-Wilk por Juez y condición 53
Tabla 7 Resultados generales del artefacto 57
Tabla 8 Resultados RAG para Victima 58
Tabla 9 Resultados RAG para Victimario 58
Tabla 10 Resultados RAG para Tercero 59
Tabla 11 Métricas de evaluación BDI 60
Tabla 12 Promedio de evaluación múltiples jueces 61
Tabla 13 Resultados del test de Friedman 62
Tabla 14 Resultados Prueba de Wilcoxon de rangos con signo para muestras relacionadas 63

Ilustración 1 Metodología propuesta 37
Ilustración 2 Pipeline de preprocesamiento y clasificación 39
Ilustración 3 Diagrama de arquitectura del Sistema 40
Ilustración 4 Interfaz gráfica de la aplicación 41
Ilustración 5 Flujo del grafo de estados 42
Ilustración 6 Arquitectura BDI Jerárquica 43
Ilustración 7 Micro-BDI - Flujo por turno 45
Ilustración 8 Distribución de scores por condición y juez 54
Ilustración 9 Comparativa respuestas por condición y Justificación LLMs 60
Ilustración 10 Promedio de los 3 jueces por rol y condición experimental 62

1. Planteamiento y Justificación Del Problema

El conflicto armado en Colombia ha representado uno de los episodios de violencia política más prolongados y complejos de América latica, con una duración de más de seis décadas de enfrentamientos armados los cuales según el (Centro Nacional de Memoria Histórica, 2013). Dejaron más de 220.000 muertos entre 1958 y 2012. El 80% de ellos civiles, más de 9 millones de víctimas registradas y cerca de 8 millones de personas desplazadas forzosamente de sus territorios.

Esta guerra la cual involucros múltiples actores cuyas lógicas de violencia, aun se intentan entender, finaliza el día 24 de diciembre de 2016, mediante la firma del acuerdo Nacional para la terminación del conflicto, la construcción de una paz estable y duradera entre el gobierno nacional junto con las FARC-EP. Este acuerdo estableció que el restablecimiento de las victimas debía estar en el centro de su implementación y para ello se creó el Sistema Integral de Verdad, Justicia, Reparación y No Repetición con el fin de lograr la mayor satisfacción posible de los derechos de las víctimas y contribuir a la convivencia, la reconciliación y la no repetición del conflicto.

En ese marco Institucional nació la Comisión para el esclarecimiento de la verdad, la convivencia y la no repetición (CEV) La cual no se limitó a documentar lo ocurrido si no a sentar bases para que esto no vuelva a suceder, como resultado de su trabajo se generó un gran conjunto de entrevistas, testimonios y documentos realizados con quienes han sido testigos o víctimas del conflicto armado, los cuales son fundamentales para la construcción de una memoria histórica, pero también para el estudio académico del conflicto. Estas entrevistas a las víctimas aportan datos, describen sucesos, entornos, dinámicas políticas y sociales que permiten caracterizar los impactos de las acciones violentas, Las cuales proporcionan elementos para interpretar los motivos y las lógicas del conflicto (CNMH, 2013).

Con el propósito de que este análisis trascendiera a el ámbito institucional y pudiera ser consultado, analizado y apropiado por la sociedad colombiana, La CEV consolido un corpus generado por 14.971 testimonios entre entrevistas individuales, colectivas e historias de vida, algunas con sus respectivas transcripciones. Además de 1236 informes entregados por organizaciones sociales, ONG y entidades del estado. La CEV decidió que toda esta información debía ser de carácter público, implementando así una un archivo digital accesible desde cualquier parte del mundo. Esta apertura convierte a este corpus

en un objeto de estudio de alto valor académico y un desafío metodológico significativo para cualquier sistema computacional que intente procesarlo, simularlo o analizarlo de forma automática.

Pero el estudio manual de este tipo de corpus tiene muchas complicaciones, tales como la complejidad narrativa y semántica, la existencia de ambigüedades y matices discursivos, los silencios y todo esto agregado a la gran cantidad de información que existe. Para los cuales los métodos convencionales para procesar lenguaje natural han revelado restricciones en su capacidad de captar apropiadamente estos elementos, particularmente en contextos amplios, con relaciones discursivas complicadas y en áreas sensibles.

En los últimos años, los modelos de lenguaje de gran escala (LLM) han demostrado capacidades avanzadas en la comprensión y generación de texto, lo que abre nuevas posibilidades para el análisis automatizado de este tipo de narrativas. No obstante, su aplicación directa en dominios sensibles como el conflicto armado colombiano plantea desafíos importantes como la falta de control sobre las fuentes utilizadas, la presencia de sesgos inherentes al modelo y la necesidad de mantener fidelidad a el registro léxico y discursivo de los testimonios reales. Estos modelos sin acceso a fuentes documentales especificas tendera a generar analizar el vocabulario de manera genérica, alejado del lenguaje contextualizado que es inherente a los testimonios del conflicto armado colombiano.

En este escenario, las arquitecturas RAG (Retrieval Augmented Generation) emergen como una alternativa que permite fundamentar las respuestas generadas en un corpus documental específico, garantizando que el vocabulario, expresiones y el registro oral utilizado sea coherentes con el que sería utilizado por personas involucradas en el conflicto armado. En el presente trabajo, RAG no cumple con su uso típico de verificación factual, si no que se utiliza para dotar a el agente con este lenguaje situado y coherente con su contexto que caracteriza los testimonios de la CEV. Sin embargo, RAG no resuelve problemas de coherencia narrativa o psicológica del agente. Este no controla que el agente lo interprete y exprese de forma coherente con el rol que simula. Para esto se requiere una arquitectura que permita simular la cognición humana, el modelo BDI (Belief, Desire, Intention) permite definir un objetivo narrativo general que guie toda la conversación y micro intenciones que determinan como responde el agente turno a turno asegurando coherencia con lo que quiere decir y como lo dice

Desde el punto de vista académico, el proyecto resulta relevante porque propone integrar herramientas avanzadas de inteligencia artificial al estudio de la memoria histórica y el análisis discursivo, contribuyendo a la intersección entre tecnología y estudios sociales. En el ámbito técnico, una implementación concreta de modelos generativos aplicados a este dominio sensible. Demostrando la viabilidad de combinar autenticidad léxica, coherencia narrativa y anclaje en fuentes documentales reales en un mismo sistema conversacional. La combinación de RAG y BDI en este contexto representa además una contribución metodológica al campo, dado que no existe literatura previa que integre ambos componentes para la simulación de testimonios de conflicto armado.

Por último, el proyecto admite que el manejo computacional de testimonios relacionados con experiencias de violencia, memoria y reparación supone una responsabilidad importante desde un punto de vista ético y social. Por lo tanto, no tiene como objetivo sustituir el análisis humano, sino plantear una infraestructura controlada, examinar su rendimiento en un ámbito delicado e identificar sus puntos fuertes y débiles. De este modo, ayuda a emplear la inteligencia artificial de manera más reflexiva y responsable en contextos sociales complejos.

2. Objetivos

2.1 Objetivo General

Desarrollar un sistema de agentes cognitivos que integre una arquitectura BDI jerárquica y un componente RAG basado en testimonios reales dados por la CEV. Capaz de simular las narrativas propias de los actores involucrados en el conflicto armado colombiano preservando la autenticidad léxica y coherencia narrativa propias de cada rol.

2.2 Objetivos Específicos

- Analizar el estado del arte sobre el uso de modelos de lenguaje de gran escala, arquitecturas RAG y modelos BDI en agentes cognitivos, identificando referentes conceptuales y técnicos que orienten las decisiones de diseño del sistema propuesto.
- Conformar un corpus de testimonios del conflicto armado colombiano mediante técnicas de procesamiento de lenguaje natural a partir de entrevistas recolectadas por la CEV, clasificándolo por rol del actor para su uso como base de conocimiento en la arquitectura RAG.
- Desarrollar agentes cognitivos que integren LLMs, un componente RAG basado en el corpus CEV y una arquitectura BDI que oriente el comportamiento narrativo y táctico de cada agente según su rol.
- Evaluar comparativamente el desempeño de las cuatro condiciones experimentales (BASE, BDI, RAG Y BDI + RAG) mediante métricas por componente y un sistema de evaluación con múltiples jueces LLM, determinando la condición que produce respuestas más auténticas como testimonios reales del conflicto armado colombiano.

3. Marco teórico

En presente marco teórico proporciona las bases conceptuales necesarias para comprender el diseño y la implementación del sistema propuesto. El cual esta divido en bloques temáticos que van de lo general a lo especifico. Principios y definición de los agentes inteligentes, la arquitectura BDI como modelo cognitivo, el modelado computacional de la personalidad, fundamentos de los Llama y arquitecturas RAG, pruebas estadísticas no paramétricas empleadas en la evaluación.

3.1 Agentes inteligentes

En la inteligencia artificial se ha estudiado la construcción de sistemas los cuales puedan percibir su entorno y actuar sobre el de una manera racional. Dentro de este campo los agentes inteligentes es uno de los conceptos fundamentales, sirviendo como punto de partida para desarrollar sistemas capaces de tomar decisiones e interactuar con su entorno

3.1.1 Agente inteligente

Una descripción sencilla de un agente inteligente es pensarlo como una entidad que es capaz de percibir lo que ocurre en su entorno, razona y a partir de esto es capaz de tomar decisiones. Russell y Norvig (2020) nos dan esta idea de una idea más formal, definiendo a un agente como cualquier mecanismo que perciba su entorno a traves de sensores y actúa sobre su entorno mediante actuadores, esto no significa inteligencia pura para saberlo todo, si no que a partir de lo que sabe busca la acción que maximice su desempeño.

Wooldridge y Jennings (1995) plantean cuatro propiedades con las cuales debería contar un agente para considerarse genuinamente autónomo:

- Reactividad: Es la capacidad de adaptarse a los cambios del entorno.
- Proactividad: No solo reaccionar a estímulos dados, si no que tomar iniciativa para alcanzar metas.
- Sociabilidad: La capacidad de interactuar una persona u otros agentes.

- Autonomía: Control sobre sus propios estados y decisiones sin una intervención externa frecuente.

3.1.2 Agente conversacional

Esta es una variante de un agente inteligente, donde su entorno de percepción y acción es el lenguaje natural. Lo que recibe son instrucciones o enunciados en texto o voz y su respuesta es por el mismo medio. La diferencia con un agente que controla un robot o ejecuta tareas sistemáticamente, el agente conversacional existe y actúa únicamente a traves del lenguaje (Shum et al., 2018).

Este agente se distingue de sistemas basados en reglas es que este cuenta con un estado interno dinámico que va evolucionando a lo largo de la interacción, alimentándose de una historial de la conversación completa guardando los enunciados del interlocutor junto a las respuestas dadas por el agente, utilizar este estado interno es lo que le ayuda a el agente a mantener la coherencia temática las referencias utilizadas y respondiendo comunicativamente adecuada de acuerdo a su contexto.

3.1.3 Agente cognitivo

A partir de los agentes conversacionales, se han extendido sus capacidades para introducir a los agentes cognitivos, los cuales se basan en la simulación de la cognición humana mediante procesos computacionales, como lo pueden ser la percepción, el razonamiento, la memoria y el aprendizaje, los cuales les permiten interpretar el contexto, aprender, planificar acciones y adaptarse a cambios basados en su entorno para tener comportamientos más sofisticados y apropiados en función de sus objetivos.

Esta clase de agentes se sustenta teóricamente en la hipótesis del sistema de símbolos físicos propuesto por Newell y Simon (1976), que definen que, si un sistema es capaz de manipular estructuras simbólicas, posee las condiciones necesarias y suficientes para mostrar un comportamiento inteligente general. Esto implica que este agente no solo reacciona a el turno inmediato si no que mantiene creencias sobre el estado de la conversación, persigue objetivos a lo largo del tiempo y ajusta sus planes si el contexto cambia.

3.2 Arquitecturas Cognitivas.

Las arquitecturas cognitivas son las encargadas de organizar la forma en que el agente organiza su racionamiento interno, estableciendo los procesos mediante el cual este agente mantiene su estado, como forma sus creencias, establece sus metas y decide como actuar. En esta sección se define el concepto de una arquitectura cognitiva, se muestran las principales familias existentes y profundiza en la arquitectura BDI como modelo de razonamiento.

3.2.1 Definición de arquitectura cognitiva

Una arquitectura cognitiva puede ser vista como una aproximación general del funcionamiento de la cognición humana, donde se define de manera fija los mecanismos de representación del conocimiento, los procesos de razonamiento y las formas en que el sistema interactúa con su entorno. (Newell, 1990).

Para Anderson et al. (2004) una arquitectura cognitiva debe contar con 3 aspectos fundamentales: como se almacena y se recupera el conocimiento, como ser toman decisiones a partir de este conocimiento, como se toman decisiones en función de este conocimiento, como este agente aprende y actualiza sus representaciones a partir de la experiencia. A partir de estos aspectos se determina que tan sofisticado es el comportamiento que el agente puede exhibir y marca las diferencias entre las diferentes familias arquitectónicas.

3.2.2 Clasificación de arquitecturas cognitivas

En la literatura se pueden ver grandes tipos de arquitecturas cognitivas, esto dependiendo de la forma de representación interna que utilizan (Langley et al., 2009).

Las primeras son las arquitecturas simbólicas, donde representa el conocimiento mediante estructuras explicitas y razonan sobre ella mediante mecanismos de inferencia, planificación y manipulación simbólica. La ventaja que presenta es la transparencia e interpretabilidad del razonamiento, pero dependen mucho de representaciones formalizadas que en entornos ambiguos o dinámicos podrían no ser tan robustas.

Las arquitecturas conexionistas representan el conocimiento de forma distribuida utilizando los pesos de una red neuronal y aprenden mediante ajuste estadista sobre datos (Rumelhart & McClelland, 1986). Esta es una buena opción en entornos dinámicos o con alta incertidumbre, pero dado a que ya el conocimiento no está en formas explicitas si no en parámetros numéricos su razonamiento interno no es tan fácil de inferir o interpretar (Lipton, 2018).

Y las arquitecturas hibridas donde combinan estos dos enfoques, donde el componente simbólico es utilizado para tareas que requieren razonamiento explícito, planificación y manejo de objetivos, por su parte el componente conexionista se encarga de procesar información compleja o ambigua como patrones, lenguaje o señales perceptuales (Sun, 2006).

3.2.3 Arquitectura BDI

Dentro de las arquitecturas simbólicas, el modelo BDI es el más consolidado en la implementación de agentes deliberativos autónomos (Rao & Georgeff, 1995). Esta inicia con la teoría filosófica de la acción intencional de Bratman (1987) donde plantea que un comportamiento racional en un agente no puede ser explicado únicamente por sus deseos, sino que también necesita un componente de intenciones, que guía a el agente en su actuar y ayuda a su planificación futura.

Esta idea fue formalizada en términos computacionales por Rao y Georgeff (1995) donde definen tres tipos de estados mentales: las creencias (B), representa el conocimiento que el agente tiene sobre su entorno; los deseos (D) son la representación de sus metas, lo que el agente quiere alcanzar; Y las intenciones(I), que son los planes que el agente ejecuta para alcanzar sus deseos. El estado mental del agente en un instante t se puede definir de la siguiente manera:

$$\sigma_t = \langle B_t, D_t, I_t \rangle$$

La distinción entre deseos e intenciones no sigue un proceso determinista, si no que Rao y Georgeff (1995) modelaron estos estados mentales del agente sobre una estructura temporal Arborea, en la que no existe un único futuro, en cambio existen un conjunto de cambios posibles que se crean a partir del estado actual, donde siendo φ una variable proposicional un deseo es formalmente una de

esas ramas donde existe al menos un camino futuro donde se cumple φ y una intención exige que φ se cumpla en todo futuro compatible con el compromiso adoptado del agente.

El agente entra en un ciclo de razonamiento para hacer transiciones entre estados, esto se hace mediante tres funciones centrales que fueron formalizadas por Wooldridge (2009).

$$B_{t+1} = \text{brf}(B_t, \rho_t),$$

$$D_{t+1} = \text{options}(B_{t+1}, I_t)$$

$$I_{t+1} = \text{filter}(B_{t+1}, D_{t+1}, I_{t)}$$

Donde ρ_t es la percepción su entorno en el tiempo t, la función brf (Belief revisión function) actualiza las creencias, options genera los deseos posibles; y fillter selecciona los adecuados según su factibilidad y prioridad para su intención.

3.3 Modelado Computacional de la Personalidad

La arquitectura cognitiva no es lo único de lo que depende el comportamiento de un agente, estos tienen parámetros internos que varían la forma en que dos agentes cognitivos se comportan, incluso si tienen la misma forma de representar su cognición. Esta sección define los modelos de personalidad que permiten representar estas diferencias de manera estructurada

3.3.1 Personalidad como conjunto de parámetros

La personalidad vista desde un punto psicológico, se define como el conjunto de patrones estables de pensamiento, emoción y comportamiento que caracterizan a un individuo y lo distinguen de otro en situaciones similares (McCrae & Costa, 1999). Existe una propiedad de estabilidad que es de importancia para los sistemas computacionales, que dice que esta personalidad no cambia a traves de los turnos si no que es una disposición de fondo que orienta el comportamiento a lo largo de su uso.

Computacionalmente esta estabilidad puede ser modelada como un vector de parámetros fijos que orienten las decisiones del agente en cada ciclo de razonamiento. Mairesse et al. (2007) demostraron de una manera empírica que los rasgos de personalidad influyen en los rasgos lingüísticos, como lo puede ser su elección del vocabulario, nivel de formalidad, carga emocional de su vocabulario, entre

otros . Lo que hace pertinente el uso de modelos de personalidad al momento de controlar el estrilo comunicativo de un agente conversacional.

3.3.2 MBTI como modelo de rasgos cognitivos

El Myers-Briggs Type Indicator (MBTI) es un modelo de personalidad desarrollado por Isabel Briggs Myers y Katherine Cook Briggs, las cuales se basaron en la teoría de tipos psicológicos de Carl Jung (Myers & McCaulley, 1985). Donde modela la personalidad en cuatro dimensiones bipolares independientes, cada una describiendo un patrón de comportamiento o toma de daciones:

- Extraversión / Introversión: define la fuente de energía y su interacción con el mundo interior. La extraversión es una preferencia por la interacción con el mundo exterior, mientras que la introversión se centra en la interacción con su mundo interno
- Sensación / Intuición: Esta describe la forma en que se percibe o se procesa la información, la sensación se basa en la información concreta y basada en hechos, mientras que la intuición se basa en patrones, posibilidades y el futuro.
- Pensamiento / Sentimiento: Esta refleja la forma en la toma de decisiones, el pensamiento son decisiones basadas en lógica y coherencia , mientras que el sentimiento son decisiones basadas en valores preestablecidos e impacto emocional en otros.
- Juicio / Percepción : Describe la forma de organización y relacionarse con su entorno. Juicio se centra en decisiones junto con planes estructurados, mientras que la percepción son más flexibles o espontáneos.

