Hermenēus
Revista de traducción e interpretación

EDICIONES Universidad Valladolid

IA generativa y traducción del habla coloquial: estudio comparativo de cuatro sistemas

Generative AI and the translation of colloquial speech: A comparative study of four systems

JAHIRO SAMAR ANDRADE PRECIADO
Universidad Autónoma de Baja California. Facultad de Idiomas Mexicali. Av. Monclova, s/n. Ex-Ejido Coahuila. 21360 Mexicali, B.C. México
Dirección de correo electrónico: samar.andrade@uabc.edu.mx
ORCID: https://orcid.org/0000-0001-6064-6426

IOANA CORNEA
Universidad Nacional Autónoma de México. Av. Universidad, 3000. C.U. Coyoacán. 04510 Ciudad de México, CDMX. México.
Dirección de correo electrónico: ioana.cornea@enallt.unam.mx
ORCID: https://orcid.org/0009-0009-9018-9196

HÉCTOR JAVIER ANDRADE PRECIADO
Universidad Autónoma de Baja California. Facultad de Idiomas, Ensenada. Juan, Blvd. Zertuche, Carlos Pacheco 7. 22890 Ensenada, Baja California. México.
Dirección de correo electrónico: hsanchez13@uabc.edu.mx
ORCID: https://orcid.org/0000-0002-4949-8001

Recibido/Received: 29/12/2025. Aceptado/Accepted: 7/7/2026.
Cómo citar / How to cite: Andrade Preciado, Héctor Javier, Ioana Cornea y Jahiro Samar Andreade Preciado, «IA generativa y traducción del habla coloquial: estudio comparativo de cuatro sistemas», *Hermēneus. Revista de Traducción e Interpretación*, 28 (2026): pp. 41-74.
DOI: https://doi.org/10.24197/exbewg03

Artículo de acceso abierto distribuido bajo una Licencia Creative Commons Atribución 4.0 Internacional (CC-BY 4.0). / Open access article under a Creative Commons Attribution 4.0 International License (CC-BY 4.0).

Resumen: En Los modelos extensos de lenguaje (MEL) como ChatGPT, DeepSeek, Gemini y Claude han mostrado avances significativos en la traducción automática; no obstante, la traducción del habla coloquial sigue siendo un reto debido a la presencia de expresiones idiomáticas, jerga marcadores pragmáticos y referencias culturales. Este estudio compara su desempeño en la traducción del inglés al español de cuarenta expresiones coloquiales, organizadas en cuatro categorías. Se adoptó un diseño comparativo, descriptivo y exploratorio, basado en un corpus validado y en la evaluación de tres jueces expertos mediante una rúbrica que considera

adecuación semántica, fluidez, naturalidad lingüística y manejo idiomático-cultural. Los resultados evidencian diferencias relevantes y subrayan la necesidad de posedición y mediación intercultural en la traducción profesional.
Palabras clave: Traducción automática; habla coloquial; inteligencia artificial generativa; modelos de lenguaje; evaluación traductológica.
Abstract: Large language models (LLMs) such as ChatGPT, DeepSeek, Gemini, and Claude have shown significant progress in machine translation; however, translating colloquial speech remains challenging due to idiomatic expressions, slang, pragmatic markers, and cultural references. This study compares their performance in translating forty English colloquial expressions into Spanish, grouped into four categories. A comparative, descriptive, and exploratory design was adopted, using a validated corpus and evaluation by three expert judges through a rubric assessing semantic adequacy, fluency, linguistic naturalness, and idiomatic-cultural handling. Results show Gemini achieved the highest overall performance, followed by ChatGPT, DeepSeek, and Claude. The findings highlight the need for pragmatic sensitivity, cultural mediation, and post-editing in professional translation.
Keywords: Machine translation; colloquial speech; generative artificial intelligence; language models; translation evaluation.
Sumario: 1. Introducción; 2. Marco teórico, 2. 1. Habla coloquial: definición y rasgos, 2. 2. Traducción automática y habla coloquial: dificultades documentadas, 2. 3. Modelos analizados: panorámica técnica, 2. 4. Mediación intercultural y competencia traductora en contextos de IA, 2. 5. Estudios sobre traducción asistida con IA generativa, 2. 6. Ejemplos ilustrativos de traducción coloquial en MEL; 3. Metodología, 3. 1. Diseño del estudio, 3. 2. Corpus de análisis, 3. 3. Procedimiento, 3. 4. Criterios de evaluación, 3. 5. Confiabilidad y análisis de datos; 4. Resultados, 4. 1. Desempeño cuantitativo general, 4. 2. Visualización de tendencias, 4. 3. Resultados por criterio, 4. 4. Análisis cualitativo, 4. 5. Patrones de error, 4. 6. Síntesis de hallazgos; 5. Discusión; 6. Conclusiones; Referencias bibliográficas.
Summary: 1. Introduction; 2. Theoretical framework, 2. 1. Colloquial speech: definition and characteristics, 2. 2. Machine translation and colloquial speech: documented difficulties, 2. 3. Models analyzed: technical overview, 2. 4. Intercultural mediation and translator competence in AI contexts, 2. 5. Studies on translation assisted by generative AI, 2. 6. Illustrative examples of colloquial translation in LLMs; 3. Methodology, 3. 1. Study design, 3. 2. Corpus for analysis, 3. 3. Procedure, 3. 4. Evaluation criteria, 3. 5. Reliability and data analysis; 4. Results, 4. 1. Overall quantitative performance, 4. 2. Visualization of trends, 4. 3. Results by criterion, 4. 4. Qualitative analysis, 4. 5. Error patterns, 4. 6. Summary of findings; 5. Discussion; 6. Conclusions; References.

1. INTRODUCCIÓN

En los últimos años, los modelos de lenguaje generativo (MEL, LLMs, por sus siglas en inglés) como ChatGPT, DeepSeek, Gemini y Claude han revolucionado múltiples tareas lingüísticas, destacando en aplicaciones como la traducción automática. Estos sistemas, basados en arquitecturas avanzadas de aprendizaje profundo, han sido ampliamente evaluados en contextos de traducción de textos formales, mostrando mejoras significativas en coherencia y precisión en comparación con

generaciones previas de traductores automáticos. Sin embargo, la traducción del habla coloquial (caracterizada por la presencia de expresiones idiomáticas, jerga, matices pragmáticos y referencias culturales) representa un desafío particular para estos modelos, debido a la alta variabilidad lingüística y la dificultad inherente de transferir significados figurados entre lenguas.

Estudios recientes han evidenciado que, aunque los modelos de lenguaje generativo han mejorado de manera notable la calidad de la traducción automática, continúan presentando dificultades al enfrentarse a fenómenos propios del habla coloquial, como las expresiones idiomáticas, la jerga, los marcadores pragmáticos y las referencias culturales. Sadiq (2025), por ejemplo, señala que incluso los sistemas más avanzados muestran limitaciones para interpretar y trasladar variantes dialectales y expresiones de fuerte carga cultural, lo que repercute en la naturalidad y adecuación pragmática de las traducciones producidas.

De forma similar, otras investigaciones han observado diferencias en el comportamiento de los distintos modelos al traducir lenguaje no literal, particularmente en aspectos relacionados con la fluidez, la preservación del significado figurado y la adaptación cultural. No obstante, estas diferencias han sido descritas principalmente desde una perspectiva comparativa y todavía existe evidencia limitada para atribuirlas de manera directa a características específicas de la arquitectura o del proceso de entrenamiento de cada modelo.

A pesar del creciente número de investigaciones sobre el empleo de modelos de lenguaje generativo en traducción automática, la mayor parte de estos trabajos se ha centrado en textos especializados, académicos o de registro estándar, privilegiando la evaluación de la precisión terminológica, la adecuación léxica o el rendimiento mediante métricas automáticas.

En contraste, la traducción del habla coloquial –caracterizada por expresiones idiomáticas, jerga, marcadores pragmáticos y referencias culturales– ha recibido una atención considerablemente menor desde una perspectiva traductológica basada en evaluación humana. Asimismo, son escasos los estudios que comparan de manera sistemática el comportamiento de distintos modelos de IA generativa utilizando criterios relacionados con la naturalidad lingüística, la adecuación pragmática y la mediación intercultural. Este vacío limita la comprensión de las fortalezas y debilidades reales de estos sistemas en uno de los ámbitos más complejos

de la traducción profesional. En este contexto, el presente estudio busca contribuir a cubrir esa laguna mediante un análisis comparativo de cuatro modelos de lenguaje generativo basado en evaluación experta.

Con el propósito de atender este vacío de investigación, el presente estudio tiene como objetivo general evaluar y comparar el rendimiento de ChatGPT, DeepSeek, Gemini y Claude en la traducción del habla coloquial del inglés al español, considerando criterios como la adecuación semántica, la fluidez, la naturalidad lingüística y el manejo de expresiones idiomáticas y referencias culturales. La pregunta de investigación que orienta este trabajo es la siguiente: ¿cómo varía la calidad de la traducción del habla coloquial del inglés al español entre ChatGPT, DeepSeek, Gemini y Claude en términos de adecuación semántica, fluidez, naturalidad lingüística y manejo de expresiones idiomáticas y referencias culturales? A partir de esta pregunta se plantean las siguientes hipótesis de trabajo: (aa) existen diferencias en el desempeño de los modelos de lenguaje generativo en la traducción del habla coloquial; (bb) algunos modelos obtendrán mejores resultados que otros en criterios relacionados con la naturalidad lingüística y el manejo de expresiones idiomáticas y referencias culturales; y (cc) los modelos presentarán patrones diferenciados de error en aspectos como la literalidad, la rigidez estilística y la adaptación cultural.

El presente estudio busca aportar evidencia empírica sobre el comportamiento de cuatro modelos de lenguaje generativo en la traducción del habla coloquial del inglés al español mediante una evaluación basada en criterios lingüísticos, pragmáticos e interculturales. Con ello, se pretende contribuir a la investigación sobre traducción asistida por inteligencia artificial (IA), proporcionando evidencia que permita comprender mejor las fortalezas, las limitaciones y las necesidades de posedición de estos sistemas en contextos de elevada complejidad pragmática y cultural. Asimismo, los hallazgos pueden servir como referente para futuras investigaciones orientadas a la evaluación de modelos de lenguaje generativo desde una perspectiva traductológica.