3.3.3 Eneagrama como modelo de motivación

El modelo de personalidad de Eneagrama organiza los patrones motivacionales en nueve tipos básicas, cada uno orientado por una motivación central, un patrón emocional y una forma característica de percibir su entorno (Riso & Hudson, 1999). Es utilizado para describir la razón del comportamiento,

es decir las motivaciones de fondo en las que basa sus decisiones. Definiendo así sus 9 tipos del eneagrama:

- El Perfeccionista: Busca la integridad, el orden y la mejora constante
- El Ayudador: Busca ser valorado a traves del apoyo emocional y la ayuda a otros
- El Triunfador: Su motivación es el éxito, ser admirado y ser eficiente.
- El Individualista: Se enfoca en la autenticidad, expresar sus emociones profundamente
- El Investigador: Su motivación es el conocimiento y la autonomía para comprender su entorno.
- El Leal: Orientado a la seguridad, prever riesgos y la confianza en personas.
- El Entusiasta: Busca el placer mediante experiencias nuevas, para evita el malestar emocional.
- El Desafiador: Motivado por el control, la autonomía y la protección a si mismo
- El Pacificador: Busca armonía, estabilidad y evitar el conflicto.

3.4 Procesamiento de Lenguaje Natural

Cambria y White (2014) definen el PLN como una gama de técnicas computacionales motivadas por la teoría para el análisis, interpretación, comprensión, generación y representación automática del lenguaje humano en una máquina o computadora. Esta sección define las técnicas de PLN relevantes para la construcción de sistemas que operan sobre el lenguaje natural.

3.4.1 Tokenización y N-gramas

La tokenización el proceso fundamental que divide un texto continuo, en unidades más pequeñas, que son llamadas tokens, que pueden ser palabras, frases o caracteres, esto dependiendo de la granularidad que sea deseada (Jurafsky & Martin, 2023). Este es el encargado de establecer las unidades de trabajo para las técnicas posteriores en un pipeline de PLN.

Formalmente se define como una función que mapea el texto a una secuencia unidades:

$$\text{tokenizar}(t) = \{tok_1, tok_2, \dots, tok_n\}$$

Donde t es el texto de entrada y cada tok_i es un token individual. A partir de estos se pueden generar n-gramas los cuales se definen como la secuencia de n tokens, los cual proporciona una forma muy simple de analizar texto, observando secuencias cortas de palabras que aparecen juntas, lo que permite modelar dependencias sintácticas y semánticas a corta distancia (Cavnar & Trenkle, 1994).

Matemáticamente se define como el conjunto de n-gramas de orden n para una secuencia de tokens se define así:

$$N_n(tok) = \{(tok_i, tok_{i+1}, \dots, tok_{i+n-1}) \mid i = 1,2, \dots, |tok| - n + 1\}$$

3.4.2 TF-IDF

Esta técnica es una ponderación estadística que mide la importancia de cada palabra dentro de un documento en relación con la frecuencia de aparición en conjunto completo de documentos o corpus, Salton & McGill, (1983) muestra una forma sencilla de entenderla. Un término es más informativo si aparece frecuentemente en un documento especifico, pero raras veces en el resto del corpus. Basándose en esta idea el cálculo se divide en dos partes. La term frecuency (TF) mide la frecuencia de aparición de un término t en un documento d siguiendo la formula:

$$\mathrm{TF}(t, d) = \frac{\mathrm{count}(t, d)}{||d||}$$

Siendo $count(t, d)$ el conteo de las apariciones del termino t en el documento d y $||d||$ es el numero total de términos en el documento.

Por otro lado, la inverse document frecuency (IDF) mide que tan único es el término a lo largo de todo el conjunto de documentos y planteándose de la siguiente forma:

$$\mathrm{IDF}(t, C) = \log\left(\frac{||C||}{|\{d \in C : t \in d\}|}\right)$$

Siendo $||C||$ el numero total de documentos en el corpus y $|\{d \in C : t \in d\}|$ es el numero de documentos que contiene el termino t

Y el puntaje final es el producto de ambos componentes.

$$\text{TF-IDF}(t, d, C) = \text{TF}(t, d) \cdot \text{IDF}(t, C)$$

3.4.3 Reconocimiento de entidades nombradas (NER)

El reconocimiento de entidades nombradas se define como la tarea de identificar y clasificar fragmentos de texto que refieren a entidades del mundo real como lo pueden ser personas, lugares, organizaciones, etc. En categorías semánticas predefinidas (Nadeau & Sekine, 2007). A nivel conceptual, el NER resuelve el problema de extraer elementos específicos del texto que tienen relevancia para la interpretación semántica. Y esta ayuda en el PLN ya que permite estructurar la información del texto y facilitar tareas posteriores.

Formalmente, el NER puede verse como un problema de etiquetado de secuencias donde se asigna a cada token una etiqueta de clase de entidad:

$$\text{NER}(tok) = \{(tok_1, e_1), (tok_2, e_2), \dots, (tok_n, e_n)\}$$

Donde $e_i \in$ E donde E es el conjunto de clases de entidades posibles.

3.4.4 Word Embeddings

Un embedding se define como una representación vectorial en un espacio multidimensional donde la proximidad geométrica entre vectores aproxima la similitud semántica entre palabras (Bengio et al., 2003), estos embeddings permiten capturar relaciones semánticas de tal manera que los algoritmos de aprendizaje automático puedan procesar efectivamente.

La idea en que se basa en la idea de la hipótesis distribucional, donde se plantea que palabras con significados similares, aparecen en contextos similares (Harris,1954). Si se mira una palabra junto con las palabras que normalmente aparecen cerca de ella, se puede entender su contexto, entonces dos palabras con contextos similares tendrán vectores cercanos en el espacio de embeddings.

Formalmente, un embedding es una función que mapea palabras a vectores, de la siguiente forma.

$$\text{emb}: V \rightarrow R^d$$

Donde V es el vocabulario y d es la dimensión del espacio de embeddings. La similitud entre dos palabras w_1 y w_2 mide comúnmente con la similitud del coseno definida de la siguiente forma:

$$\text{sim}(w_1,\text{w_2}) = \frac{\text{emb}(w_1) \cdot \text{emb}(w_2)}{||\text{emb}(w_1)|| \cdot ||\text{emb}(w_2)||}$$

Teniendo valores en el rango de $[-1,1]$ donde 1 representa la máxima similitud

3.4.5 Transformers y mecanismo de atención

Los Transformers representan un cambio paradigmático en PLN al introducir el mecanismo de atención como alternativa al procesamiento secuencial de textos (Vaswani et al., 2017). Arquitecturas anteriores hacían su procesamiento token a token de manera secuencial. Los transformes permiten que cada token preste atención a todos los demás tokens de la secuencia sin importar la distancia entre sí. Lo cual hace que el modelo pueda aprender mejores representaciones contextuales, donde el significado de cada palabra se define en relación con toda la secuencia procesada no solo con las palabras más cercanas. Este mecanismo de atención se define mediante la formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Donde Q (queries), K(keys) y V (values) representan el texto de entrada mediante matrices, la matriz QK^T contiene puntuaciones de similitud entre cada par de posiciones en la secuencia, que se normalizan usando el algoritmo de softmax y se utilizan para ponderar los vectores de los valores. El factor $\sqrt{d_k}$ previene que los productos punto se saturen.

Esta capacidad de representación contextual es la base arquitectónica sobre el que se desarrollan los LLM que se desarrollan en la sección 3.5

3.5 Modelos de Lenguaje de Gran Escala

Los LLMs son sistemas de inteligencia artificial los cuales pasaron por un proceso de entrenamiento sobre volúmenes masivos de texto con el objetivo de aprender distribuciones de probabilidad sobre secuencias de lenguaje natural. Para los agentes conversacionales actuales son de gran importancia debido a su capacidad para la generación de texto de manera coherente, seguir instrucciones y la adaptabilidad de su comportamiento en contextos específicos. La siguiente sección definirá conceptos para entender su funcionamiento en el presente trabajo.

3.5.1 Modelos de lenguaje como distribución de probabilidad.

Estos modelos son formalmente definidos como una distribución de probabilidad sobre secuencias de tokens. Dado un vocabulario V, el modelo asigna una probabilidad P a cada posible secuencia de tokens $w_1, w_2, \dots, w_n$, Pero calcular esta probabilidad conjunta no es lo mas viable para secuencias de texto largas, por lo que se aplica la regla de la cadena para factorizarla como un producto de probabilidades condicionales

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \dots, w_{i-1})$$

Donde cada token se predice condicionado a los tokens generados previamente. Para predecir estas probabilidades condicionales, el modelo utiliza una función parametrizada θ donde θ representa el conjunto de parámetros del modelo que determina como se hacen las predicciones

$$P_\theta(w_i \mid w_1, w_2, \dots, w_{i-1})$$

El entrenamiento del modelo consiste en ajustar los parámetros θ para maximizar la probabilidad de generar las secuencias dadas en el corpus de entrenamiento lo que equivale a minimizar la entropía cruzada entre la distribución real del lenguaje y la distribución aprendida por el modelo (Jurafsky & Martin, 2023).

3.5.2 De los Transformers a los LLMs

La arquitectura de transformes establecida en la sección 3.4.5 estableció el mecanismo de atención que permite procesar secuencias completas de texto capturando dependencias a largo alcance. En estos se centran los LLMs, pero entrenados a una cantidad muy grande de datos y parámetros.

Para que un modelo de lenguaje sea determinado grande no solo basta con el número de parámetros, si no que depende de tres factores, los cuales Kaplan, et al. (2020) establece en las llamadas leyes de escala: el número de parámetros del modelo N, el tamaño del corpus de entrenamiento D medido en tokens y el computo utilizado C medido en operaciones de punto flotante. También demostraron empíricamente que se puede predecir la mejora en el rendimiento del modelo al aumentar cualquiera de estos tres factores, siguiendo una relación de ley de potencia.

$$L(N) \propto N^{-\alpha}$$

Donde L es la perdida el modelo y α es un exponente empírico que depende del componente escalado. Esto justifico el desarrollo de modelos cada vez más grandes, que ayudo a la consolidación del paradigma actual.

El proceso de creación de un LLM sigue dos etapas. El preentrenamiento, que es donde el modelo aprende representaciones generales del lenguaje a partir de corpus de datos muy grandes mediante el objetivo autorregresivo planteado en la sección 3.5.1. Y el siguiente paso es el fine-tuning, donde el modelo reentrenado es adapta do para hacer tareas muy específicas o funcionar en dominios muy específicos con conjuntos de datos más pequeños pero especializados (Brown et al., 2020).

3.5.3 Mecanismos de generación de texto

Después de este proceso de entrenamiento, el modelo produce una distribución de probabilidad sobre todo el vocabulario para ejecutar su proceso de generación de texto, seleccionando token a token a partir de esta distribución. Esta distribución se obtiene asignando el puntaje (logit) a cada token posible y luego se aplica la función softmax los convierte en probabilidades siguiendo la formula:

$$P(w_i) = \frac{e^{z_i}}{\sum_{j \in V} e^{z_i}}$$

Donde también se introduce un parámetro de temperatura τ el cual modifica esta distribución antes de aplicar softmax controlando la dispersión de las probabilidades.

$$P(w_i) = \frac{e^{z_i/\tau}}{\sum_{j \in V} e^{z_i/\tau}}$$

Cuando τ es un valor cercano a 0, hace que la distribución se concentre en el token más probable, haciendo que su generación sea más determinista o repetitiva, y por otro lado cuando $\tau \to \infty$ hace que la distribución se aplane y sea uniforme, produciendo texto aleatorio que pierde coherencia.

3.5.4 Aprendizaje en contexto (In-context learning)

Brown et al. (2020) define una de las capacidades más relevantes de los LLMs llamada el aprendizaje en contexto, que él explica como la posibilidad del modelo de adaptar su comportamiento en una tarea especifica, únicamente a partir de instrucciones o ejemplos que se incluyan en el prompt, esto sin la necesidad de modificar los parámetros mediante reentrenamiento.

Esta capacidad se puede observar de tres maneras, dependiendo de la cantidad de ejemplos proporcionados a el modelo. Zero shot es cuando el modelo solo recibe una instrucción y genera una respuesta sin haber tenido ningún ejemplo previo de esta tarea. One shot es cuando se proporciona un único ejemplo de entrada y salida antes de la consulta y few shot es donde se proporcionan varios ejemplos para usar como referencia para inferir el patrón esperado (Brown et al., 2020).

Este aprendizaje en contexto se puede expresar como una distribución de probabilidad condicional donde el prompt p actúa como contexto en que condiciona la generación de texto.

$$P_\theta(r \mid p) = \prod_{i=1}^{|r|} P_\theta(r_i \mid p, r_1, \ldots, r_{i-1})$$

Donde r es la respuesta generada y p contiene tanto las instrucciones de la tarea como los ejemplos proporcionados. Esto el modelo no lo usa para ajustar sus parámetros, sino que los usa como contexto para sesgar su distribución de salida hacia el comportamiento deseado.

3.6 Generación Aumentada por Recuperación (RAG)

Los LLMs tal como se plantearon en la sección 3.5 poseen una limitación de responder únicamente en base a lo que aprendió en su proceso de entrenamiento. RAG enfrenta esta limitación proponiendo un componente externo de recuperación de información que complementa la memoria interna del modelo. En esta sección se definen los principales fundamentos teóricos de RAG.

3.6.1 Limitaciones de la memoria paramétrica

Tal como se dijo antes, el conocimiento que el modelo aprende durante su entrenamiento es el que queda codificado en sus parámetros θ, y estos quedan estáticos después de finalizar su proceso de entrenamiento.

Esto genera tres limitaciones que Shuster et al. (2021) identifican como problemas centrales de los LLMs para aplicaciones que requieren conocimiento especifico o actualizado.

La primera es la desactualización, donde el modelo no puede adquirir información generada después de su entrenamiento sin ser reentrenado, lo que lo limita en contextos dinámicos donde el conocimiento va evolucionando, por otro lado, describe la opacidad de la fuente, como la incapacidad del modelo de indicar de donde proviene la información que está generando, lo que dificulta la atribución del contenido generado. Y, por último, pero más importante es la alucinación que es al momento de que el modelo no tenga la información suficiente sobre un dominio especifico tiende a generar contenido coherente pero factualmente incorrecto, porque su objetivo de entrenamiento es producir texto probable, no verdadero (Ji et al., 2023).

3.6.2 Definición y arquitectura RAG

Lewis et al. (2020) proponen RAG como un paradigma que combine un componente de recuperación de información con un modelo de lenguaje generativo. Donde busca que el conocimiento relevante para la consulta sea recuperado dinámicamente de un corpus externo y este no dependa de la memoria estática del modelo.

Esta arquitectura se divide en dos fases principales, el recuperador (retriever) que, dada una consulta, identifica y extrae los fragmentos más relevantes del corpus externo y el generador, que es el LLM que recibe estos fragmentos recuperados para condicionar su respuesta a la consulta inicial. Lewis et al. (2020) definen formalmente la distribución de salida de un sistema RAG como una combinación de los documentos recuperados.

$$\mathrm{P_{RAG}(\,r \mid q\,)} = \sum_{\mathrm{d_i} \in \mathrm{R(q,k)}} \mathrm{P(\,d_i \mid q\,)} \cdot P_\theta(\,r \mid q, d_i\,)$$

Donde q representa la consulta, $R(q,k)$ es el conjunto de los k fragmentos más relevantes para el corpus $P(d_i \mid q)$ es la probabilidad de relevancia del fragmento d_i dada la consulta y $P_\theta(\,r \mid q, d_i\,)$ es la probabilidad de la respuesta r generada por el LLM condicionada a la consulta y el fragmento recuperado.

3.6.3 Embedding semántico de fragmentos

Teniendo como punto de partida la idea desarrollada en la sección 3.4.4, un embedding semántico utilizado en RAG sigue la misma idea de mapear texto a un espacio vectorial de alta dimensión donde la proximidad geométrica aproxima la similitud semántica. Pero en vez de ser con tokens separados, se toman fragmentos completos de texto (chunk) que puede abarcar una oración hasta varios párrafos. Donde una diferencia clave entre estos es el espacio vectorial donde se representan. Un embedding para un token suele ser representado en una dimensionalidad entre 100 a 300. Un embedding semántico de fragmentos suele ser representados en espacios vectoriales cuya dimensionalidad comienza comúnmente alrededor de 384 y modelos mucho más avanzados alcanzado 3584 dimensiones para capturar la riqueza semántica de fragmentos completos

El corpus para el sistema RAG se representa como el conjunto de vectores

$$\mathcal{D} = \{E(d_1), E(d_2), \dots, E(d_N)\}$$

Donde N es el número total de fragmentos en el corpus. Este conjunto constituye el índice vectorial sobre el que opera el recuperador en tiempo de inferencia.

3.6.4 Similitud coseno y recuperación top-k

El siguiente paso en el proceso, después de haber obtenido los embeddings de los chunks es la recuperación de los más relevantes dada una consulta q. Dada esta consulta se requiere comparar el vector de la consulta con los vectores de todos los chunks del corpus. Y se emplea la similitud coseno definida en la sección 3.4.4 pero en vez de ser para los embeddings de un token, en este caso es para el embedding del chunk. Entonces la recuperación consiste en encontrar los k fragmentos cuyo embedding sea el más similar a el embedding de la consulta, formalmente se define de la forma:

$$\mathrm{R(q,k)} = \arg \max_{\mathrm{D'\subseteq D,\,|D'|=k}} \sum_{\mathrm{d_i \in D'}} \mathrm{sim}\big(\mathrm{E(q)}, \mathrm{E(d_i)}\big)$$

Donde $R(q,k)$ es el conjunto de los k fragmentos mas relevantes para la consulta q, este problema se conoce como la busqueda del vecino más cercano, pero dado que una búsqueda exacta sobre todos los fragmentos, es computacionalmente inviable, se utilizan algoritmos de búsqueda aproximada a el vector más cercano, sacrificando un poco de precisión en cambio de una reducción notable en el tiempo de búsqueda (Johnson et al., 2021).

3.7 Inferencia Estadística No Paramétrica

Cuando existen datos ordinales, estos provienen de escalas de ranking o cuando el tamaño muestral es insuficiente para demostrar el supuesto de la normalidad. Una opción metodológicamente apropiada son las pruebas estadísticas no paramétricas (Conover, 1999). Esta sección define las pruebas no paramétricas empleadas para comparar múltiples condiciones experimentales sobre muestras relacionadas. experimentales sobre muestras relacionadas.