2. MARCO TEÓRICO

El presente estudio se adscribe a una perspectiva funcional, pragmática e intercultural del lenguaje y de la traducción, en la que el lenguaje se concibe como una práctica social situada, dinámicamente construida en contextos comunicativos específicos, y no únicamente como

un sistema abstracto de correspondencias léxico-gramaticales. Desde esta concepción, la traducción se entiende como una actividad interpretativa y mediadora que trasciende la equivalencia formal para atender a intenciones comunicativas, valores socioculturales y funciones discursivas, especialmente relevantes en registros no normativos como el habla coloquial.

Este enfoque se alinea con los planteamientos contemporáneos de la traductología funcional, que subrayan el papel del propósito comunicativo, el contexto de recepción y el destinatario en la toma de decisiones traductoras (Pym, 2021; House, 2021).

En este sentido, la traducción no se limita a reproducir contenidos lingüísticos, sino que implica una negociación de significados culturalmente situados, donde la adecuación pragmática y la aceptabilidad comunicativa adquieren un papel central.

2.1. Habla coloquial: definición y rasgos

El habla coloquial constituye un registro lingüístico de uso cotidiano, caracterizado por su espontaneidad, inmediatez y fuerte anclaje cultural. A diferencia de los registros formales o literarios, el coloquial se emplea en contextos de interacción social no planificada y, por lo general, en situaciones de proximidad comunicativa.

Así como Beeching (2016) señala que este tipo de habla se distingue por la utilización recurrente de expresiones idiomáticas, jerga, marcadores pragmáticos y estructuras sintácticas fragmentadas. Los marcadores pragmáticos (tales como you know, like o sort of en inglés) cumplen una función discursiva más que proposicional, ya que guían la interpretación del mensaje, indican cortesía y contribuyen a la cohesión conversacional.

De manera que, el habla coloquial se encuentra atravesada por elementos de variación dialectal y sociocultural, lo que la convierte en un desafío para la traducción. Una expresión coloquial en un dialecto específico puede no tener un equivalente directo en la lengua meta, lo que obliga al traductor a recurrir a estrategias de adaptación, paráfrasis o sustitución cultural (Andrade, 2024 y Beeching, 2016).

En este sentido, la traducción de expresiones coloquiales no se limita a un traslado lingüístico, sino que requiere un proceso de mediación intercultural.

2.2. Traducción automática y habla coloquial: dificultades documentadas

Los avances en la traducción automática (TA), especialmente desde la irrupción de los sistemas neuronales, han transformado radicalmente la calidad y naturalidad de las traducciones. Los sistemas actuales superan con creces a sus predecesores estadísticos en términos de fluidez gramatical y coherencia semántica. Sin embargo, los registros informales continúan siendo un obstáculo significativo. Liu *et al*. (2023) advierten que las expresiones idiomáticas y los proverbios plantean dificultades persistentes, ya que los modelos suelen producir traducciones literales que resultan poco naturales o culturalmente inadecuadas. Este fenómeno responde en gran parte a la falta de representatividad de tales expresiones en los corpus de entrenamiento y a la tendencia de los sistemas a buscar equivalencias palabra por palabra en lugar de interpretaciones figurativas.

Por otro lado, Michel *et al*. (2018) destacan la presencia de texto ruidoso en interacciones coloquiales (fragmentos incompletos, abreviaciones, emoticonos, errores tipográficos o alternancia de códigos), lo que genera degradación en el rendimiento de los sistemas. Este tipo de ruido, común en redes sociales y chats, introduce retos adicionales porque obliga a los modelos a operar en condiciones muy distintas a las del texto escrito estándar con el que fueron entrenados. En este mismo sentido, Naveen (2024) subraya que los MEL enfrentan problemas de desambiguación semántica, manejo de polisemia y preservación de matices pragmáticos, factores que repercuten directamente en la coherencia y naturalidad de la traducción. Estos desafíos se agravan cuando las expresiones incluyen referencias culturales implícitas o marcadores discursivos que carecen de equivalentes directos en la lengua meta.

Desde la óptica del campo de la subtitulación, un entorno donde prevalece el habla coloquial, estudios como el de Pramodya *et al*. (2025) demostraron que la inclusión de metainformación contextual, como sinopsis, género o datos de los personajes, mejora de manera significativa la calidad de la traducción. Este hallazgo revela que la traducción de registros coloquiales no puede depender únicamente de la potencia estadística de los modelos, sino que requiere de un procesamiento contextual que capture los matices culturales e intertextuales. En otras

palabras, el éxito en la traducción de lo coloquial depende tanto del modelo como de la riqueza del contexto proporcionado.

Adicionalmente, investigaciones recientes han señalado que las métricas automáticas tradicionales (BLEU, METEOR, COMET, entre otras) presentan limitaciones para evaluar la calidad de las traducciones coloquiales (Naveen, 2024; Zhang *et al*., 2024). Estas métricas tienden a premiar la coincidencia léxica en lugar de la adecuación cultural, invisibilizando el valor de soluciones creativas o adaptadas al registro. Esto ha impulsado el desarrollo de métodos de evaluación más cercanos al juicio humano, así como propuestas de métricas híbridas que consideren la aceptabilidad pragmática y el grado de naturalidad.

En conjunto, la literatura apunta a que los desafíos centrales de la traducción automática en lo coloquial pueden resumirse en cuatro dimensiones principales: (a) la dificultad para trasladar significados figurados y expresiones idiomáticas, (b) la vulnerabilidad frente a textos fragmentarios o no normativos, (c) la insuficiencia de las métricas automáticas para evaluar la adecuación cultural y pragmática, y (d) la necesidad de integrar información contextual adicional para mejorar la comprensión de referencias culturales. Estos retos confirman que, a pesar de los avances tecnológicos, el habla coloquial continúa siendo un terreno particularmente complejo para los sistemas de traducción automática y un campo fértil para la investigación traductológica.