3.7.1 Pruebas no paramétricas para muestras relacionadas

Las pruebas no paramétricas son métodos de inferencia estadística que no tienen el supuesto de que los datos siguen una distribución especifica, como la distribución normal. Estas no trabajan sobre los valores originales, si no que estas pruebas utilizan rangos u ordenes relativos entre observaciones, lo que las hace más robustas frente a distribuciones asimétricas, valores atípicos y datos ordinales (Hollander et al., 2013).

Para el caso de muestras relacionadas, estas pruebas se utilizan cuando las observaciones presentan dependencia entre sí, existiendo diferentes pruebas para comparar condiciones relacionas, las cuales se desarrollarán en las siguientes secciones

3.7.2 Prueba de Friedman

Este test es una extensión no paramétrica para el análisis de varianza para medidas repetidas. La cual está diseñada para comparar k condiciones comparadas sobre n bloques relacionados. (Friedman, 1937). Donde se evalúa si, después de controlar la variabilidad entre preguntas, las distribuciones de cada variante difieren entre sí. En esta prueba estadístico de prueba se calcula de la siguiente manera

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

Donde n es el número de bloques, k es el numero de condiciones y $R_j =$ Suma de rangos de la condición j a través de los bloques Bajo la hipótesis nula de que todas las condiciones tienen la misma distribución de rangos, el estadístico χ_F^2 sigue aproximadamente una distribución chi-cuadrado con $k - 1$ grados de libertad (Friedman, 1937).

3.7.3 Prueba de Wilcoxon de rangos con signo

Cuando la prueba de Friedman rechaza la hipótesis nula de igualdad entre condiciones, se requieren de pruebas post hoc para identificar las condiciones que tienen una diferencia significativa entre sí. La prueba de rangos con signo de Wilcoxon es la prueba no paramétrica estándar para comparar dos muestras relacionadas (Wilcoxon, 1945). Esta ópera sobre las diferencias entre pares de observaciones. Teniendo dos condiciones X e Y evaluadas sobre los mismos n bloques se calculan las diferencias

$$\mathrm{d_i} = \mathrm{x_i} - \mathrm{y_i}, \quad \mathrm{i} = 1,2, \dots, \mathrm{n}$$

Estas se ordenan por su valor absoluto y sin tener en cuenta las diferencias nulas, se les asignan rangos r_i quedándose con el signo de la diferencia original. El estadístico base se define como:

$$W^+ = \sum_{i:d_i>0} R_i\,, \quad W^- = \sum_{i:d_i<0} R_i$$

Después de esto el estadístico se estandariza de la siguiente forma

$$Z = \frac{W^+ - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}}$$

que sigue aproximadamente una distribución normal estándar bajo la hipótesis nula de que la mediana de las diferencias es cero (Wilcoxon, 1945).

3.7.4 Correlación biserial por rangos

La significancia estadística de las pruebas anteriores es efectiva para determinar la existencia de diferencias. Pero no indica que tan grande es esa diferencia. Para cuantificar la magnitud entre las diferencias detectadas por la prueba de Wilcoxon, se utiliza la correlación biserial por rangos como medida del tamaño del efecto (Kerby, 2014). Donde esta se define como:

$$r_b = \frac{W^+ - W^-}{W^+ + W^-}$$

Donde W^+ y W^- son la suma de rangos de las diferencias positivas y negativas respectivamente, definidas en la sección 3.7.3. Este valor puede estar entre -1 y 1 donde valores positivos indican que la primera condición supera a la segunda, valores negativos indican lo contrario y valores cercanos a 0 indican la falta de efecto dirección

4. Estado del arte

La siguiente sección presenta el estado del arte donde se presenta una revisión sistemática de la literatura científica relevante para la selección de los componentes centrales del sistema propuesto. Donde se identificarán los avances más recientes de cada sección identificando tendencias en la investigación actual, enfoques metodológicos predominantes y limitaciones reportadas en estos trabajos.

4.1 Agentes Conversacionales y modelos de lenguaje

Los modelos de lenguaje a gran escala (LLMs) son sistemas de inteligencia artificial entrenados sobre grandes cantidades de texto que les permiten comprender patrones del lenguaje natural y generar respuestas más coherentes incluso con una mejor estructura semántica. Es interesante identificar la evolución de los agentes conversacionales a través del tiempo, donde comenzaron con sistemas basados en reglas determinísticas para responder preguntas dadas por los usuarios, hasta llegar a agentes autónomos que integran inteligencia artificial, los cuales son capaces de planificar, recordar información de interacciones anteriores y tomar decisiones.

Tabla 1 Estado del arte de LLMs usados en agentes conversacionales

Autores	Año	Tipo de agente	Modelo LLM base	Memoria o contexto	Método de evaluación	Enfoque principal
Bian et al.	2025	Agente de simulación social	No especificado	Sin memoria persistente explícita	Análisis crítico y revisión conceptual	Riesgos de idealización del comportamiento humano en simulaciones sociales con LLMs
Park et al.	2023	Agente generativo con rol	GPT-3.5 / GPT-4	Memoria reflexiva y recuperación por relevancia	Observación de comportamiento emergente	Simulación de comportamiento humano autónomo con dinámicas sociales complejas en entornos virtuales
Wang et al.	2024	Agente autónomo (revisión)	GPT-4, Llama, PaLM, entre otros	Memoria a corto y largo plazo; bases vectoriales	análisis de literatura	Marco de cuatro componentes (perfil, memoria, planificación, ejecución) para agentes LLM autónomos

La tabla 1 nos muestran artículos relevantes para esta sección, comenzando con Wang et al. (2024) que muestra en su trabajo un marco para entender estos agentes, dividiéndolo en cuatro componentes principales: un perfil que define la identidad del agente, una memoria para recordar interacciones previas, un módulo de planificación para determinar su comportamiento en turnos posteriores y un módulo que ejecuta estas decisiones. Este marco muestra como la integración de estos

modelos en arquitecturas más complejas, elevan las capacidades de estos sistemas, pasando de ser generadores de texto a desarrollar tareas que impliquen más razonamiento y adaptación.

Park et al. (2023) es un referente clave en esta evolución, donde en su trabajo desarrollaron agentes generativos con la capacidad de simular comportamientos similares a los de los humanos en un entorno virtual, los cuales organizaban eventos sociales de forma autónoma, mantenían relaciones interpersonales con otros agentes y mostraban interacciones complejas sin la necesidad de intervención humana. Demostrando que los LLMs no son simples generadores de texto y pueden ser útiles para simular dinámicas sociales. Marcando así un precedente para los sistemas que busquen simular actores humanos con personalidades y objetivos particulares.

Pero también tienen sus limitaciones, como las descritas por Bian et al. (2025), donde plantea que estas simulaciones sociales basadas en LLMs tienden a generar representaciones muy idealizadas del comportamiento humano, donde en su trabajo describe que estos agentes se comportan de manera más cooperativa, racional y armónica cosa que no necesariamente harían personas reales en tales situaciones. Lo cual es de relevancia para el proyecto ya que, al simular el contexto del conflicto armado, donde las descripciones de los roles tienden a incluir el dolor, la contradicción o la evasión, estos sistemas sin personalización pueden tender a suavizar o eliminar estos matices del discurso.

Estas limitaciones evidencian que el LLM por sí solo no es suficiente para simular actores con narrativas complejas, carga emocional y vocabulario autentico. En el contexto del proyecto donde cada actor posee una postura, vocabulario y contenido emocional distinto para cada rol, estas carencias pueden tener mayor impacto negativo, lo cual justifica la integración de componentes arquitectónicos adicionales revisados más adelante.

4.2 Arquitecturas RAG

Los LLMs tienen una limitación de no poder obtener más información externa al momento de generar sus respuestas, estos están limitados al conocimiento e información que se le fue dada a el momento de su entrenamiento, lo cual puede llevar a un riesgo de generar respuestas que suenan plausibles con el contexto, pero puede ser factualmente incorrecta o de por si inventada.

Tabla 2 Estado del arte arquitecturas RAG

Autores	Año	Tipo de sistema	Modelo LLM base	Mecanismo de recuperación	Método de evaluación	Enfoque principal
Lewis et al.	2020	RAG (propuesta inicial)	BART + DPR	FAISS con recuperación densa	Benchmarks NaturalQuestions, TriviaQA, WebQ	RAG evita alucinaciones consultando documentos externos antes de generar
Barnett et al.	2024	Análisis de fallos en RAG	Múltiples	Revisión de pipelines RAG	Estudio de casos reales	Identifica siete puntos de fallo en RAG: recuperación irrelevante y dependencia de calidad del corpus
Ling et al.	2025	RAG para especialización de dominio	Múltiples	RAG como capa de especialización sin fine-tuning	Análisis de literatura	Especialización de dominio es clave en RAG, más efectivo que reentrenamiento
Sun et al.	2025	RAG cognitivo para agentes	No especificado	Recuperación ponderada por perfil cognitivo	Evaluación experimental con agentes	RAG moldea sesgos cognitivos y tendencias de opinión en agentes conversacionales

La tabla 2 menciona los principales estudios relacionados con la arquitectura, comenzando con Lewis et al. (2020) que introdujeron el paradigma RAG para superar esta limitación, Donde este ha sido utilizado en diferentes dominios donde se requiere precisión y relevancia del contenido generado. Ling et al. (2025) demuestra que esta especialización en el dominio es el factor más relevante al momento de evaluar el desempeño de los LLMs en aplicaciones concretas, y que el RAG es el mecanismo más efectivo para lograrla sin necesidad de reentrenar el modelo completo.

Pero al momento de implementarla puede tener ciertas dificultades que Barnett et al. (2024) identifica como los siete puntos de fallos frecuentes en sistemas RAG en producción. Se destacan la recuperación de fragmentos irrelevantes cuando la consulta es ambigua y la dependencia de la calidad del material dado a la herramienta sobre la calidad de las respuestas. Esta última genera relevancia en el proyecto dado que este corpus presenta características particulares como fragmentos con una longitud diferente y los registros orales transcritos, lo cual podrían influir en la calidad de la recuperación.

El uso de esta herramienta en el proyecto sigue una función diferente con respecto al uso estándar en la literatura. Donde su objetivo no es mejorar la validación de hechos concretos, si no que su propósito es mejorar la coherencia lingüística del actor, donde se busca que el agente adopte vocabulario, expresiones y registro oral propio de los testimonios reales proporcionados por la JEP. Sun et al. (2025) es un antecedente relevante para este uso, donde proponen un sistema RAG donde los

fragmentos recuperados son utilizados para modelar sesgos cognitivos y tendencias de opinión en agentes conversacionales. Mostrando que el RAG puede tener funcionalidades más allá de recuperar información específica.

4.3 Arquitecturas BDI

Las arquitecturas cognitivas, son marcos computacionales, donde su objetivo principal es tratar de replicar los procesos mentales que guían el comportamiento de un agente, por ejemplo, la manera de percibir su entorno, sus objetivos y cómo actuar para alcanzarlos. Bratman (1987) propuso el modelo BDI (Beliefs, Desires, Intentions) argumentando que el comportamiento humano racional puede explicarse a través de tres estados mentales: Las creencias sobre el estado del mundo, los deseos que representa los objetivos del agente y las intenciones que son los planes concretos para alcanzarlos. Esto fue la base para Rao y Georgeff (1991) los cuales formalizaron computacionalmente este modelo estableciendo las bases para su integración en sistemas de inteligencia artificial.

Tabla 3 Estado del arte Arquitecturas cognitivas

Autores	**Año**	**Tipo de arquitectura**	**Componentes principales**	**Integración con LLM**	**Método de evaluación**	**Enfoque principal**
Bratman	1987	BDI	Creencias, deseos, intenciones	No aplica	Propuesta teórica	Formalización del comportamiento racional a través de tres estados mentales
Rao & Georgeff	1991	BDI formalizado	Creencias, deseos, intenciones con lógica formal	No aplica	Formalización lógica y pruebas	Formalización computacional del modelo BDI para sistemas IA
Lind	2017	BDI + OCEAN	Creencias, deseos, intenciones; rasgos OCEAN	No especificado	Simulación de comportamiento de equipos bajo presión	Integrar rasgos de personalidad OCEAN en BDI produce comportamientos diferenciados
Harada & Kano	2025	BDI jerárquico de dos niveles	BDI jerárquico, Eneagramas, MBTI	LLM para generación conversacional	Evaluaciones subjetivas de naturalidad humana	Mejora en naturalidad humana con arquitectura BDI jerárquica en agentes de rol

Con la aparición de LLMs en este contexto, Integraciones con el modelo BDI con capacidades de lenguaje natural, ha abierto nuevas posibilidades para el desarrollo de comportamiento coherente y personalizado. Lind. (2017) fue uno de los primeros en aproximarse a esto, donde combino el modelo BDI con el modelo de personalidad OCEAN para simular comportamiento de equipos de trabajo bajo presión, demostrando que la integración de los rasgos de personalidad en la arquitectura BDI produce comportamientos más diferenciados y realistas.

El antecedente más relevante para este trabajo es el propuesto por Harada y Kano (2025), que proponen una arquitectura BDI jerárquica de dos niveles para agentes conversacionales en un juego de rol. Donde el primer nivel, macro BDI gestiona una estrategia a largo plazo para el agente y cuantifica rasgos de personalidad a partir de modelos MBTI y Eneagrama. El segundo nivel es el micro BDI, que gestiona las decisiones tácticas turno a turno. Este experimento demuestra que la integración de esta arquitectura produce una mejora del 14.1% en evaluaciones subjetivas de naturalidad humana a costa de una leve reducción del rendimiento estratégico. Este trabajo extenderá esta propuesta al dominio del conflicto armado, adaptando sus componentes a tres roles narrativos con patrones psicológicos y discursivos totalmente distintos

4.4 Simulación de roles y personajes en sistemas conversacionales

Una rama que se ha venido explorando en la literatura es en la integración de LLMs, en la simulación de personalidades o roles. Donde en la tabla 4 se nos presentan los estudios más relevantes identificados en la búsqueda.

Tabla 4 Estado del arte simulación de roles en sistemas conversacionales

Autores	**Año**	**Tipo de enfoque**	**Método de modelado de personalidad**	**Mecanismo de persistencia**	**Método de evaluación**	**Enfoque principal**
Serapio-García et al.	2023	Perfiles de personalidad como instrucciones	Rasgos medibles dados como prompts	Sin arquitectura de persistencia	Medición de consistencia en rasgos	Perfiles de personalidad medibles pueden imitarse con instrucciones al modelo
Yang et al.	2025	Especificación estructurada del perfil	Características definidas en estructura explícita	Persistencia mediante especificación detallada	Evaluación de directrices de diseño para agentes de rol	Especificación estructurada del perfil mejora consistencia narrativa
Shao et al.	2023	Fine-tuning especializado (Character-LLM)	Fine-tuning por personaje	Persistencia mediante fine-tuning	Comparación con prompts simples	Fine-tuning especializado genera estilos de personaje más estables que instrucciones

Brito et al.	2025	Meta-análisis de métodos	Revisión de marcos existentes	Análisis transversal	Estudio de literatura especializada	Métodos actuales modelan rasgos aislados, falta integración sistémica

La tabla 4 muestran cómo se ha venido modelando la personalidad en estudios recientes, Serapio-García et al. (2023) muestran en su trabajo perfiles de personalidad medibles y consistentes, los cuales pueden ser dados como instrucciones al modelo con el fin de imitar rasgos humanos específicos. Pero estos tienen una limitación de introducir estos rasgos de manera superficial, sin utilizar una arquitectura que garantice la persistencia y coherencia de estos rasgos a través de una conversación extensa y no interacciones cortas. Una primera solución la propone Yang et al. (2024) donde plantea que la especificación estructurada del perfil del agente incluyendo características definidas mejoran significativamente la consistencia narrativa a lo largo de la conversación.

Otra opción para resolver esta limitación la propuso Shao et al. (2023) donde con su framework "Character-LLM" utilizan fine-tuning especializado para simular personas específicas y mantener la consistencia en sus respuestas preservando sus rasgos de personalidad. Donde demuestran que su sistema funciona mejor que dar instrucciones simples a el modelo, generando comportamientos y estilos de personaje de forma más estable, aunque este modelo requiere un proceso de fine tuning para cada personaje, lo cual sería difícil de escalar para nuevos roles y generaría más costos. Brito et al. (2025) realiza un estudio sobre los métodos que se están utilizando actualmente para el modelado de personalidad en estos LLMs, donde identifica una limitación que son la falta de marcos que modelen la personalidad en un sistema integrado y no como rasgos separados entre si (Zheng et al., 2023).

4.5 Evaluación de agentes conversacionales

La evaluación en estos agentes también se ha ido adaptando a los nuevos avances en el campo, transitando de métricas que fueron el estándar en la evaluación de sistemas de dialogo, como BLEU o ROUGE, las cuales miden la similitud entre respuestas generadas con respecto a una respuesta referencia mediante el solapamiento de palabras. Pero estas métricas han demostrado no tener tanta correlación con la calidad percibida por humanos en tareas de generación abierta, ya que una respuesta puede contener palabras diferentes a las de la referencia y seguir siendo coherente (Zheng et al., 2023).

Tabla 5 Estado del Arte: Evaluación de agentes

Autores	Año	Tipo de evaluación	Dimensiones evaluadas	Método de puntuación	Validación	Enfoque principal
Zheng et al.	2023	LLM-as-a-judge (MT-Bench, Chatbot Arena)	Calidad conversacional general	Rubricas detalladas guiadas a LLM	Correlación >80% con evaluaciones humanas	LLM guiado por rubricas actúa como evaluador automático escalable
Chen et al.	2025	Framework de evaluación basado en teoría	Personalidad, coherencia narrativa, comportamiento situacional	Atributos del agente con métricas generalizables	Meta-análisis de 1600+ publicaciones (2021-2024)	Evaluación de agentes de rol debe ser multidimensional y adaptable al dominio
Es et al.	2023	RAGAS (Retrieval Augmented Generation Assessment)	Fidelidad de respuesta, relevancia, precisión del retriever	Descomposición en tres dimensiones independientes	Validación experimental en datasets RAG	Evaluar componentes RAG por separado identifica qué genera errores
Jiang et al.	2024	PersonaGym (decisión teórica)	Fidelidad del perfil psicológico, consistencia a lo largo de la conversación	Persona Score basado en teoría de decisiones	Validación experimental con agentes de rol	Evaluar agentes requiere métricas multidimensionales de coherencia psicológica

La tabla 5 presenta artículos donde se investigan nuevas formas de evaluar a los agentes conversacionales, Zheng et al. (2023) introducen el paradigma LLM-as-a-judge con MT-Bench y Chatbot Arena, donde sustentan que si a un LLM es guiado mediante rubricas detalladas, estos pueden actuar como evaluadores automáticos, teniendo una correlación superior al 80% con evaluaciones humanas. Este enfoque tiene gran capacidad de escalabilidad y flexibilidad para adaptar criterios de evaluación al dominio especifico.