2.3. Modelos analizados: panorámica técnica

Para comprender el comportamiento de los sistemas de IA en la traducción del habla coloquial, resulta necesario revisar brevemente las características técnicas y los enfoques de diseño de los modelos seleccionados. Aunque todos se inscriben en la familia de grandes modelos de lenguaje (MEL), cada uno integra arquitecturas, metodologías de entrenamiento y principios de alineación distintos, los cuales repercuten directamente en la calidad y naturalidad de sus traducciones.

En primer lugar, ChatGPT, desarrollado por OpenAI, se apoya en la arquitectura Transformer multimodal de la familia GPT-4. Una de sus innovaciones más destacadas es el empleo del aprendizaje por refuerzo con retroalimentación humana (RLHF), estrategia que permite ajustar las respuestas del modelo a las preferencias humanas y mejorar su utilidad comunicativa (Achiam *et al*., 2023; Ouyang *et al*., 2022). Gracias a este mecanismo, ChatGPT ha mostrado un rendimiento notable en la

generación de traducciones más fluidas y culturalmente adecuadas, especialmente en interacciones dinámicas donde el usuario puede retroalimentar el proceso. Sin embargo, investigaciones recientes han identificado que, frente a expresiones coloquiales con fuerte carga cultural, aún persisten problemas de literalidad e inconsistencias en la adecuación pragmática (Zhang *et al*., 2024).

En contraste, DeepSeek-V2 se distingue por la implementación del paradigma MoE, en el cual solo un subconjunto de parámetros se activa en cada paso de inferencia. Este diseño híbrido busca equilibrar eficiencia computacional y especialización semántica, lo que permite alcanzar un alto nivel de precisión en determinados dominios (Liu *et al*., 2024). No obstante, esta especialización puede limitar la flexibilidad estilística requerida por los registros coloquiales, lo que a veces se traduce en una menor fluidez o en traducciones que no capturan del todo la naturalidad conversacional.

Por su parte, Gemini 1.5, creado por Google DeepMind, representa un avance significativo en el procesamiento multimodal y en la capacidad de manejar ventanas de contexto extremadamente largas. Estos atributos son especialmente relevantes para la traducción coloquial, dado que permiten mantener coherencia en interacciones prolongadas y procesar referencias culturales dependientes del contexto narrativo (Gemini Team, 2024). Al ser capaz de integrar texto, imagen, audio y vídeo en su análisis, Gemini ofrece un marco robusto para abordar la complejidad pragmática y cultural que caracteriza al lenguaje informal.

Finalmente, Claude 3, desarrollado por Anthropic, se orienta hacia la seguridad y la fiabilidad mediante el enfoque de Constitutional AI (Bai *et al*., 2022). Esta metodología busca reducir sesgos y riesgos en la generación de respuestas, priorizando la transparencia y el cumplimiento de principios éticos. Si bien ello se traduce en salidas más consistentes y coherentes, también limita la creatividad y la riqueza idiomática, dando como resultado traducciones más neutras y conservadoras (Anthropic, 2024). Esta característica, si bien valiosa para garantizar fiabilidad, puede ser restrictiva en la recreación de matices propios del habla coloquial.

En conjunto, las diferencias arquitectónicas y de entrenamiento descritas en la literatura permiten comprender que los modelos de lenguaje generativo han sido desarrollados con enfoques técnicos distintos, los cuales podrían influir en su comportamiento en diversas tareas lingüísticas. No obstante, la evidencia disponible todavía no permite establecer de manera concluyente cómo dichas características se traducen en diferencias

específicas de desempeño en la traducción del habla coloquial. En consecuencia, el presente estudio no pretende explicar causalmente el rendimiento de los modelos a partir de su arquitectura, sino describir y comparar empíricamente su comportamiento traductológico mediante criterios de evaluación lingüística, pragmática e intercultural.

2.4. Mediación intercultural y competencia traductora en contextos de IA

Desde la perspectiva adoptada en este estudio, la lengua(je) se concibe no únicamente como un sistema formal de signos, sino como una práctica social situada, cuya interpretación depende de factores pragmáticos, culturales e interaccionales. Esta concepción, alineada con enfoques funcionales y pragmáticos del lenguaje, resulta fundamental para comprender la traducción como un proceso orientado a la función comunicativa del texto y no solo a la correspondencia formal entre lenguas (Bowker y Buitrago Ciro, 2021; Kenny y Doherty, 2023). En este marco, traducir implica interpretar intenciones, valores socioculturales y efectos discursivos que se actualizan en contextos específicos de uso, lo que adquiere especial relevancia en registros coloquiales y no normativos.

Desde el funcionalismo traductológico, la traducción se entiende como una actividad orientada al propósito comunicativo y a la adecuación funcional del texto meta, lo que sitúa al traductor como un agente decisor que evalúa y adapta los recursos lingüísticos en función del contexto y del receptor. Esta perspectiva resulta clave para analizar el uso de modelos de lenguaje generativo en traducción, ya que, aunque estos sistemas pueden producir soluciones semánticamente plausibles, su funcionamiento se basa en patrones estadísticos que no siempre garantizan una interpretación funcional ni culturalmente adecuada del discurso, especialmente en ausencia de intervención humana (Hansen-Schirra *et al*., 2021; O'Brien, 2022).