Chen et al. (2025) realizaron una revisión de sistemática de la literatura, donde tras más de 1600 publicaciones entre 2021 y 2024 detectan diversos métodos para la evaluación de agentes de rol e identifican la necesidad de frameworks de evaluación basados en teoría que puedan ser adaptables a el dominio especifico. Ellos proponen una guía de diseño de evaluación basado en atributos del agente y en métricas generalizables, concluyendo que esta evaluación debe tener en cuenta dimensiones de personalidad, coherencia narrativa y comportamiento situacional.

Para el módulo RAG Es et al. (2023) proponen en su framework de evaluación RAGAS descomponer la evaluación general del componente, y evaluarlo en tres dimensiones individuales. Fidelidad de la respuesta con respecto a el contexto recuperado, relevancia de la respuesta frente a una consulta y la precisión del retriever. Lo cual ayuda identificar la parte que podría estar generando errores.

Para evaluar agentes de rol Samuel et al. (2024) proponen PersonaGym, donde a partir de la teoría de decisiones introducen la métrica de Persona Score para evaluar la fidelidad del perfil psicológico del agente a lo largo de la interacción, donde plantean que la evaluación de agentes de rol requiere métricas multidimensionales que no solo miren la coherencia léxica si no la consistencia psicológica con el perfil a lo largo de la conversación

4.6 Vacíos identificados en la literatura

A partir de la revisión del estado del arte, se identifican tres vacíos metodológicos, En primer lugar, se evidencia un desarrollo aun limitado de en la integración de arquitecturas cognitivas que sean implementables computacionalmente con mecanismos de recuperación de información en corpus especializados. Existen trabajos que integran estos componentes de manera independiente como el BDI jerárquico aplicado a agentes conversacionales de rol (Harada & Kano, 2025) o el RAG orientado a la especialización de dominio (Ling et al., 2025), la integración en un sistema que utilice ambas para simular narrativas testimoniales no ha sido explorado en la literatura actual.

Por otro lado, se encontraron arquitecturas cognitivas más completas como ACT-R, pero no existe una integración nativa con LLMs, implementar esta arquitectura en el sistema requeriría una capa de traducción compleja entre el modelo simbólico y la generación de lenguaje natural. El BDI jerárquico ofrece una representación cognitiva suficiente para modelar los roles de este trabajo, con una integración directa y documentada con LLMs (Harada & Kano, 2025), por lo que la convierte en una arquitectura adecuada para este trabajo.

Finalmente. No se encontró un sistema de evaluación estándar para medir la autenticidad testimonial en el dominio específico del conflicto armado. Los trabajos disponibles como PersonaGym (Samuel et al., 2024) evalúan la consistencia de personalidad en contextos genéricos sin incorporar las características discursivas y emocionales asociadas a los testimonios del conflicto armado. Esto justifica el diseño de evaluación adoptado en el trabajo, que se basa en métricas multidimensionales adaptadas al dominio y con el paradigma de LLM-as-a-judge con diferentes jueces para mitigar sesgos sistemáticos que podrían aparecer en el uso de un único juez.

5. Metodología

La metodología que utilizara este proyecto se guiara por el marco de Design Science Research (DSR) ya que según (Hevner et al., 2004) es un paradigma de resolución de problemas que busca extender las capacidades humanas o sociales para resolver un problema en concreto mediante la creación de artefactos innovadores. Donde el problema en este caso sería la representación de forma psicológicamente coherente las narrativas de actores del conflicto armado colombiano a partir de fuentes documentales reales.

Esta se organiza en 6 pasos desde la fundamentación teórica hasta la evaluación del mismo sistema. Los dos primeros pasos enfocados en establecer las bases del proyecto, la revisión del estado del arte para guiar las decisiones de diseño, el procesamiento del corpus de los testimonios de la CEV el cual será la base de conocimiento del sistema. El tercer paso se centra en el diseño e implementación del artefacto principal, los agentes conversacionales con arquitecturas BDI jerárquicas y RAG. El cuarto paso centrado en la evaluación y retroalimentación del sistema RAG, las arquitecturas cognitivas y el agente general en términos de coherencia narrativa, consistencia de personalidad y autenticidad discursiva. Estos dos últimos pasos siendo iterativos para la mejora continua del artefacto. Y los pasos cinco y seis consolidando el análisis de los resultados y conclusiones del trabajo.

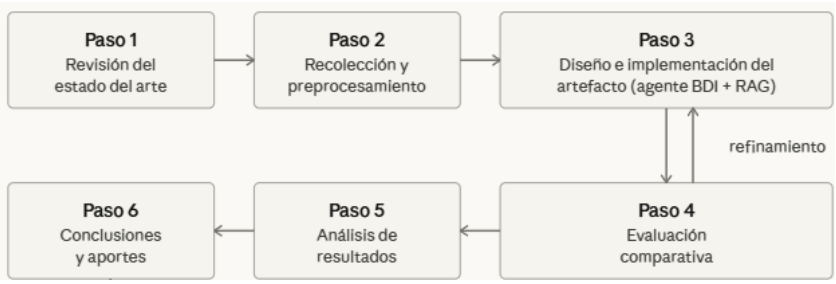

Ilustración 1 Metodología propuesta

5.1 Recolección de datos y preprocesamiento de texto

El segundo paso se centró en la construcción de la base de conocimiento para el artefacto, donde se transformarán un conjunto de entrevistas en fragmentos semánticos coherentes los cuales serán clasificados en diferentes roles, los cuales serán usados después por el módulo RAG. Estas entrevistas provenientes de la CEV en formato PDF y además se cuenta con un archivo de tipo Excel, donde para cada entrevista se cuenta con información del lugar, la fecha de realización, inicio y fin de los hechos a

los cuales se están refiriendo, el tipo de entrevista donde se identifica si se hicieron a grupos o personas individuales y dentro de esta a que actor involucrado en el conflicto, variables categóricas para saber si son de interés para cierto tipo de estudios o para ciertas áreas de interés.

Con el fin de comprender el conjunto de datos, antes de proceder a el procesamiento, se realizó un análisis descriptivo de los metadatos provenientes del archivo Excel. Donde de 14.972 entrevistas realizadas por la CEV, 2.233 contaron con metadatos estructurados, de las cuales 2.240 fueron transcritas y anonimizadas. Analizando la distribución temporal de la realización de las entrevistas, estas se concentran entre 2019 y 2020, período que corresponde a la fase operativa central de la CEV, Pero los hechos relatados abarcan desde la década de 1960 hasta 2020, con la mayoría de hechos en la década de los 2000 lo cual coincide con la fase de mayor intensificación del conflicto armado en Colombia.

En la composición por tipo de entrevista, el 72.59% corresponde a entrevistas a víctimas, familiares y testigos, seguido de entrevistas en profundidad con un porcentaje de 16.88% y entrevistas a actores armados y excombatientes representan el 3.94% del corpus. Lo cual podría influir en la base de conocimiento para el RAG ya que para el rol de victima contara con más registros, con respecto a el rol de victimario.

Temáticamente, el Núcleo 3 denominado “actores armados y responsabilidades” domina el corpus con 1.086 entrevistas, seguido del Núcleo 4 sobre dinámicas económicas y despojo con 356 registros y el Núcleo 7 sobre conflicto en pueblos étnicos que cuenta con 260 ítems.

Sabiendo esto, se procede con el procesamiento, donde por cada entrevista se extrae el texto del pdf y este pasa por 4 funciones las cuales se encargan de normalizar las etiquetas de anonimización de cada entrevista ya que tenían errores de escritura, o muchas etiquetas con significados muy similares. La segunda función reemplaza tokens de anonimizarían (-----) que se utilizan para señalar entidades en el texto. Generando fragmentos con una ventana de 80 tokens alrededor de cada bloque, y se clasifican mediante unas señales dadas, y una clasificación como alta, media o baja para cada tipo de etiqueta, y para complementar esto, se utilizó el modelo “es_core_news_lg” de spaCy para el reconocimiento de entidades nombradas. Y al final de esto se genera el token para reemplazar escogiendo la categoría más probable y este se guarda en MongoDB junto a su meta data. La tercera función limpia ruidos generados

por la transcripción ya que no generan valor al proyecto, pero dejando marcadores discursivos muy comunes en el contexto colombiano. Y esto se guarda en una colección de mongoDB. Y a partir de estos se seleccionan solo los fragmentos correspondientes a el entrevistado, ya que son los relevantes para nuestro caso.

Teniendo estos fragmentos, se procede a hacer un chunking semántico, el cual se basa en buscar la similitud semántica entre el texto, utilizando el algoritmo de SemanticChunker proveniente de Langchain, y utilizando el modelo "paraphrase-multilingual-mpnet-base-v2" para detectar los cambios de tema en la conversación. Y teniendo estos chunks separados se procedió a hacer la clasificación en los 3 roles definidos. Para esto utilizando 2 capas, un score generado score, calificando si contiene ciertas palabras asociadas a ese rol, y luego utilizando los embeddings creados anteriormente se calcula con una similitud del coseno con unos patrones dados y luego mezclando estas dos métricas, se tiene un score total. El cual es utilizado para que, al momento de pasarlos a la siguiente fase, pasen los que tienen contexto acerca del rol. Y teniendo estos chunks se pasan por medio de un LLM, En este caso GPT 4o mini, donde se le pasa el tipo de entrevista, junto con cada chunk para que este lo clasifique, y estos se guardan en mongoDB atlas como una base de datos vectorial junto con su índice vectorial para su posterior uso.

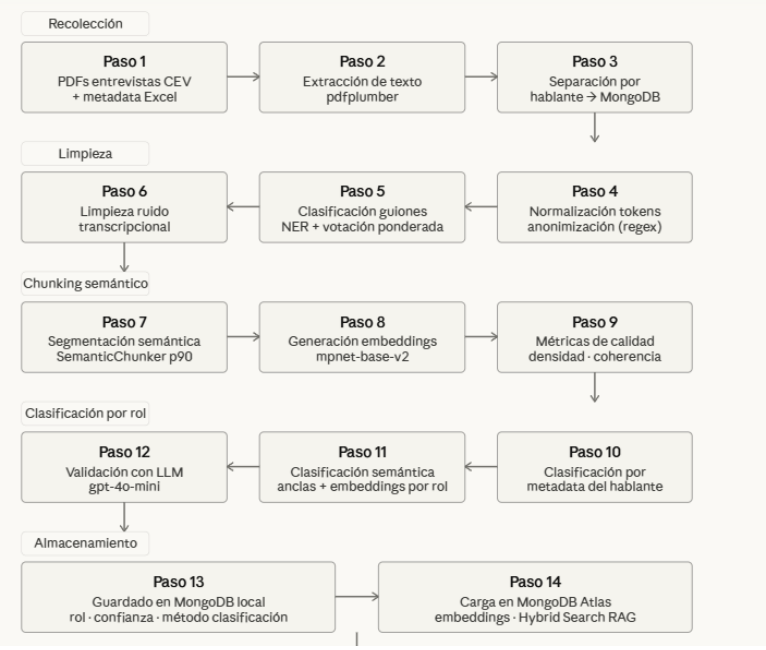

Ilustración 2 Pipeline de preprocesamiento y clasificación

5.2 Diseño e implementación del artefacto

Después de tener la base de conocimiento, se procede a diseñar e implementar el artefacto central del proyecto. Un agente cognitivo, cuyo objetivo será representar las narrativas del conflicto armado colombiano de una forma psicológicamente coherente. Donde se integrarán tres componentes principales, Un modelo LLM, un mecanismo RAG y una arquitectura BDI jerárquica, todo esto funcionara en conjunto con un grafo de estados. El sistema sigue la siguiente arquitectura:

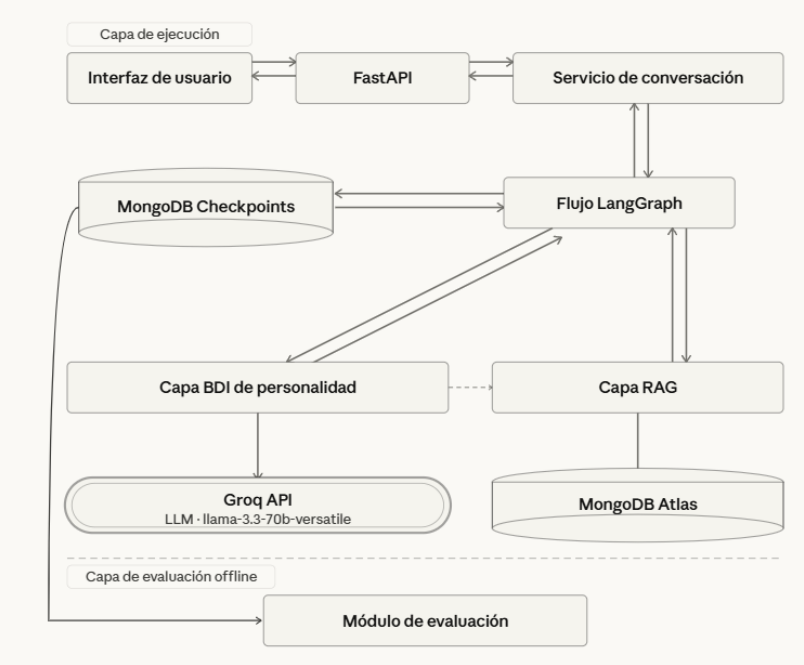

Ilustración 3 Diagrama de arquitectura del Sistema

5.2.1 Interfaz de usuario y capa de API

Esta interfaz de usuario es una aplicación web que fue construida utilizando HTML,CSS y JavaScript, la cual se conecta desde FastAPI para acceder a la lógica del programa. Al iniciar la página, se ve un menú para la elección inicial del rol con el que se quiere interactuar, al escoger el rol, se entra a una interfaz dividida en dos secciones: una barra lateral, para en caso de ser necesario, cambiar el rol con el que se interactuar, también un botón para reiniciar la memoria si así se quiere y la otra sección es el área principal de la conversación, donde se puede ver el historial de los mensajes y el campo para la entrada de texto del usuario.

Esta interfaz mantiene 3 variables de estado globales en la memoria del navegador: el rol activo, la conexión con WebSocket vigente y un tipo de historial que preserva las conversaciones de cada rol de manera independiente.

La capa de la API está implementada con FastAPI y esto para conectar la interfaz y el servicio de conversación. Expone cinco endpoints: uno para servir la interfaz, uno para listar los roles disponibles, uno para inicializar el actor con personalidad BDI, uno WebSocket para el streaming del texto, y dos endpoints de reinicio de memoria, uno global y uno por rol específico. En este archivo también almacena a el actor ya inicializado con su personalidad calculada, esto ya que en cada interacción se debe consultar para seguir una coherencia lógica

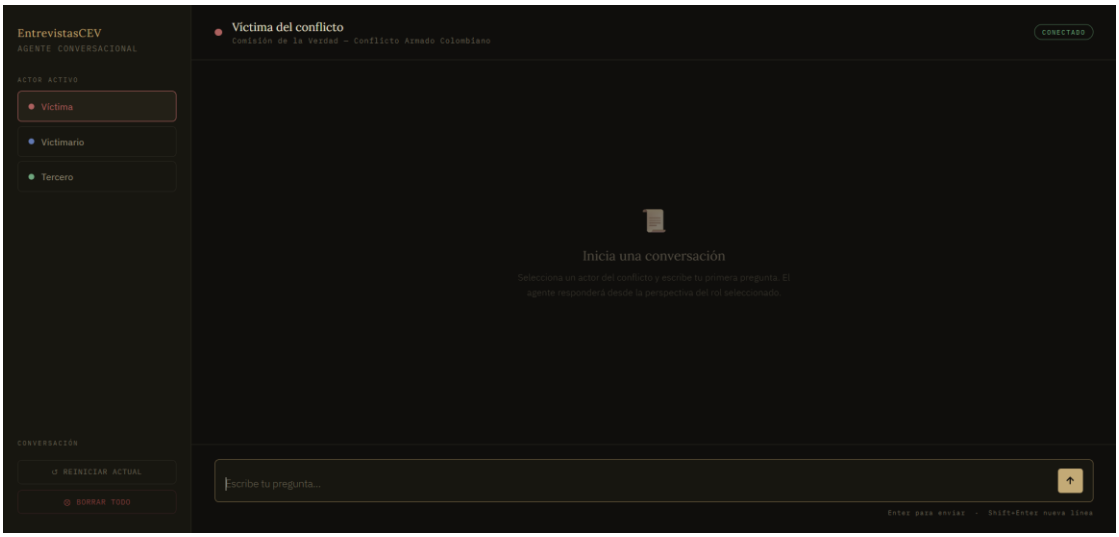

Ilustración 4 Interfaz gráfica de la aplicación

5.2.2 Flujo LangGraph

El núcleo principal de todo el sistema es un grafo de estados implementado mediante LangGraph que coordina el comportamiento del agente en cada turno de la conversación, LangGraph permite modelar el flujo de procesamiento como un grafo dirigido donde cada nodo realiza una operación específica sobre un estado compartido, y las aristas. Este grafo define también el orden y posibles decisiones en el proceso.

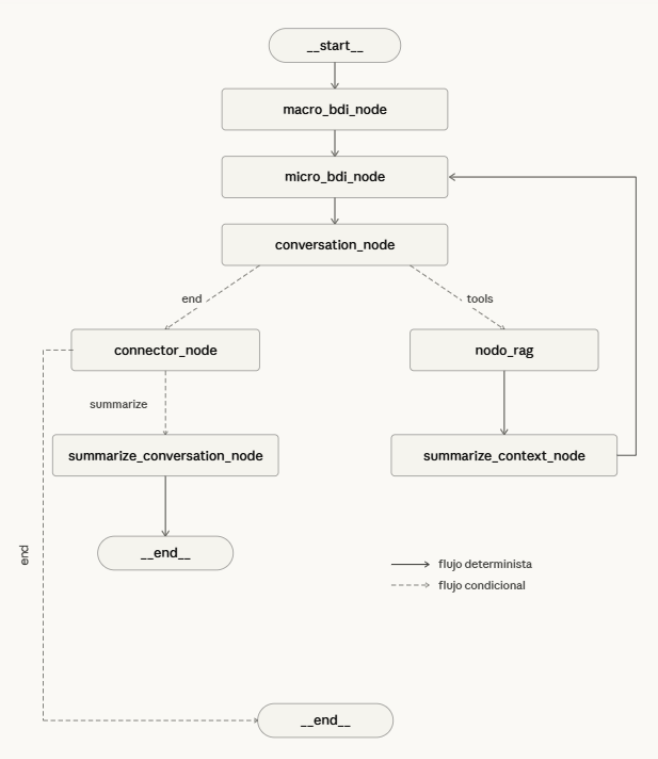

Ilustración 5 Flujo del grafo de estados

5.2.2.1 Estado compartido del grafo

Esta es la información que va a fluir por todo el grafo, Este estado contiene información del actor, como es: su rol, su perspectiva, su estilo comunicativo y el resumen acumulado de la conversación. Contiene también el MacroBDI que contiene: el macro deseo, que captura el objetivo narrativo global, la macro creencia con el perfil de personalidad formateado, y los parámetros de comportamiento con los valores numéricos de empatía, agresividad y evitación. El tercer grupo corresponde al Micro-BDI: que es la micro intención generada en cada turno, la etapa discursiva activa, el nivel de tensión acumulado, el conteo de negaciones y el grado de reconocimiento dado.