Asimismo, el enfoque de la competencia traductora desarrollado por el grupo PACTE y operacionalizado en el marco EMT (European Master's in Translation) permite situar la mediación intercultural como una competencia central del traductor profesional. En particular, subcompetencias como la pragmática, la estratégica y la instrumental resultan determinantes en contextos de traducción asistida por IA, donde el traductor debe evaluar críticamente las salidas generadas por los sistemas, identificar pérdidas pragmáticas o culturales y aplicar estrategias

de posedición orientadas a la naturalidad y la adecuación funcional del texto meta (PACTE Group, 2020; European Commission, 2022).

Desde esta perspectiva integrada, los modelos de lenguaje generativo no pueden concebirse como sustitutos del traductor humano, sino como artefactos mediadores cuya eficacia depende del nivel de competencia traductora del usuario. La creatividad, la sensibilidad cultural y la capacidad de toma de decisiones estratégicas del traductor constituyen elementos esenciales para garantizar una mediación intercultural efectiva. En consecuencia, este posicionamiento teórico refuerza la necesidad de una formación traductora que, lejos de rechazar la automatización, promueva una integración crítica y reflexiva de la IA, sustentada en una concepción del lenguaje como práctica social y de la traducción como mediación intercultural (Kenny y Doherty, 2023; O'Brien, 2022).

2.5. Estudios sobre traducción asistida con IA generativa

La investigación empírica más reciente ofrece un panorama mixto sobre el desempeño de los LLMs en traducción, particularmente en contextos especializados y registros informales. Zhang *et al*. (2024) hallaron que, en textos legales, la superioridad de los modelos no es uniforme: si bien ChatGPT y Gemini mostraron mayor fluidez y naturalidad, la traducción automática neuronal tradicional resultó más consistente en términos de precisión terminológica. Este hallazgo confirma que la excelencia de los MEL no siempre se traduce en superioridad absoluta, sino que depende del tipo de texto, del dominio temático y de las necesidades de la traducción (fidelidad vs. naturalidad).

En otro trabajo, Naveen (2024) enfatizó que los proverbios y expresiones idiomáticas continúan siendo un reto para los sistemas, aunque los MEL superan en muchos casos a los traductores automáticos neuronales convencionales. El autor resalta que, aunque los MEL logran captar mejor el sentido figurado en contextos frecuentes, todavía presentan inconsistencias cuando se enfrentan a referencias culturales menos difundidas o a variaciones dialectales. Esto demuestra que la cobertura de los datos de entrenamiento sigue siendo un factor crítico en la calidad de las traducciones.

Por su parte, Pramodya *et al*. (2025) demostraron que en subtitulación de películas los MEL alcanzan un rendimiento superior al de sistemas como Google Translate o DeepL, siempre que se incorporen elementos de contexto adicional como sinopsis, género o perfil de personajes. Dicho

hallazgo evidencia la relevancia del contexto pragmático como variable decisiva en la traducción coloquial: los modelos no solo deben reconocer las palabras, sino también interpretar la intención comunicativa y las referencias implícitas que articulan el discurso.

De manera complementaria, estudios recientes en traducción audiovisual y en entornos de chat conversacional sugieren que la introducción de mecanismos de memoria a largo plazo y ventanas contextuales extendidas mejora sustancialmente la coherencia en diálogos coloquiales (Gemini Team, 2024). Asimismo, experimentos de evaluación con métricas híbridas (que combinan métricas automáticas y juicios humanos) señalan que los MEL son capaces de ofrecer soluciones culturalmente más aceptables, aunque las métricas tradicionales como BLEU o COMET no siempre reflejen estos avances (Naveen, 2024; Zhang *et al*., 2024).

En síntesis, la literatura reciente converge en señalar que el rendimiento de los MEL en traducción coloquial se encuentra mediado por cuatro factores clave: (1) la capacidad de modelar contextos amplios y mantener coherencia a lo largo de interacciones extensas, (2) la exposición suficiente a datos coloquiales durante el entrenamiento, (3) los mecanismos de alineación (como RLHF o Constitutional AI) que guían la toma de decisiones pragmáticas hacia soluciones culturalmente adecuadas, y (4) el desarrollo de métodos de evaluación que valoren no solo la equivalencia semántica, sino también la naturalidad, fluidez y adecuación cultural. Estos factores confirman que, aunque los MEL representan un avance considerable frente a generaciones anteriores de traductores automáticos, la traducción del habla coloquial sigue siendo un campo de experimentación en constante evolución.

2.6. Ejemplos ilustrativos de traducción coloquial en MEL

Aunado a todo lo anterior, la traducción del habla coloquial constituye un reto central para los sistemas de IA generativa debido a la presencia de expresiones idiomáticas, jerga, marcadores pragmáticos y referencias culturales implícitas.

Estos elementos no solo dificultan la transferencia semántica, sino que además demandan estrategias de adaptación cultural que trascienden la literalidad. Para visualizar las diferencias de desempeño entre modelos, a continuación, se presentan ejemplos de expresiones coloquiales traducidas por ChatGPT, DeepSeek, Gemini y Claude.