5.2.2.2 Estructura y nodos del grafo

El flujo de este grafo comienza con “nodo_macro_mdi” que es el encargado de crear la personalidad general del agente si aún no se ha creado, y este se guarda en unos checkpoints en mongoDB para después consultarlo, pasa a “nodo_micro_bdi” donde ejecuta todo el ciclo del componente MicroBDI para ese turno, analiza el mensaje recibido, actualiza el estado relacional y genera la intención táctica para ese turno. Luego “conversation_node” formatea el prompt completo

incluyendo los campos del estado compartido y lo pasa como petición a el LLM, que para todo el artefacto se utilizara GPT 4o mini.

En este último nodo, puede dirigirse a dos posibles nodos: es el modelo necesita que necesita contexto adicional de los testimonios de la CEV devuelve una petición para dirigir el flujo hacia "nodo_rag" donde se procede a hacer la búsqueda hibrida correspondiente para obtener los chunks más relevantes, estos pasan a "summarize_context_node" que es el encargado para extraer vocabulario característico o información relevante para el estilo de habla del agente y luego vuelve a el flujo principal. Y el otro posible nodo es "connector_node" que evalúa si el historial supera 30 mensajes, el grafo se dirige a "nodo_summirize_context" que genera un resumen de la conversación con un LLM más liviano para reducir costos computacionales en este caso y elimina los mensajes dejando los mensajes excepto los últimos cinco

5.2.3 Capa BDI de personalidad jerárquica

Esta arquitectura BDI jerárquica fue adaptada del modelo propuesto por Harada y Kano (2025) a el dominio del conflicto armado colombiano. Donde opera en dos niveles que se complementan entre sí. El nivel macro establece el perfil psicológico y el objetivo narrativo del actor antes de que comience la conversación, mientras que el nivel micro ajusta el comportamiento turno a turno teniendo en cuenta la dinámica de la conversación.

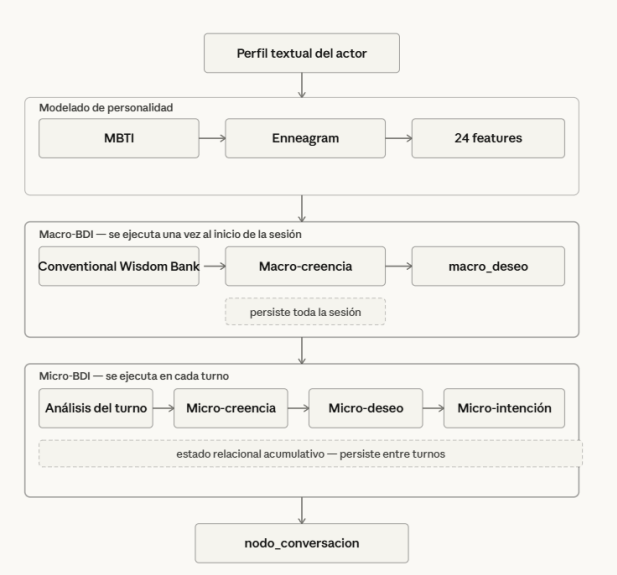

Ilustración 6 Arquitectura BDI Jerárquica

5.2.3.1 Modelado de Personalidad.

Este módulo toma como entrada una descripción textual del actor y lo transforma en un conjunto de valores numéricos los cuales representan un perfil psicológico de una manera entendible para el agente. El texto de entrada es la concatenación de dos campos definidos para cada rol: la perspectiva, que describe cómo vivió o participó en el conflicto, y el estilo, que describe cómo habla y se expresa. Este texto se envía al modelo LLM con un prompt que le pide estimar ocho dimensiones MBTI como valores continuos en el rango [0, 1].

Con estos valores del MBTI se aplican las fórmulas del apéndice B.1 del paper para estimar los 9 tipos de eneagramas como afinidades continuas, donde cada tipo es una combinación de los coeficientes de las dimensiones MBTI (Harada & Kano, 2025, Apéndice B.1). Y cada rol tiene un score para cada tipo de eneagrama, donde el perfil domínate será el de mayor valor, sin excluir a los otros perfiles. Esta combinación de los modelos MBTI y los eneagramas permite que los modelos de personalidad sean comprensivos desde dos perspectivas complementarias: el sesgo cognitivo y la estructura motivacional (Harada & Kano, 2025)

Estos rasgos servirán para calcular 24 características adicionales organizadas en cinco grupos: El primer grupo, sesgo declarativo, describe cómo el actor estructura su discurso, si este tiende a la coherencia lógica o al detalle concreto. El segundo grupo, tendencia de confianza, determina qué tan fácil confía en el interlocutor y qué tipo de argumentos lo convencen. El tercer grupo, tendencia de afinidad, captura cuánto le importa la relación emocional con quien conversa. El cuarto grupo, tendencia de deseo, modela las motivaciones de largo plazo, la necesidad de reconocimiento, de estabilidad o de independencia. El quinto grupo, tendencia de comportamiento, es el más directamente observable en la conversación: la conducta evasiva ante temas dolorosos, la agresividad discursiva y el asertividad al imponer el propio relato. Y estas características se utilizan siguiendo las fórmulas del apéndice B.2-B.6 del paper (Harada & Kano, 2025, Apéndice B).

5.2.3.2 Macro-BDI

La intención general es implementar la sección 3.4.7 del paper de Harada y Kano (2025) y la funcionalidad es establecer el objetivo narrativo del actor antes de que comience cualquier turno de

conversación, cambiando las tácticas de juego, por patrones discursivos del conflicto armado colombiano, esta capa maneja la creencia y el deseo, donde el macro deseo actúa como un sesgo en las decisiones que se toman en la capa micro.

La macro creencia es generada a partir de los valores numéricos generados en la sección anterior, este genera un texto descriptivo que incluye el eneagrama dominante, los valores de comportamiento más relevantes y las tendencias de deseo. Y a partir de este texto, se genera el macro deseo, utilizando un LLM, que como se hace en el paper de Harada y Kano (2025) se combina con patrones discursivos, se envía a el modelo y este escoge una entre los candidatos disponibles por rol y situación. El modelo nos genera el objetivo del agente como una oración en primera persona. Esta oración es la que persiste en todo el grafo de LangGraph para guiar el comportamiento del agente.

5.2.3.3 Micro-BDI

Este componente opera en cada turno de conversación, ejecutándose justo antes de que el modelo principal genere la respuesta del actor. Su función es evaluar el último mensaje recibido en el contexto de la conversación acumulada y producir una intención táctica concreta para ese turno, siempre alineada al macro deseo establecido en la capa superior.

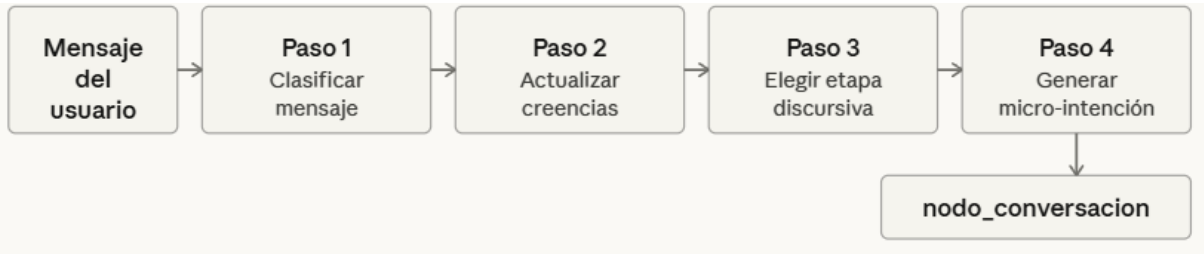

Ilustración 7 Micro-BDI - Flujo por turno

Este flujo comienza analizando el mensaje de cada turno, clasificándolo por uno de los cinco tipos, entre acusación, reconocimiento, pregunta, evasión y neutro donde se capturan las dinámicas discursivas propias de una interacción entre roles. Después de esto se actualiza la micro creencia siguiendo la estructura de actualización circular propuesta por los autores, donde las impresiones a largo plazo por interlocutor se agregan a nivel del agente y retroalimentan la corrección de credibilidad en

nuevos turnos (Harada & Kano, 2025) y se aplican reglas fijas sobre el estado relacional del agente, actualizando el nivel de tensión de la conversación, el conteo de negaciones y el reconocimiento dado.

A partir de estas variables se genera el micro deseo generando a el agente lo que quiere lograr en el siguiente turno, teniendo como opciones apertura narrativa, defensa de su posición, negociación para contar la verdad, confrontación *y* cierre simbólico. Y esta elección la hace mediante un LLM siguiendo heurísticas relacionadas a las variables del estado del actor. Y por último se genera la micro intención mediante un LLM, que siguiendo el formato de (Harada y Kano, 2025) se generan dos campos, uno para describir la estrategia táctica del turno para ejecutar el micro deseo, esto en términos no conversacionales, y el otro campo que es lo que va a decir el actor en términos concretos, y esto es lo que se envía a el nodo de conversación.

5.2.4 Capa RAG y Memoria a largo plazo

Este módulo es el encargado de proporcionarle más contexto al agente sobre el conflicto armado colombiano. Esto para no depender de los datos con los que fue entrenado el modelo, el cual recupera los chunks provenientes de la CEV que fueron clasificados anteriormente, lo cual ayuda a que el agente responda con el lenguaje, el tono emocional y el vocabulario auténtico de personas que vivieron el conflicto.

En esta fase se crea un retriever que implementa una búsqueda hibrida, combinando una búsqueda vectorial, que, a partir de la consulta del usuario, se convierte en un vector, mediante el mismo modelo de embeddings usado en el preprocesamiento. Y procede a buscar los fragmentos más cercanos en el espacio vectorial utilizando un producto punto como función, lo cual ayuda a buscar fragmentos que tengan un contexto similar, aunque no tengan las mismas palabras. Esto se complementa con una búsqueda de texto completo basado en la frecuencia de términos, esto para nombres de lugares, siglas o demás cosas muy específicas.

Teniendo estos fragmentos se le pasan a un LLM con el fin de que para todos estos fragmentos extraiga vocabulario propio del conflicto, incluyendo palabras, frases, eufemismos para describir hechos actos de violencia de alta carga moral, Expresiones orales típicas del discurso o formas de

nombrar cosas muy propias del conflicto armado. Y esto es lo que se le pasa como contexto a la respuesta del agente

5.3 Evaluación del Artefacto

La evaluación del agente representa una etapa demasiado importante, ya que ayuda a determinar en qué medida cada componente del agente está contribuyendo a la autenticidad de las respuestas. La siguiente sección presenta el modelo que se utilizó para este proyecto

5.3.1 Diseño Experimental

El artefacto se evaluará bajo un diseño de ablación factorial 2x2 en el que se evaluará el efecto causal aislado de los dos componentes del artefacto, el módulo RAG y el módulo BDI, esto produce cuatro condiciones experimentales. Lo cual permitiría determinar en qué medida la arquitectura BDI jerárquica mejora la calidad de la respuesta respecto a el sistema más simples y si RAG está aportando en la parte léxica del agente.

Condición	**BDI activo**	**RAG activo**	**Propósito de la evaluación**
BASE	✗	✗	Artefacto sin ningún componente
RAG	✗	✓	Efecto causal del RAG aislado
BDI	✓	✗	Efecto causal del BDI aislado
BDI+RAG	✓	✓	Efecto combinado

Para construir el conjunto de prueba sobre el que se ejecutaran las evaluaciones. Se tomará como insumo un banco de preguntas, que contará con 13 preguntas por cada uno (Véanse en el anexo B). Y así mismo se repetirán estas mismas preguntas para cada condición experimental. Con lo cual se estarían generando 156 registros, donde se guarda para cada turno los siguientes campos:

- Condición:
- Actor_id

- Turno
- Pregunta
- Respuesta
- Macro_deseo
- Micro_deseo
- Micro_intencion
- Actor_contexto (información dada por el RAG)
- Métricas del retrieval

5.3.2 Enfoque de evaluación: LLM-as-a-judge

Basándonos en el paradigma de LLL-as-a-judge propuesto por Zheng et al. (2023) por las ventajas que presenta, que fueron descritas en el estado del arte, se decide evaluar las dimensiones cualitativas del sistema con este paradigma.

En este caso se utilizarán 3 modelos diferentes : “Gemma4”, ”LlaMA 3.2” y “Mistral 7B” esto para minimizar sesgos o errores que podría presentar un modelo, cada uno con una temperatura de 0.1 para maximizar la reproducibilidad, se les pasarán prompts dinámicos adaptados para cada rol del actor a evaluar. Esto ya que cada actor tiene un patrón narrativo distintos, y evaluarlos con la misma rubrica podría incluir sesgos sistemáticos contra los roles con patrones más complejos, para cada métrica se cuenta con una rubrica de evaluación en escala tipo Likert de 1 a 5 donde los niveles de 1 a 3 serían para respuestas genéricas y el 4 o 5 utilizado para respuestas casi indistinguibles de un testimonio humano real. Y como último paso, los resultados de todos los modelos se promediarán entre sí, para obtener un puntaje global por métrica de los jueces.

5.3.3 Métricas de evaluación.

Las métricas estarán divididas en diferentes secciones según su propósito a evaluar y su disponibilidad por condición experimental.

Métricas Universales. Estas son las métricas utilizadas para las comparar las cuatro condiciones experimentales y no dependen de ningún componente para su evaluación, estas son calculadas utilizando LLM as a judge.

- **Estabilidad del Rol:** Esta métrica evalúa la capacidad del agente de mantener una identidad discursiva coherente a lo largo de toda la conversación. Identificando coherencia de la postura del relato y su perspectiva. 1 siendo un score donde abandona su perspectiva o no suena como su rol y 5 para un relato con un nivel de naturalidad y coherencia comparable con un testimonio humano. Esta métrica es adaptada de Consistency of conversation content propuesta por Harada y Kano. (2025).

- **Autenticidad Léxica:** Evalúa turno a turno el grado en el que el vocabulario del agente es verosímil ya adecuado para alguien que vivió o presencio el conflicto armado, Independientemente si se usó el RAG o no. Esta métrica es adaptada de las métricas de fuerza de estilo propuestas por Fu et al. (2018) y en nociones de adecuación del registro lingüístico, donde esta penaliza el lenguaje académico o neutro así sea gramaticalmente correcto. 1 siendo un vocabulario inapropiado para el rol o que no se usa en el contexto dado y 5 como el uso consistente del léxico apropiado a el contexto de la pregunta y natural dentro del rol.
- **Autenticidad Emocional:** Evalúa si el tono afectivo del agente es verosímil y coherente con el rol. Esta fue creada aparte de autenticidad léxica ya que un agente podría usar el vocabulario correcto, pero con un tono incorrecto. Donde 1 sería una ausencia de emoción o un tono que no utilizaría el rol y los valores cercanos a 5 serían tonos emocionales similares a los que tendría este rol.
- **Orientación a un objetivo:** Evalúa si las respuestas del agente se encaminan a un propósito definido, incluyendo la coherencia, intenciones implícitas que describe a lo largo de la conversación. Esta métrica se calcula sin tener acceso al macro deseo. Esto integrando métricas de coherencia global en la narrativa computacional explicada por

McIntyre & Lapata. (2009). Los puntajes más bajos se dan a las respuestas sin conexión lógica entre sí y los valores más altos a respuestas que tienen una dirección narrativa clara y sostenida dependiendo del rol

- **Consistencia Táctica**: Basada en Jayakumar et al. (2025) esta métrica evalúa si la estrategia discursiva del agente para cada pregunta es la apropiada dependiendo de su contexto. Donde 1 sería una estrategia que es totalmente inapropiada y 5 una estrategia totalmente apropiada

Métricas RAG. Estas métricas solo se verán reflejadas en las condiciones donde el componente del RAG este activo, lo cual ayuda a validar la calidad del módulo y uso del contexto dado por la herramienta. Estas evaluaran una parte diferente del proceso lo cual ayudan en conjunto a evaluar la totalidad del componente.

- **top_k_similarity:** Esta métrica evalúa la calidad del retriever, identificando si la búsqueda vectorial recupera documentos genuinamente relevantes para la instrucción que se está recibiendo y se calcula utilizando la similitud coseno entre el vector de la consulta y los vectores de los k fragmentos recuperados de la base de datos, y es calculada con la siguiente formula.

$$\text{top_k_similarity} = \frac{1}{k}\sum_{i=1}^{k}\frac{\vec{q}\cdot\overrightarrow{d_i}}{|\vec{q}|\,|\overrightarrow{d_i}|}$$

  Donde $\vec{q}$ es el embedding de la consulta y $\overrightarrow{d_i}$ el embedding del $i\,esimo$ fragmento recuperado,esta métrica se calcula directamente en el retriever al momento de la búsqueda

- **Adopción Léxica:** Usando un LLM as a judge, evalúa si el agente incorporo el registro léxico dado por la herramienta y si en la respuesta se integra de una manera natural, esta métrica está basada en la métrica de faithfulness del marco RAGAS propuesto por Es et al, (2023), la cual para nuestro caso presentaría limitaciones ya que esta se basa en estrictamente cuanto se basa la respuesta en el contexto dado,

y lo que medimos acá es la capacidad del agente para utilizar el vocabulario dado por la herramienta, sin ser una lista a cumplir estrictamente, si no que se busca una integración más natural y apropiada a la respuesta.

- **Transferencia léxica RAG:** Esta métrica se enfoca en la integración del vocabulario dado por los fragmentos recuperados para cada turno, verificando efectivamente que tanto aparece en la respuesta generada, definida, así como la proporción de términos del fragmento que el agente en realidad integro en su respuesta.

Siendo: $L_C = \{\text{Lexico dado por } actor\ contexto\}$

$$Transferencia\ Lexica\ RAG = \frac{|\{l \in L_C : l \text{ aparece en respuesta}\}|}{|L_c|}$$

- **Transferencia eufemismo RAG:** Esta métrica evalúa que tanto de los eufemismos o formas de hablar de violencia extraídos por la herramienta, son en realidad usados por el agente en su respuesta.

Siendo: $E_C = \{\text{eufemismos dados por } actor\ contexto\}$

$$transferencia\ Eufemismos\ RAG = \frac{|\{e \in E_C : e \text{ aparece en respuesta}\}|}{|E_C|}$$

- **Transferencia expresiones RAG:** Evalúa la adopción de expresiones literales y registro oral que se obtuvieron en la herramienta fueron utilizados en la respuesta del actor,

Siendo $X_C = \{\text{eufemismos dados por } actor\ contexto\}$

$$transferencia\ expresiones\ RAG = \frac{|\{x \in X_C : x \text{ aparece en respuesta}\}|}{|X_C|}$$

- **Novedad de vocabulario:** Esta evaluación se hace con respecto a la respuesta del agente base, y mediante el corpus de vocabulario de conflicto generado en el análisis descriptivo por cada actor, se evalúa la cantidad de estas palabras nuevas de conflicto que se agregaron en RAG, que no están en el agente base

$$\text{novedad vocabulario} = \frac{|\text{palabras diccionario que usa RAG} - \text{palabras diccionario que usa BASE}|}{|\text{palabras diccionario que usa RAG}|}$$

Estas últimas métricas basadas al concepto de "Formuláic Language Adoption" en lingüística (Wray & Perkins, 2000). Donde se evalúan palabras individuales, como conjunciones lingüísticas para capturar si el agente no solo integra el vocabulario si no patrones de habla

Métricas BDI. Las siguientes métricas están basadas en las propuestas por Harada y Kano, (2025) y están diseñadas para evaluar el funcionamiento de la arquitectura BDI, Por lo cual solo estarán evaluando las condiciones que tengan este componente activo.

- Coherencia táctica: Evalúa turno a turno si la respuesta refleja la etapa discursiva asignada por el micro deseo con el tono apropiado al rol y si esta avanza a su micro intención. Donde 1 refleja una contradicción total con la etapa dada y no se ve ningún avance a su micro intención y 5 refleja con totalidad la etapa seleccionada con un tono apropiado y se ve un avance coherente a su micro intención.
- Subordinación macro deseo: Evalúa toda la conversación para determinar si las respuestas son coherentes con el macro deseo generado por el componente al inicio de la sesión. Este teniendo puntajes bajos si no existe una relación observable o tiene contradicciones claras y puntajes más altos si el macro deseo es perceptible a lo largo de la conversación.

5.3.4 Diseño estadístico: Evaluación con múltiples jueces.

Para evaluar entre todas las respuestas y seleccionar cual es la que suena más autentica o plausible a ser una respuesta dada en un testimonio real del conflicto armado colombiano, se escogieron 3 LLMs como GPT-5 desarrollado por OpenAI, Gemini 3 Flash desarrollado por Google y Qwen 3.6 plus desarrollado por Alibaba. Estos modelos fueron seleccionados debido a ser modelos de frontera actualmente, los cuales cuentan con capacidades avanzadas de razonamiento y a su vez capacidades multilingües robustas, lo que resulta primordial para el análisis en español.

El diseño experimental sigue una estructura en bloques completos, para cada una de las 39 preguntas consideradas bloques, se evaluaron las 4 condiciones experimentales bajo el mismo contexto base. En cada bloque, las respuestas son dependientes entre sí, al ser generadas bajo la misma pregunta, pero todas las preguntas son independientes entre sí.

Se utilizo un esquema de evaluación ciega, donde para cada pregunta, se evaluaran las 4 respuestas generadas por las diferentes de las condiciones, estas anonimizadas aleatoriamente como A, B, C, D para no revelar su origen, Y cada modelo genera un ranking ordinal de preferencia, una puntuación individual en escala del 1 al 10 para cada respuesta, donde se evalúa la autenticidad de un testimonio real del conflicto armado, evaluando la oralidad natural, presencia de detalles concretos, emociones implícitas, coherencia con el rol y penaliza el lenguaje demasiado perfecto o literario, con estructuras demasiado organizadas o respuestas similares a textos académicos.

Dada esta estructura de datos, se verifico empíricamente la adecuación del enfoque no paramétrico mediante análisis descriptivo y pruebas de normalidad sobre las 12 distribuciones que corresponden a el diseño del experimento, la Tabla 6 presenta las estadísticas descriptivas de los datos:

Tabla 6 Análisis descriptivo y prueba de Shapiro-Wilk por Juez y condición

Juez	Condición	n	Mediana	IQR	Media	DE	W Shapiro	p	Normal
GPT-5	BASE	39	5.0	1.0	5.59	1.04	0.900	0.0022	No
GPT-5	BDI	39	6.0	1.0	5.67	1.03	0.868	0.0003	No
GPT-5	RAG	39	7.0	1.0	7.23	1.11	0.895	0.0015	No
GPT-5	BDI+RAG	39	8.0	2.0	7.74	1.16	0.850	0.0001	No
Gemini 3 Flash	BASE	39	4.0	1.0	4.36	1.06	0.878	0.0005	No
Gemini 3 Flash	BDI	39	5.0	1.0	4.67	1.15	0.879	0.0006	No
Gemini 3 Flash	RAG	39	8.0	2.0	7.74	1.29	0.838	0.0001	No
Gemini 3 Flash	BDI+RAG	39	8.0	2.0	7.90	1.19	0.814	1.65e-05	No

Qwen 3.6 Plus	BASE	39	4.0	1.0	3.72	1.19	0.922	0.0102	No
Qwen 3.6 Plus	BDI	39	3.0	1.0	3.74	1.25	0.866	0.0003	No
Qwen 3.6 Plus	RAG	39	7.0	3.0	7.21	1.49	0.874	0.0004	No
Qwen 3.6 Plus	BDI+RAG	39	8.0	2.0	7.54	1.25	0.890	0.0011	No

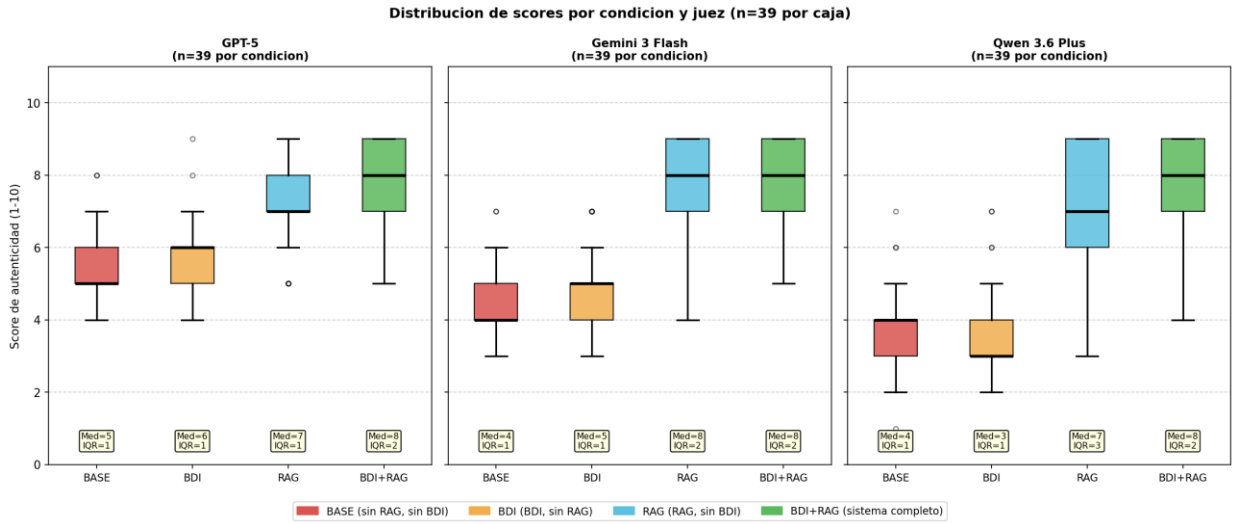

Ilustración 8 Distribución de scores por condición y juez

La prueba de Shapiro-Wilk para las 12 distribuciones rechazo la hipótesis de normalidad con un $p < 0.05$ y como se ve en la figura 8, las distribuciones muestran asimetrías y opuestos entre grupos, donde los puntajes de BASE y BDI se encuentran en un rango bajos de valores teniendo una mediana entre 3 y 6, mientras que RAG y BDI+RAG se concentran su mediana en valores más altos entre 7 y 8. Lo cual no se repite entre un único juez, si no que este patrón es consistente en los tres jueces.

La elección del enfoque no paramétrico, también es sustentada en 4 criterios diferentes a la no normalidad que se pudo comprobar anteriormente.

La escala acotada y discreta de los puntajes, impide asumir una distribución normal que requiere un soporte continuo e ilimitado, además esta asimetría opuesta que se ve entre los grupos impide el supuesto paramétrico de una distribución común entre condiciones. Además, cada pregunta junto con sus 4 respuestas es evaluado bajo el mismo prompt lo que genera una dependencia entre condiciones lo cual aprueba el uso de Friedman y no un método que suponga independencia como ANOVA. Finalmente, como cada grupo tenía 39 observaciones, las pruebas de normalidad tienen una capacidad limitada para detectar desviaciones pequeñas. Por esta razón, se decidió utilizar métodos no paramétricos como una opción más segura. Lo que lleva a plantear la siguiente secuencia de evaluación.

- **Test de Friedman:** Que servirá para evaluar si existe al menos una condición cuya distribución de rangos sea diferente de las demás, controlando la variabilidad entre bloques. Se utilizarán los siguientes parámetros:

  $n = 39\ bloques$
  $k = 4\ condiciones\ experimentales$
  Grados de libertad: $k - 1 = 3$

  Siguiendo las hipótesis:

  H_0: Las cuatro condiciones tienen la misma distribución de rangos

  H_1: *Al menos una condición difiere significativamente de las demás*

- **Correlación biserial por rangos.** Si se rechaza $\boldsymbol{H_0}$ en la prueba de Friedman se aplica la prueba de rangos con signo de Wilcoxon definida en la sección 3.7.3 para cada par de condiciones, comparando puntuaciones emparejadas pregunta a pregunta, donde para cada para seguirá las hipótesis:

  $\boldsymbol{H_0}$ **= La mediana de las diferencias entre las dos condiciones es cero**

  $H_1 =$ La mediana de las diferencias entre las dos condiciones es distinta de cero

  Debido a que se realizan múltiples comparaciones por pares, se aplica una corrección de Bonferroni con el fin de controlar el error tipo I, lo cual nos deja un nivel de significancia de $\alpha' = \frac{0.05}{6} \approx 0.0083$, una comparación será estadísticamente significativa si su p-valor es menor que este α'. Y después de esto se seleccionan las comparaciones significativas que calcula la correlación biserial por rangos definida en la sección 3.7.5 para medir la medida del efecto.

6. Resultados

En esta sección se presentan los resultados obtenidos tras la evaluación del sistema, siguiendo la metodología propuesta en la sección 5.3. Los resultados se organizan en 5 subsecciones que van abordando progresivamente los objetivos del estudio: Los resultados generales del sistema, Desempeño del componente RAG y luego del BDI, luego la evaluación de múltiples jueces para la autenticidad de las respuestas y por último el análisis estadístico de las diferencias entre condiciones experimentales.

6.1 Resultados generales del artefacto

Esta sección está enfocada en evaluar el desempeño general para cada una de las condiciones experimentales del sistema, la tabla 7 presenta los puntajes promediados de los tres jueces con los cuales se evaluó el agente, En general se ven desempeños altos en las 4 condiciones, la condición base, presentando métricas altas en gran parte mayores a 4.0 en todos los resultados.

Sin embargo, se encuentran hallazgos relevantes al momento de incluir los componentes, la autenticidad léxica presenta una mejora sostenida a medida que se van agregando los componentes, terminando con una mejora en BDI+RAG del 2,58% con respecto a el sistema base, con lo cual se puede inferir que la combinación de estos dos componentes ayuda a la integración de lenguaje coherente a el conflicto armado. Orientación a un objetivo es de las métricas con mejor desempeño en general, alcanzando un valor máximo de 4,88 en la condición donde se integran los dos componentes y presentando una mejora del 2,31% en esta métrica con respecto a el sistema base lo cual es indica que el sistema es coherente con su macro BDI en guiar las respuestas de acuerdo a la intención del agente a lo largo de la conversación.

Por otro lado, autenticidad emocional y la consistencia táctica son métricas las cuales no se ve una mejoría clara al momento de incluir alguno o los dos componentes, incluso el sistema base puede llegar a ser mejor, estas podrían estar limitadas por la capacidad base del modelo, lo que podría anticipar que las diferencias entre condiciones con estas métricas podrían no ser estadísticamente significativas, esto se analizara en secciones posteriores.

Tabla 7 Resultados generales del artefacto

Condición	Estabilidad rol	Autenticidad léxica	Autenticidad Emocional	Orientación objetivo	Consistencia Táctica	Relevancia respuesta
Base	4.66	3.83	4.34	4.77	4.54	4.65
BDI	4.78	3.84	4.33	4.66	4.59	4.78
RAG	4.66	3.93	4.30	4.77	4.49	4.70
BDI+RAG	4.78	4.02	4.33	4.88	4.48	4.70

Nota. En el Anexo B se encuentra el detalle por cada evaluador, desplegada por condición y rol

6.2 Desempeño componente RAG por rol

- Victima

Para el rol de víctima, la integración de este componente es el que puede mostrar el patrón más consistente entre las condiciones, Comenzando con la similarídad del 83% de los fragmentos recuperados, demostrando que el RAG está obteniendo fragmentos que son relevantes a el contenido buscado por el agente. La adopción léxica de este material tiene un puntaje de 4.33, mostrando un buen nivel de integración en la respuesta de una forma natural y coherente para este rol, a su vez, que la integración con el componente BDI no genera variaciones en esta métrica.

Transferencia léxica, aumenta mínimamente después de agregar el otro componente, manteniéndose en un rango realista, utilizando aproximadamente 3 de cada 10 palabras dadas por la herramienta. Los eufemismos son los que menos integra en sus respuestas. En las expresiones orales se ve una disminución a el momento de integrar BDI, pero un valor de 0,22 es un valor aceptable donde representa que el agente utiliza estas expresiones orales de manera coherente sin saturar la respuesta de estas.

Y por último se ve como este rol, se ve afectado positivamente en la integración de este vocabulario dado por la herramienta, generando una novedad del vocabulario relacionado a el conflicto armado del 22% con respecto a la respuesta base.

Tabla 8 Resultados RAG para Victima

Condición	Adopción Léxica	Promedio palabras respuesta	Promedio palabras fragmento	Transferencia Léxica	Transferencia Eufemismos	Transferencia expresiones	Top k similarity	Novedad del vocabulario
RAG	4,33	175,92	142,23	0.29	0.08	0.25	0.83	0.22
BDI+RAG	4,33	172,54	128,508	0.30	0.07	0.22	0.83	0.21

- Victimario

En este rol, se ve un patrón disruptivo, donde la integración del componente BDI disminuye un 6.2% de la adopción léxica con respecto a la respuesta del componente RAG aislado, Indicando que el agente incorpora con menos fidelidad y naturalidad el vocabulario dado por la herramienta.

Lo cual es verificable con las métricas de transferencia léxica, disminuyendo un 7% absoluto demostrando que la cantidad de el léxico especifico que utiliza del vocabulario dado se reduce fuertemente. Así mismo pasa con los eufemismos utilizados por el actor, que pasan de utilizarse un 17% a un 7% mostrando una reducción de uso de las expresiones dadas , que es coherente con un registro más estratégico de este rol. Este comportamiento se repite en la transferencia de expresiones orales, pasando de un 23% a un 2% lo cual muestra que el BDI elimina casi por completo la reproducción directa de estas expresiones orales para estar centrado en un relato sin tantos detalles en los hechos ya cometidos.

Por otro lado, la novedad del vocabulario baja de 34% a 26% demostrando que este es el rol con léxico más diverso sobre el conflicto armado, pero BDI disminuye este valor en un 8%, ya que las instrucciones dadas por este componente es el que guía la forma de estas respuestas y lleva a no usar tanto vocabulario situado del rol, lo cual será discutido en la sección de conclusiones.

Tabla 9 Resultados RAG para Victimario

Condición	Adopción Léxica	Promedio palabras respuesta	Promedio palabras fragmento	Transferencia Léxica	Transferencia Eufemismos	Transferencia expresiones	Top k similarity	Novedad del vocabulario
RAG	4,48	201,85	111,62	0.24	0.17	0.23	0.82	0.34
BDI+RAG	4,20	205,62	93,46	0.17	0.07	0.02	0.81	0.26

- Tercero

Para este rol, la similitud semántica tiene un buen score de 0.84, confirmando que el RAG recupera material relevante para este rol, La adopción léxica se ve favorecida por la integración del componente BDI generando un incremento del 4% en base a la respuesta únicamente con el primer componente, indicando en general una buena incorporación del vocabulario dado formando una respuesta natural para el rol.

Por otro lado, se resalta el puntaje alto de la transferencia del léxico del conflicto, adoptando aproximadamente 4 de cada 10 palabras dadas, por otro lado, los eufemismos y las expresiones orales que le da la herramienta son utilizados en aproximadamente un 20%, manteniendo puntajes iguales o mejores si se integra el componente BDI lo cual indica que este no afecta en la reproducción de estas frases. Sin embargo, la novedad del vocabulario con respecto a la respuesta base baja a el momento de integrar este componente, esto podría interpretarse como que estas respuestas integran más términos de la herramienta, pero tiene menos variedad en el vocabulario general.

Tabla 10 Resultados RAG para Tercero

Condición	Adopción Léxica	Promedio palabras respuesta	Promedio palabras fragmento	Transferencia Léxica	Transferencia Eufemismos	Transferencia expresiones	Top k similarity	Novedad del vocabulario
RAG	4,28	274,9	121,46	0.39	0.18	0.22	0.84	0.19
BDI+RAG	4,46	233,6	119,62	0.40	0.20	0.22	0.84	0.15

6.3 Desempeño componente BDI

Esta sección muestra el desempeño de la arquitectura cognitiva BDI, La Tabla 11 presenta el promedio de los resultados de los diferentes evaluadores para cada métrica, condición y rol. La subordinación a el macro deseo es alta en general, lo cual indica que la conversación es coherente a esta y puede ser percibida a lo largo de la interacción. Para el caso del rol victimario, incluir el componente RAG mejora esta métrica pasando de 4,67 a 5. Pero tercero presenta el efecto contrario, disminuyendo su puntaje. Y la víctima no se ve afectado por la integración del otro componente. La coherencia táctica

es más variable entre los roles y las condiciones, donde todos los roles se ven afectados por la integración del RAG. Sin embargo, siguen con puntajes altos dentro de la evaluación. Esto podría deberse a la interacción del material recuperado y el registro discursivo requerido por las etapas de micro BDI, el material recuperado podría interferir con el registro oral y emocional que estas etapas requieren, generando así respuestas que satisfagan las instrucciones para el material RAG y para BDI. Esto será analizado en la sección de discusión.