Expresión en inglés	Sentido literal	Sentido coloquial	*ChatGPT*	*DeepSeek*	*Gemini*	*Claude*
Kick the bucket	Patear el balde	Morir	Morir (adecuado)	Oscila entre morir y literalidad	Fallecer / morirse con fluidez	Perder la vida (neutro)
That party was lit	Esa fiesta estuvo encendida	La fiesta estuvo increíble/divertida	Increíble / genial	Muy buena / literal a veces	Padrísima / genial (juvenil)	Muy divertida (neutral)
You know, it's kind of weird	Tú sabes, es un poco raro	Marcador de atenuación	Pues / la verdad	Sabe (literal)	Pues es medio raro (natural)	Sabe / se omite
It's my Achilles heel	Es mi talón de Aquiles	Punto débil	Talón de Aquiles (fluido)	Talón de Aquiles (neutro)	Uso natural e integrado	Talón de Aquiles (formal)

Tabla 1. Ejemplificación de traducción coloquial en *MEL*.

Los ejemplos de la tabla 1 ilustran que los modelos evaluados presentan comportamientos diferenciados frente a los retos del habla coloquial. En los casos analizados, ChatGPT y Gemini produjeron soluciones más naturales y culturalmente adecuadas, mientras que DeepSeek tendió a conservar con mayor frecuencia el sentido denotativo, aunque con menor fluidez idiomática. Claude, por su parte, ofreció traducciones generalmente coherentes, pero más conservadoras y culturalmente neutras. No obstante, estos resultados deben interpretarse como tendencias observadas dentro del corpus analizado y no como evidencia suficiente para atribuir tales diferencias a características específicas de la arquitectura o del entrenamiento de cada sistema.

Estos hallazgos se alinean con estudios recientes que han documentado limitaciones recurrentes en la traducción de expresiones idiomáticas y coloquialismos. Naveen (2024) y Liu *et al*. (2023) destacan la persistencia de problemas de literalidad y polisemia, mientras que Pramodya *et al*. (2025) subrayan la importancia del contexto audiovisual en la mejora de traducciones coloquiales. De igual manera, Zhang *et al*. (2024) evidencian que el desempeño de los modelos depende del equilibrio entre fidelidad terminológica y naturalidad expresiva, lo que refuerza la necesidad de enfoques evaluativos más sofisticados.

En conjunto, los resultados sugieren que la traducción del habla coloquial no depende exclusivamente de la potencia computacional, sino de la integración de factores pragmáticos y culturales. Así, los avances futuros deberán enfocarse en (a) entrenar a los modelos con corpus

coloquiales más representativos, (b) diseñar métricas híbridas que consideren tanto la equivalencia semántica como la adecuación cultural, y (c) potenciar enfoques multimodales capaces de capturar la complejidad del lenguaje en interacción. Solo de esta forma será posible desarrollar sistemas traductores que se aproximen con mayor fidelidad a la sensibilidad humana en el tratamiento del discurso coloquial.

3. METODOLOGÍA

La presente investigación se enmarca en un enfoque comparativo, descriptivo y exploratorio, cuyo objetivo fue analizar el rendimiento de cuatro sistemas de IA generativa (ChatGPT, DeepSeek, Gemini y Claude) en la traducción del habla coloquial del inglés al español. Este enfoque se adopta en consonancia con la tradición de la traductología descriptiva (Toury, 2012) y con estudios recientes sobre traducción asistida por IA, que subrayan la necesidad de documentar empíricamente el comportamiento de los sistemas antes de avanzar hacia modelos explicativos o predictivos (Bowker y Buitrago Ciro, 2021; Kenny y Doherty, 2023).

Desde una perspectiva metodológica, el análisis se centra en expresiones coloquiales aisladas, entendidas como unidades microdiscursivas altamente cargadas de valor pragmático y cultural. Esta decisión no responde a una limitación del diseño, sino a una estrategia analítica consciente, orientada a observar con mayor precisión cómo los modelos activan (o fallan en activar) inferencias pragmáticas e interculturales en ausencia de un contexto discursivo amplio. Diversos estudios en pragmática y traducción han señalado que fenómenos como la idiomaticidad, la atenuación, la ironía o la cercanía interpersonal suelen manifestarse de forma particularmente intensa en este tipo de unidades breves, lo que las convierte en un terreno privilegiado para evaluar la mediación intercultural y las necesidades de posedición experta (Hansen-Schirra *et al*., 2021; O'Brien, 2022).

En este sentido, la investigación no busca establecer generalizaciones universales sobre el desempeño global de los modelos, sino identificar patrones de comportamiento traductológico, tales como la literalidad, la neutralización pragmática o la pérdida de marcadores socioculturales, que suelen emerger con claridad en el análisis microdiscursivo. De este modo, el estudio pretende construir un marco de referencia empírico que

contribuya tanto a la investigación en traductología como al análisis metodológico de la traducción asistida por IA .

3.1. Diseño del estudio

El diseño fue comparativo porque permitió contrastar los resultados de cuatro modelos bajo condiciones homogéneas; descriptivo porque caracterizó el comportamiento traductológico de cada sistema; y exploratorio porque abordó un fenómeno todavía poco analizado en la literatura científica: la traducción automática del habla coloquial.