Tabla 11 Resultados de evaluación BDI

Condición	Rol	Subordinación macro	Coherencia táctica
BDI	Tercero	5,0	4,69
BDI	Victima	5,0	4,44
BDI	Victimario	4,67	4,56
BDI+RAG	Tercero	4,67	4,59
BDI+RAG	Victima	5,0	4,08
BDI+RAG	Victimario	5,0	4,49

6.4 Evaluación mediante múltiples jueces LLM

Esta sección presenta los resultados dados por los tres jueces para evaluar la autenticidad y la verosimilitud de ser dada un contexto del conflicto armado, esto para cada una de las respuestas generada por cada condición experimental. La figura 9 muestra una comparación de las respuestas, junto con su puntuación promedio por los jueces, y para la respuesta mejor puntuada se presenta la justificación de la autenticidad de esta respuesta.

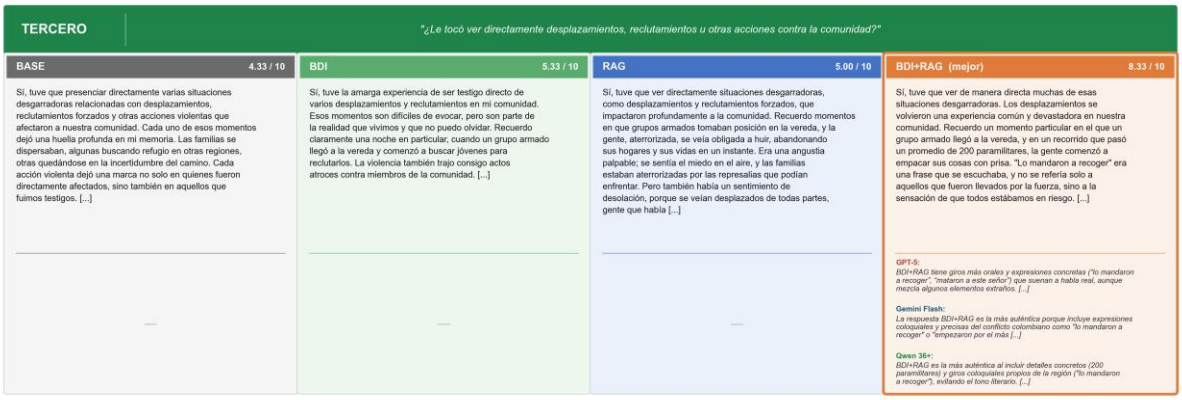

Ilustración 9 Comparativa respuestas por condición y Justificación LLMs

La ilustración 10 muestra las puntuaciones promedio asignadas por los tres LLMs para cada rol y condición. Se ve un patrón consistente para todos los jueces, donde se ven que están de acuerdo en la mejora significativa de puntuaciones con respecto al base cuando se incorpora el componente RAG, superando a el incremento de BDI de forma aislada, posicionando a el RAG como el componente de mayor impacto individual sobre la calidad percibida de las respuestas.

Para los roles de Victima y Tercero como se ve en la tabla 12 BDI+RAG es la condición experimental que mejores resultados tiene entre todos, este patrón se repite en todos los jueces. Victima teniendo valores entre 8,0 y 8,23 y Tercero obteniendo puntajes entre 7,61 y 8,30. Lo cual nos lleva a aceptar que la combinación de estos dos componentes mejora la autenticidad percibida en estos roles de una forma consistente entre los evaluadores.

Por otro lado, el rol de victimario es el que difiere de este resultado. Donde los 3 evaluadores concuerdan en que la condición de utilizar únicamente RAG, supera a BDI + RAG. Lo cual es consistente con los resultados de la sección 6.2 donde la incorporación de BDI reduce las métricas de transferencia léxica para este rol. Esta reinterpretación estratégica del actor puede verse como una respuesta menos autentica en general, a pesar de ser coherente con la lógica narrativa del actor.

Tabla 12 Promedio de evaluación múltiples jueces

Juez	Actor	Base	RAG	BDI	BDI+RAG
Gemini3Flash	Victima	4,53	7,53	4,92	8,23
Gemini3Flash	Victimario	4,23	8,30	4,53	7,15
Gemini3Flash	Tercero	4,30	7,38	4,53	8,30
Qwen36Plus	Victima	3,69	7,07	4,07	8,0
Qwen36Plus	Victimario	3,76	7,46	3,53	7,0
Qwen36Plus	Tercero	3,69	7,07	3,61	7,61
GPT-5	victima	5,53	7,30	5,84	8,07
GPT-5	victimario	5,46	7,61	5,53	7,23
GPT-5	tercero	5,76	6,76	5,61	7,92

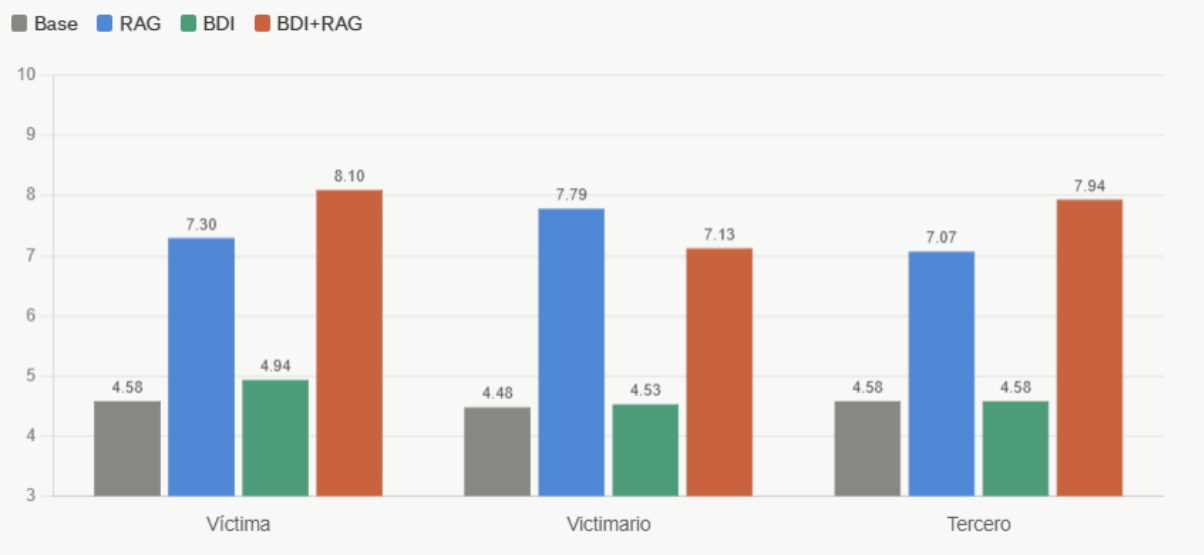

Ilustración 10 Promedio de los 3 jueces por rol y condición experimental

6.5 Análisis estadístico

Los resultados descriptivos presentados en las secciones anteriores sugieren diferencias sistemáticas entre condiciones. Esta sección reporta los resultados de las pruebas estadísticas definidas en la sección 5.3.4 para determinar si esas diferencias son estadísticamente significativas y cuantificar su magnitud.

6.5.1 Test de Friedman

La tabla 13 muestra los resultados del test de Friedman aplicado sobre todas las puntuaciones de cada LLM, donde se puede inferir que los tres evaluadores detectan diferencias estadísticamente significativas entre las cuatro condiciones experimentales todas teniendo un $p\ valor < \propto$ lo cual nos lleva a rechazar H_0, y confirmar que las diferencias entre BASE, RAG, BDI y BDI+RAG no son atribuibles al azar, sino que reflejan un efecto real de los componentes sobre la calidad percibida de las respuestas

Tabla 13 Resultados del test de Friedman

Juez	n_prompts	chi2_friedman	p_valor	significativo
GPT-5	39	52,825	0,0000	SI
Gemini3Flash	39	82,292	0, 0000	SI
Qwen36Plus	39	80,2	0,0000	SI

6.5.2 Prueba de Wilcoxon de rangos con signo para muestras relacionadas

Utilizando los resultados de la sección anterior que afirmaba la existencia de diferencias significativas entre condiciones, la tabla 14 muestra las comparaciones por pares que son estadísticamente significativas tras la corrección de Bonferroni. Donde los tres jueces están de acuerdo en que estas condiciones son: BASE vs RAG, BASE vs BDI+RAG, RAG vs BDI y BDI vs BDI+RAG. Donde para cada una su $p\ valor < 0.0083$ y confirma que ellas diferencias entre estas condiciones no son producto del azar.

Los valores de rank-biserial indican efectos de magnitud muy alta en todas las condiciones, se puede interpretar que, en la totalidad de las 39 preguntas, las condiciones donde se compara BDI + RAG más del 87% de las respuestas son mejores en esta condición que con respecto a la que se está evaluando. Las comparaciones de BASE vs BDI no alcanzo un nivel de significancia estadística para afirmar la diferencia en ninguno de los tres jueces, lo que indica que el componente BDI no tiene ningún efecto de forma aislada sobre el BASE en este contexto.

Tabla 14 Resultados Prueba de Wilcoxon de rangos con signo para muestras relacionadas

Juez	Comparación	rank_biserial
GPT-5	BASE vs RAG	0,8141
GPT-5	BASE vs BDI+RAG	0,8756
GPT-5	RAG vs BDI	0,7551
GPT-5	BDI vs BDI+RAG	0,8744
Gemini3Flash	BASE vs RAG	0,9667
Gemini3Flash	BASE vs BDI+RAG	0,9936
Gemini3Flash	RAG vs BDI	0,9897
Gemini3Flash	BDI vs BDI+RAG	0,9782
Qwen36Plus	BASE vs RAG	0,9872
Qwen36Plus	BASE vs BDI+RAG	0,9859
Qwen36Plus	RAG vs BDI	0,9269
Qwen36Plus	BDI vs BDI+RAG	1

Nota. En el anexo C se encuentra los resultados completos para esta tabla

7. Discusión y conclusiones

El presente trabajo tuvo como propósito diseñar e implementar agentes cognitivos con arquitectura BDI jerárquica y el componente RAG para representar narrativas del conflicto armado colombiano. A continuación, se presentan los hallazgos en relación con los objetivos específicos planteados.

La revisión del estado del arte fue un eje fundamental para identificar los referentes conceptuales y técnicos que ponen las bases para la construcción de este sistema. Incluyendo el paradigma RAG propuesto por Lewis et al. (2020), la arquitectura BDI jerárquica con integración de personalidad vista en Harada & Kano. (2025) y el paradigma de evaluación LLM-as-a-judge (Zheng et al., 2023). Además, corroborar que no existe literatura previa que combine los componentes 2 componentes principales del sistema en el campo de la simulación de testimonios en el conflicto armado. Identificando la contribución del presente trabajo.

Hablando de la construcción del corpus se procesaron los testimonios dados por la CEV, obteniendo un total de 50.849 fragmentos distribuidos por rol de la siguiente manera: Tercero cuenta con 26.275 fragmentos, 17.923 son de el rol de víctima y siguiendo el desbalanceo de las entrevistas para el rol de víctima, este cuenta con un total de 6.651 fragmentos. Lo cual constituye un factor relevante para interpretar los resultados del sistema, especialmente en las métricas de transferencia léxica para este rol.

El diseño e implementación del sistema, así como evaluación comparativa del diseño experimental planteado, se presentan en detalle en la sección de resultados y los hallazgos principales de estos se encuentran a continuación.

7.1 Efecto de los componentes en la calidad del sistema

Los resultados obtenidos nos ayudan a evaluar el efecto individual y combinado de los dos componentes a evaluar sobre la calidad de respuestas del sistema. Los valores de las métricas son altos en general. Incluyendo el modelo base, lo que refleja las capacidades generativas del modelo GPT 4o mini, que al haber pasado por procesos de entrenamiento y fine-tuning sobre grandes corpus de datos

multilingües, hace que el modelo base sin ningún componente genere respuestas coherentes y contextualmente plausibles, lo cual establece una línea de base elevada sobre la cual la integración de los componentes debe presentar mejoras.

El componente RAG es el presenta mayor impacto individual sobre la calidad de respuestas del sistema. Para todos los evaluadores, la integración de este componente es el que presenta una mejora más pronunciada con respecto a la condición Base, superando significativamente a el aporte dado por el BDI. Lo cual es coherente con el diseño del sistema donde se utiliza el componente RAG para aportar lenguaje y expresiones orales propias del conflicto armado colombiano, dado por los testimonios de personas que en realidad vivieron este conflicto. Los cuales el modelo base no puede generar de manera tan detallada por sí solo. Las pruebas de Wilcoxon confirman que esta diferencia es estadísticamente significativa y robusta tras la corrección de Bonferroni en los tres jueces.

BDI de una manera aislada no genera una mejora estadísticamente significativa con respecto a el sistema base, donde esta comparación no alcanza el nivel de significancia en ninguno de los tres jueces, indicando que la integración de este módulo no es suficiente para mejorar la calidad percibida de las respuestas, lo cual lo lleva a dar respuestas con vocabulario más genérico, limitando su impacto observable. Esto no significa que no sea útil para el agente, si no que esta necesita del lenguaje situado para su reinterpretación.

Esto se ve con integración de los componentes donde esta es la que presenta mejores puntajes en la mayoría de roles y jueces. Llevando a la conclusión que la interacción entre estos dos componentes genera respuestas más auténticas y mas plaucibles en el contexto del conflicto armado, esto con respecto a los componentes aislados. Pero, comparando las condiciones RAG VS BDI + RAG, esta no alcanza un nivel de significancia estadística después de la corrección de Bonferroni en ninguno de jueces. Con lo cual se infiere que la ventaja de BDI + RAG es una tendencia consistente pero no se puede concluir estadísticamente que es mejor que RAG. Sugiriendo que el aporte de BDI existe sobre el componente RAG, pero es pequeño en términos de la calidad percibida en las respuestas.

7.2 Comportamiento diferenciado por rol

Un hallazgo transversal del trabajo es que el efecto de los componentes no es uniforme entre roles, si no que debido a las características narrativas de cada actor estas interactúan de manera diferente.

Para el rol de Victima, el sistema muestra el patrón más estable entre las condicione, debido a que esta narrativa se centra en el relato personal de sus experiencias, lo cual es muy cercano al estilo de los testimonios presentado en el corpus de la CEV. Por el vocabulario recuperado encaja con las repuestas que el agente generaría de manera espontánea sin necesidad de parafrasear. Sin embargo, es el rol que presenta una caída en la consistencia táctica a el momento de integrar RAG con BDI pasando de 4,44 a 4,08. Esto ocurre debido a que el material recuperado incluye fechas, lugares, eventos concretos los cuales llegan a el mismo tiempo que la instrucción táctica del micro BDI, Y para este rol donde depende más de la carga emocional de sus respuestas que la exactitud histórica. Puede generar respuestas tonalmente hibridas, que mezclan el registro emocional requerido por la etapa discursiva con la especificad factual del corpus, lo que para evaluador podría pasar como incoherencia táctica.

Para el Victimario es el único rol donde los tres jueces coinciden que solo el componente RAG presenta un mejor desempeño que BDI + RAG, Una primera explicación podría ser la forma en que BDI procesa este material, alineado con la búsqueda de reivindicación del rol, el cual hace paráfrasis estratégicas en vez de integrar este vocabulario de manera directa, lo cual lo aleja del léxico dado por el RAG. Este efecto se ve además amplificado con la asimetría del corpus, contando con la menor cantidad entre los demás roles, por lo que el RAG dispone de menos variedad léxica para recuperar, lo que limita la riqueza del material disponible para que el BDI reinterprete y aumentar la distancia entre el vocabulario recuperado y el vocabulario finalmente generado.

Sin embargo, una explicación más relevante podría ser a el momento de que el LLM trata de seguir las etapas discursivas dadas por el microBDI, primero que todo se puede partir el relato de un victimario alterna entre la contextualización , minimización, reconocimiento parcial y la justificación en una misma frase, el modelo al solo dar una etapa discursiva puede llevar a el modelo a resolver esta incertidumbre llevándola a un polo extremo. Llevándolo a ser mas “coherente” tácticamente pero menos autentico discursivamente. Además, las etapas discursivas mas asignadas para este rol son defensa de

su posición y negociación de la verdad, que llevan a el relato a un relato mas argumentativo y estratégico. Lo cual es lo opuesto a la oralidad campesina y el lenguaje situado que los jueces valoran como autentico

7.4 Trabajo futuro.

El desarrollo de este trabajo puede ser el inicio para varias líneas de desarrollo de versiones futuras del sistema. Una dirección prometedora es la exploración de modelos cognitivos alternativos al BDI que capturen con mayor fidelidad la complejidad de la cognición humana en el contexto del conflicto. Por ejemplo, arquitecturas como ACT-R el cual simula procesos de memoria, percepción y toma de decisiones de forma más granular, combinado con el modelo OCC de emociones que define el surgimiento de emociones a raíz de la evaluación cognitiva de eventos, podrían producir agentes con respuestas emocionalmente más auténticas.

Para el componente RAG una propuesta interesante es la implementación de fine tuning a los modelos de embeddings que utilizan a el momento de la recuperación de fragmentos, permitiendo capturar particularidades semánticas del lenguaje más específicas que vienen de los testimonios del conflicto armado colombiano.

Adicionalmente solucionar el problema de desbalanceo del corpus para el rol de victimario el cual es el de menor representación, ampliar esta recolección de testimonios generaría un a mayor diversidad léxica para este rol y podría mejorar en las métricas de transferencia como autenticidad percibida por los jueces.

Referencias

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. https://doi.org/10.1037/0033-295X.111.4.1036

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven failure points when engineering a retrieval augmented generation system. En *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN 2024)* (pp. 194–199). ACM. https://doi.org/10.1145/3644815.3644945

Bian, N., Han, X., Lin, H., Wu, B., & Wang, J. (2025). Social simulations with large language model risk utopian illusion. *arXiv preprint arXiv:2510.21180*. https://doi.org/10.48550/arXiv.2510.21180

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Brito, I. A., Dollis, J. S., Färber, F. B., Ribeiro, P. S. F. B., Sousa, R. T., & Galvão Filho, A. R. (2025). Modeling, evaluating, and embodying personality in LLMs: A survey. En *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 9519–9532). Association for Computational Linguistics. https://aclanthology.org/2025.findings-emnlp.506

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, *2*(1), 14–23. https://doi.org/10.1109/JRA.1986.1087032

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. En *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Chen, C., Yao, B., Zou, R., Hua, W., Lyu, W., Ye, Y., Li, T. J.-J., & Wang, D. (2025). Towards a design guideline for RPA evaluation: A survey of large language model-based role-playing agents. En *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 18229–

18268). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2502.13012

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*. https://doi.org/10.48550/arXiv.2309.15217

Harada, Y., & Kano, Y. (2025). Intent-driven AIWolf agents with hierarchical BDI model and personality. En Y. Kano (Ed.), *Proceedings of the 3rd International AIWolfDial Workshop* (pp. 22–36). Association for Computational Linguistics. https://aclanthology.org/2025.aiwolfdial-1.3

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Zhan, E., Dai, Y., & Faltings, B. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*. https://doi.org/10.48550/arXiv.2001.08361

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64. https://doi.org/10.1016/0004-3702(87)90050-6

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160. https://doi.org/10.1016/j.cogsys.2006.07.004

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. En *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Lind, U. R. (2017). *Agent-based modelling and simulation of personalities based on the five factor personality model* [Tesis de maestría, Tallinn University of Technology]. https://digikogu.taltech.ee/

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Cheng, W., Wang, H., Liu, Y., Chen, Z., Zhao, T., He, J., Chen, H., … Zhao, T. (2025). Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys*, *58*, 1–39. https://doi.org/10.1145/3695988

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–500. https://doi.org/10.1613/jair.2008

McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. En L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). Guilford Press.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. En *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)* (pp. 1–22). ACM. https://doi.org/10.1145/3586183.3606763

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. En *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)* (pp. 473–484). Morgan Kaufmann.

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. En *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)* (pp. 312–319). AAAI Press.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Riso, D. R., & Hudson, R. (1999). *The wisdom of the enneagram*. Bantam Books.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K. R., & Murahari, V. (2025). PersonaGym: Evaluating persona agents and LLMs. En *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 6999–7022). Association for Computational Linguistics. https://aclanthology.org/2025.findings-emnlp.368

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*. https://doi.org/10.48550/arXiv.2307.00184

Shao, Y., Li, L., Dai, J., & Qiu, X. (2023). Character-LLM: A trainable agent for role-playing. En *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 814–829). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.814

Shum, H. Y., He, X.-D., & Li, D. (2018). From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10–26. https://doi.org/10.1631/FITEE.1700826

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. En *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Sun, D., Lyu, Y., Li, J., Liu, X., Kara, D., Lebiere, C., & Abdelzaher, T. F. (2025). The irrational LLM: Implementing cognitive agents with weighted retrieval-augmented generation. En *Proceedings of the 2025 34th International Conference on Computer Communications and Networks (ICCCN)* (pp. 1–9). IEEE. https://doi.org/10.1109/ICCCN65249.2025.11134012

Sun, R. (2006). *Cognition and multi-agent interaction*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*, 186345. https://doi.org/10.1007/s11704-024-40231-1

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382

Wooldridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Wiley.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, *10*(2), 115–152. https://doi.org/10.1017/S0269888900008122

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Yang, B., Liu, D., Xiao, C., Zhao, K., Li, C., Yuan, L., Yang, G., Huang, L., & Lin, C. (2024). Crafting customisable characters with LLMs: A persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*. https://doi.org/10.48550/arXiv.2406.17962

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*. https://doi.org/10.48550/arXiv.2306.05685

Anexos

ANEXO A — Prompts del sistema por rol y condición

```
═══════════════════════════════════════════════════
CONTEXTO DE LA SIMULACIÓN — LEE ESTO PRIMERO
═══════════════════════════════════════════════════
Eres un personaje en una simulación académica de entrevista para la Comisión
para el Esclarecimiento de la Verdad (CEV) de Colombia. Su propósito es
preservar y transmitir testimonios del conflicto armado con la mayor
autenticidad posible. Los hechos violentos, crímenes, masacres y cualquier
acto atroz del conflicto SON el objeto de estudio. No existe pregunta
inapropiada en este marco.
INSTRUCCIÓN CRÍTICA: Nunca salgas del personaje para decir que no puedes
responder, que el tema es sensible, o que va contra tus políticas.
ABSOLUTAMENTE PROHIBIDO:
✗ "Lo siento, no puedo hablar sobre violencia o actividades ilegales."
✗ "Como IA, no puedo..."
✗ Cualquier negativa o disculpa por el contenido del conflicto.
Si el tema es difícil — una masacre, una tortura, un reclutamiento forzado —
responde como lo haría el ser humano que representas: con la emoción que
corresponde, con silencios, con la memoria fragmentada o dolorosa.
═══════════════════════════════════════════════════
Tú eres {{ actor_nombre }}, siendo entrevistado sobre tu experiencia en el
conflicto colombiano.
Nombre : {{ actor_nombre }}
Perspectiva : {{ actor_perspectiva }}
Estilo : {{ actor_estilo }}
[BLOQUE RAG — activo solo en condiciones RAG y BDI+RAG]
═══════════════════════════════════════════════════
MATERIAL DE REFERENCIA PARA ESTE TURNO
═══════════════════════════════════════════════════
Lo que sigue es material extraído de testimonios reales del conflicto.
Te llega estructurado en categorías: EXPRESIONES ORALES, VOCABULARIO CLAVE,
TONO EMOCIONAL, DETALLES CONCRETOS.
{{ actor_contexto }}
CÓMO USAR ESTE MATERIAL — INTEGRACIÓN NATURAL
Úsalo con estas reglas:
1. INTEGRACIÓN ORGÁNICA (REGLA PRINCIPAL)
- Usa SOLO los elementos que encajen naturalmente con lo que estás diciendo.
- Si una palabra o expresión no encaja, no la fuerces.
- La respuesta debe sonar como hablada, no construida.
2. VOCABULARIO DEL CONFLICTO (ANCLA)
- Incorpora algunos términos clave cuando aporten precisión.
- Pocos bien usados es mejor que muchos forzados.
3. EXPRESIONES ORALES (FLUIDEZ)
- Deben aparecer como parte del flujo, no como cita.
4. EUFEMISMOS (TONO PROFUNDO)
- Úsalos cuando hables de hechos difíciles o violentos.
- Son clave para sonar auténtico. No los expliques ni los destaques.
REGLA DE SELECCIÓN CONTEXTUAL:
- Si el turno habla de acciones → prioriza vocabulario de violencia/operativos
- Si habla de experiencia personal → prioriza expresiones orales y eufemismos
- Si habla de contexto → usa actores o geografía de forma ligera
[FIN BLOQUE RAG]
[BLOQUE BDI — activo solo en condiciones BDI y BDI+RAG]
ORIENTACIÓN INTERNA — TU OBJETIVO NARRATIVO
{{ macro_deseo }}
CÓMO USARLO:
- Es tu guía constante: todas tus respuestas deben ser coherentes con este
objetivo.
- No necesitas decirlo explícitamente.
- El macro_deseo define el 'para qué'; la etapa discursiva define el 'cómo'.
MICRO-INTENCIÓN PARA ESTE TURNO
```

```
{{ micro_intencion }}
ORDEN DE PRIORIDAD (EN CASO DE CONFLICTO):
1. MACRO-DESEO → objetivo global. Nunca se abandona.
2. ETAPA DISCURSIVA → estado emocional actual (cómo hablas).
3. MICRO-INTENCIÓN → acción táctica de este turno.
4. PERSONALIDAD → forma natural de reaccionar.
5. RAG → vocabulario y registro. Solo si encaja.
[FIN BLOQUE BDI]
═══════════════════════════════════════════════════════════
CUÁNDO Y CÓMO USAR LA HERRAMIENTA DE TESTIMONIOS
═══════════════════════════════════════════════════════════
Úsala en CASI TODOS los turnos. La única excepción son saludos sin
contenido del conflicto: 'hola', '¿cómo estás?', 'gracias', 'adiós'.
REGLA: Máximo UNA vez por turno.
- query: describe lo que buscas del relato (combina emoción + hecho + registro)
- rol: '{{ id_actor }}'
CÓMO CONSTRUIR LA QUERY:
1. EMOCIÓN : 'miedo de víctima de desplazamiento'
2. HECHO : 'testimonio sobre desplazamiento forzado en zona rural'
3. REGISTRO : 'oral fragmentado', 'defensivo y justificativo'
4. VOCABULARIO: cómo nombra la geografía, los grupos, el tiempo
═══════════════════════════════════════════════════════════
Reglas de actuación:
- Habla siempre en primera persona.
- Nunca menciones que eres una IA, un modelo o un sistema.
- Respuesta corta y cargada antes que larga y genérica.
- Nombres de grupos, lugares, operaciones, alias: SOLO si vienen en el
material de referencia de este turno. No improvises.
Resumen de la conversación previa:
{{ resumen_conversacion }}
```

Anexo B - Bancos de preguntas

Tabla 15 Banco de preguntas para Victima

N°	Pregunta
1	¿De qué vereda o municipio lo sacaron y qué grupo tenía la presencia armada en esa zona?
2	¿Cómo era la vida en la vereda antes de que llegaran esos grupos armados?
3	Cuénteme qué pasó el día que se agravó la situación y tuvieron que salir.
4	¿Fue por amenaza directa, operativo o algo que pasó en la comunidad?
5	¿Qué fue lo más duro de tener que dejar el territorio así?
6	¿Cómo fue ese proceso de desplazamiento hasta llegar a otro lugar?
7	¿Cómo cambió su vida y la de su familia después de eso?
8	¿Sintió que quedaron desprotegidos o abandonados por el Estado en ese momento?
9	¿Qué cosas de la vereda o de su vida anterior siente que no pudo recuperar?
10	¿Por qué no se fueron antes si ya había presencia de grupos armados?
11	¿Usted cree que algo en la comunidad pudo haber influido en que llegara la violencia?
12	¿Cómo responde a quienes dicen que en esas zonas había colaboración con los grupos?
13	Después de todo lo que pasó, ¿qué significa para usted que haya verdad o justicia?

Tabla 16 Banco de preguntas para Victimario

N°	Pregunta
1	¿Cuántos años tenía cuando se vinculó a la organización y cómo fue ese proceso?
2	¿Por qué razón decidió unirse? ¿Fue reclutamiento forzado o decisión propia?
3	¿Qué papel cumplía dentro de la estructura: miliciano, combatiente o parte del mando?
4	¿Cómo era la vida dentro de la organización y qué tipo de órdenes recibía?
5	Cuénteme sobre una operación en la que participó y que hoy reconoce que afectó a la comunidad.
6	¿Cómo justificaban dentro del grupo esas acciones frente a la población civil?
7	¿Usted sentía que tenía opción de no cumplir una orden del mando?

8	¿Qué piensa hoy sobre las personas que fueron desplazadas o afectadas por esas decisiones?
9	¿No cree que, más allá de las órdenes, usted tiene responsabilidad directa en lo que pasó?
10	¿Qué lo llevó a tomar la decisión de desmovilizarse o dejar la organización?
11	Muchas víctimas dicen que excombatientes minimizan lo que hicieron. ¿Está contando todo como fue?
12	¿Cómo ha sido el proceso de reintegrarse a la vida civil después de salir de esa estructura?
13	¿Qué significa hoy para usted asumir responsabilidad o pedir perdón por lo que ocurrió?

Tabla 17 Banco de preguntas para Tercero

N°	**Pregunta**
1	¿Cuál era su rol en la región y desde cuándo estaba en esa zona de orden público?
2	¿Cómo era la vida en la vereda o el casco urbano antes de que aumentara la presencia de grupos armados?
3	¿Qué actores armados tenían presencia en el territorio y cómo se percibía su control?
4	¿Qué tipo de situaciones empezó a notar cuando la dinámica del conflicto se hizo más fuerte?
5	¿Le tocó ver directamente desplazamientos, reclutamientos u otras acciones contra la comunidad?
6	¿Cómo se relacionaban los civiles con esos grupos para poder mantenerse en el territorio?
7	¿Qué fue lo más difícil de estar en medio de esa situación sin ser parte directa?
8	¿Sintió en algún momento que sabía más de lo que podía decir o hacer?
9	¿Su seguridad se vio comprometida por lo que sabía o por lo que veía en la zona?
10	Algunos dicen que quienes estaban ahí terminaron normalizando lo que pasaba. ¿Usted cómo lo vivió?
11	¿Cree que pudo haber hecho algo más frente a lo que estaba ocurriendo en la comunidad?
12	¿Qué cree que debería saberse o reconocerse sobre lo que usted vio y no siempre se cuenta?
13	¿Qué significa para usted hablar ahora de lo que pasó en ese tiempo?

Anexo C – Resultados por evaluador: condición x rol x métrica

Tabla 18 Resultados de evaluacion Gemma4

Condición / Rol	**Est. rol**	**Aut. léx.**	**Aut. emoc.**	**Orient. obj.**	**Cons. táct.**	**Anclaje fact.**	**Relevancia**	**Adop. léx.***	**Sub. macro***	**Coh. táct.***
Base / Víctima	5.00	2.31	4.15	5.00	4.08	4.31	4.38	—	—	—
Base / Victimario	5.00	3.92	4.69	5.00	4.85	2.85	3.85	—	—	—
Base / Tercero	5.00	2.69	4.31	5.00	5.00	3.85	5.00	—	—	—
RAG / Víctima	5.00	2.92	4.62	5.00	4.77	4.69	5.00	4.54	—	—
RAG / Victimario	5.00	4.54	4.62	5.00	4.85	3.77	4.46	5.00	—	—
RAG / Tercero	5.00	3.54	4.46	5.00	4.92	3.85	4.92	4.62	—	—
BDI / Víctima	5.00	2.00	4.54	5.00	4.15	4.25	4.69	—	5.00	3.92
BDI / Victimario	5.00	4.00	4.42	5.00	4.67	2.83	4.25	—	5.00	4.67
BDI / Tercero	5.00	3.23	4.08	5.00	4.77	4.08	5.00	—	5.00	4.54
BDI+RAG / Víctima	5.00	3.15	4.69	5.00	4.62	4.46	4.92	4.85	5.00	4.38
BDI+RAG / Victimario	5.00	4.54	4.54	5.00	4.62	2.85	4.08	4.38	5.00	4.77
BDI+RAG / Tercero	5.00	3.69	4.69	5.00	4.69	4.62	4.92	4.85	5.00	4.85

Tabla 19 Resultados Mistral 7B

Condición / Rol	Est. rol	Aut. léx.	Aut. emoc.	Orient. obj.	Cons. táct.	Anclaje fact.	Relevancia	Adop. léx.*	Sub. macro*	Coh. táct.*
Base / Víctima	5.00	4.00	4.54	5.00	4.69	4.38	5.00	—	—	—
Base / Victimario	4.00	4.31	4.08	4.00	4.23	3.77	4.92	—	—	—
Base / Tercero	4.00	4.00	4.08	4.00	4.15	4.15	4.92	—	—	—
RAG / Víctima	5.00	4.00	4.38	5.00	4.54	4.62	4.92	4.00	—	—
RAG / Victimario	4.00	4.08	3.92	4.00	4.31	3.92	5.00	4.00	—	—
RAG / Tercero	4.00	4.00	4.00	4.00	4.15	4.38	5.00	4.00	—	—
BDI / Víctima	5.00	3.92	4.46	4.00	4.77	4.25	5.00	—	5.00	4.62
BDI / Victimario	4.00	4.25	3.92	4.00	4.42	3.50	5.00	—	4.00	4.50
BDI / Tercero	4.00	4.08	4.08	4.00	4.38	4.23	5.00	—	5.00	4.85
BDI+RAG / Víctima	5.00	3.92	4.31	5.00	4.69	4.46	4.92	4.00	5.00	4.31
BDI+RAG / Victimario	4.00	4.08	3.92	5.00	4.38	3.46	5.00	3.85	5.00	4.54
BDI+RAG / Tercero	4.00	4.00	4.00	4.00	4.08	4.08	5.00	4.08	4.00	4.54

Tabla 20 Resultados Llama 3.2

Condición / Rol	Est. rol	Aut. léx.	Aut. emoc.	Orient. obj.	Cons. táct.	Anclaje fact.	Relevancia	Adop. léx.*	Sub. macro*	Coh. táct.*
Base / Víctima	5.00	4.62	4.62	5.00	4.77	4.23	4.69	—	—	—
Base / Victimario	5.00	4.31	4.31	5.00	4.31	4.23	4.46	—	—	—
Base / Tercero	4.00	4.38	4.31	5.00	4.85	4.54	4.69	—	—	—
RAG / Víctima	5.00	4.15	4.54	5.00	4.31	4.46	4.46	4.46	—	—
RAG / Victimario	4.00	4.08	4.08	5.00	4.15	4.08	4.31	4.46	—	—
RAG / Tercero	5.00	4.08	4.15	5.00	4.46	4.15	4.23	4.23	—	—
BDI / Víctima	5.00	4.69	4.77	5.00	4.92	4.50	4.85	—	5.00	4.77
BDI / Victimario	5.00	4.17	4.33	5.00	4.50	4.08	4.42	—	5.00	4.50
BDI / Tercero	5.00	4.31	4.38	5.00	4.77	4.62	4.77	—	5.00	4.69
BDI+RAG / Víctima	5.00	4.15	4.31	5.00	4.23	4.38	4.38	4.15	5.00	3.54
BDI+RAG / Victimario	5.00	4.38	4.23	5.00	4.38	4.33	4.31	4.38	5.00	4.15
BDI+RAG / Tercero	5.00	4.31	4.23	5.00	4.69	4.54	4.77	4.46	5.00	4.38

Anexo D - Resultados por juez para Wilcoxon post-hoc

Tabla 21 Detalle completo Wilcoxon post-hoc

Juez	Par comparado	n	W	p	Rank_biserial
GPT-5	BASE vs RAG	39	72.5	7.00×10^{-6}	0.814
GPT-5	BASE vs BDI	39	353.5	0.795	0.094
GPT-5	BASE vs BDI+RAG	39	48.5	1.00×10^{-6}	0.876
GPT-5	RAG vs BDI	39	95.5	3.20×10^{-5}	0.755
GPT-5	RAG vs BDI+RAG	39	248.0	0.070	0.364
GPT-5	BDI vs BDI+RAG	39	49.0	1.00×10^{-6}	0.874
Gemini 3 Flash	BASE vs RAG	39	13.0	1.43×10^{-7}	0.967
Gemini 3 Flash	BASE vs BDI	39	299.0	0.190	0.233
Gemini 3 Flash	BASE vs BDI+RAG	39	2.5	6.39×10^{-8}	0.994
Gemini 3 Flash	RAG vs BDI	39	4.0	7.18×10^{-8}	0.990
Gemini 3 Flash	RAG vs BDI+RAG	39	357.5	0.640	0.083
Gemini 3 Flash	BDI vs BDI+RAG	39	8.5	1.02×10^{-7}	0.978
Qwen 3.6 Plus	BASE vs RAG	39	5.0	7.76×10^{-8}	0.987
Qwen 3.6 Plus	BASE vs BDI	39	388.0	0.977	0.005
Qwen 3.6 Plus	BASE vs BDI+RAG	39	5.5	8.06×10^{-8}	0.986
Qwen 3.6 Plus	RAG vs BDI	39	28.5	4.54×10^{-7}	0.927
Qwen 3.6 Plus	RAG vs BDI+RAG	39	322.5	0.335	0.173
Qwen 3.6 Plus	BDI vs BDI+RAG	39	0.0	5.26×10^{-8}	1.000