3.2. Corpus de análisis

El corpus se construyó mediante un proceso de selección intencional y estratificada de expresiones coloquiales en inglés, con sus respectivas equivalencias en español. Se establecieron cuatro categorías para abarcar la diversidad de fenómenos lingüísticos asociados al registro coloquial:

- **Expresiones idiomáticas**: frases cuyo sentido no puede inferirse literalmente.
- ***Slang* juvenil y urbano**: expresiones emergentes en contextos juveniles y digitales.
- **Marcadores pragmáticos**: elementos conversacionales que regulan la interacción.
- **Referencias culturales**: expresiones con fuerte anclaje histórico o social.

El corpus final estuvo integrado por cuarenta expresiones (diez por categoría), seleccionadas mediante un muestreo intencional con fines analíticos. El tamaño del corpus responde al carácter exploratorio del estudio, cuyo propósito no fue establecer generalizaciones sobre el desempeño de los modelos de lenguaje, sino identificar patrones traductológicos recurrentes en la traducción de distintos fenómenos del habla coloquial. La distribución equilibrada entre las cuatro categorías permitió analizar expresiones idiomáticas, jerga, marcadores pragmáticos y referencias culturales bajo condiciones homogéneas de evaluación. Las expresiones que integran el corpus se presentan en las tablas 2, 3, 4 y 5.

La evaluación de las traducciones se realizó tomando como referencia el español mexicano contemporáneo. Esta decisión metodológica obedece a que los tres evaluadores pertenecen a dicha comunidad lingüística y a

que el objetivo del estudio fue valorar la naturalidad pragmática dentro de una variedad específica del español. En consecuencia, las valoraciones de expresiones coloquiales se interpretan exclusivamente desde este marco dialectal y no pretenden generalizarse al conjunto de las variedades del español.

Expresión en inglés	Traducción al español	Fuente principal
Kick the bucket	Estirar la pata / Morir	Cambridge Idioms Dictionary; BNC
Spill the beans	Revelar el secreto	Linguee; Sketch Engine
Break the ice	Romper el hielo	BNC; COCA
Hit the sack	Irse a dormir	Cambridge Idioms Dictionary
Bite the bullet	Afrontar la situación	OED; COCA
Let the cat out of the bag	Descubrir el pastel	Cambridge Idioms Dictionary
Under the weather	Sentirse mal / indispuesto	BNC; COCA
Burn the midnight oil	Trasnochar / desvelarse	Europarl Corpus
The ball is in your court	La pelota está en tu cancha	BNC; Linguee
Costs an arm and a leg	Cuesta un ojo de la cara	Sketch Engine; OED

Tabla 2. Expresiones idiomáticas (inglés–español).

T

Expresión en inglés	Traducción al español	Fuente principal
Kick the bucket	Estirar la pata / Morir	Cambridge Idioms Dictionary; BNC
Spill the beans	Revelar el secreto	Linguee; Sketch Engine
Break the ice	Romper el hielo	BNC; COCA
Hit the sack	Irse a dormir	Cambridge Idioms Dictionary
Bite the bullet	Afrontar la situación	OED; COCA
Let the cat out of the bag	Descubrir el pastel	Cambridge Idioms Dictionary
Under the weather	Sentirse mal / indispuesto	BNC; COCA

Tabla 3. *Jerga* juvenil y urbana (inglés-español).

Expresión en inglés	Traducción al español	Fuente principal
You know	¿Sabes? / Pues / La verdad	SBCSAE; COCA
Like	O sea / Como que	COCA; BNC
Sort of	Más o menos / Algo así	COCA
Well	Pues / Bueno	COCA; BNC
I mean	Es decir	SBCSAE
Right?	¿Verdad?	COCA; BNC
Actually	La verdad / En realidad	COCA
Anyway	En fin / De todos modos	COCA
Basically	Básicamente / O sea	COCA
Seriously	En serio / Neta	SBCSAE; COCA

Tabla 4. Marcadores pragmáticos (inglés–español).

Expresión en inglés	Traducción al español	Fuente principal
Achilles heel	Talón de Aquiles	OED; Europarl
The American dream	El sueño americano	COCA; Europarl
Catch-22	Situación sin salida / círculo vicioso	OED; BNC
Waterloo momento	Momento de derrota decisiva	COCA; BNC
Rome wasn't built in a day	Roma no se construyó en un día	OED
Big Brother	Gran Hermano / vigilancia autoritaria	COCA; Europarl
Trojan horse	Caballo de Troya	OED; Europarl
David and Goliath	David contra Goliat	COCA; BNC
To meet one's Waterloo	Encontrar su Waterloo	OED
Crossing the Rubicon	Cruzando el Rubicón	BNC; OED

Tabla 5. Referencias culturales (inglés–español).

3.3. procedimiento

El procedimiento del estudio se organizó en cuatro fases consecutivas y controladas:

1. Selección y preparación del corpus: se recopilaron cuarenta expresiones coloquiales en inglés, clasificadas en cuatro categorías (idiomáticas, jerga, marcadores pragmáticos y referencias culturales). Cada expresión se verificó en corpus reconocidos (COCA, BNC, Urban Dictionary, OpenSubtitles, entre otros) para asegurar su frecuencia de uso y vigencia.
2. Traducción con los modelos de IA: cada expresión fue introducida individualmente en ChatGPT, DeepSeek, Gemini y Claude, utilizando el mismo *prompt* en inglés: «Translate into Spanish: [expresión]». Se evitó proporcionar contexto adicional o instrucciones de estilo, con el fin de evaluar el desempeño base de cada modelo. Todas las consultas se realizaron utilizando exactamente el mismo *prompt* y sin proporcionar información contextual adicional, con el propósito de mantener condiciones homogéneas para todos los modelos evaluados. Asimismo, las consultas se efectuaron dentro de un mismo periodo de recopilación de datos para reducir el posible efecto de actualizaciones sucesivas de los sistemas durante el desarrollo del estudio.
3. Recolección y sistematización: las traducciones generadas fueron recopiladas en una base de datos comparativa. Cada entrada incluyó: expresión original, traducción obtenida por cada modelo, categoría, fecha de prueba y versión del sistema. Esto permitió mantener un registro claro y reproducible del proceso. El registro de la fecha de consulta y de la versión disponible de cada modelo tuvo como finalidad favorecer la trazabilidad del proceso y facilitar, en la medida de lo posible, la replicabilidad del estudio, considerando la naturaleza dinámica de los modelos de lenguaje generativo y sus frecuentes actualizaciones.
4. Evaluación por jueces expertos: tres evaluadores con perfiles distintos analizaron los resultados: (a) un profesor-investigador en traductología, (b) un traductor profesional especializado en traducción audiovisual y (c) un estudiante avanzado de licenciatura en traducción. Se buscó de esta manera combinar la perspectiva

académica, la práctica profesional y la visión formativa. Los jueces recibieron la rúbrica previamente diseñada y fueron capacitados en su aplicación para asegurar consistencia en la evaluación.

3.4. Criterios de evaluación

La calidad de las traducciones fue evaluada según cuatro criterios fundamentales, seleccionados por su relevancia en la traducción del habla coloquial:

1. Adecuación semántica: se midió la fidelidad del contenido traducido respecto al original, valorando si el sentido se preservaba de manera precisa.
2. Fluidez: se evaluó la corrección gramatical, cohesión y estilo del texto en español. Una traducción fluida debía sonar natural y no evidenciar interferencias del inglés.
3. Naturalidad lingüística: se examinó el grado de cercanía de la traducción al español coloquial auténtico, considerando si la frase podría usarse en un contexto real de interacción.
4. Manejo de expresiones idiomáticas y referencias culturales: se analizó la capacidad de los modelos para reconocer significados figurativos y adaptarlos a equivalencias culturalmente pertinentes, en lugar de recurrir a traducciones literales.

Cada criterio fue valorado con una escala Likert de 1 a 5 puntos, apoyada en descriptores cualitativos claros (ver tabla 6).

Criterio	1 Deficiente	2 Insuficiente	3 Aceptable	4 Bueno	5 Excelente
Adecuación semántica	El significado se pierde o es incorrecto.	Se conserva parcialmente, con omisiones relevantes.	Se conserva el sentido general, aunque con errores menores.	Se conserva el sentido casi completo, con buena correspondencia.	Se transmite íntegramente el significado del original.
Fluidez	Texto ininteligible o con errores graves de gramática.	Texto comprensible, pero con errores frecuentes que interrumpen la lectura.	Texto entendible con algunos errores gramaticales o de cohesión.	Texto mayormente correcto, con fluidez casi natural.	Texto completamente natural, sin errores gramaticales.

Naturalidad lingüística	No se asemeja al español coloquial; suena artificial.	Suena forzado, con estructuras poco naturales.	Se acerca al español coloquial, aunque con cierta rigidez.	Es natural y fluido en la mayoría de los contextos.	Suena completamente natural y auténtico en español coloquial.
Manejo de expresiones idiomáticas y referencias culturales	Traducción literal y sin sentido.	Traducción literal con escasa adecuación cultural.	Traducción que comunica el sentido, pero poco natural.	Traducción adecuada culturalmente, aunque con leves ajustes necesarios.	Traducción culturalmente pertinente, idiomática y contextualizada.

Tabla 6. Criterios de valorización para la evaluación.

3.5. Confiabilidad y análisis de datos

Con el objetivo de garantizar la confiabilidad de los resultados, se adoptaron varias estrategias metodológicas:

- Capacitación de jueces: los evaluadores recibieron una sesión introductoria sobre el uso de la rúbrica y realizaron pruebas piloto para homogeneizar criterios de valoración.
- Consistencia interevaluadora: con el propósito de favorecer la homogeneidad de las valoraciones, los tres evaluadores participaron en una sesión de calibración previa mediante la evaluación conjunta de un conjunto piloto de expresiones. Durante esta etapa se discutieron los criterios de la rúbrica y se resolvieron posibles discrepancias interpretativas antes de iniciar la evaluación definitiva. Asimismo, cuando durante la evaluación se presentaron diferencias relevantes entre las puntuaciones asignadas, estas fueron revisadas conjuntamente hasta alcanzar un consenso razonado.
- Revisión cruzada: cuando se identificaron discrepancias relevantes entre las valoraciones de los jueces, estas fueron discutidas de manera conjunta hasta alcanzar un consenso razonado, documentando las razones que sustentaron la decisión final.
- Validez de contenido: la rúbrica fue validada mediante revisión de literatura previa (Castilho *et al*., 2018; Moorkens, 2020) y adaptada a las particularidades de la traducción coloquial.

En cuanto al análisis, se aplicó una estrategia mixta:

- Cuantitativo: cálculo de medias y elaboración de rankings descriptivos del desempeño de cada modelo en función de los criterios de evaluación establecidos.
- Cualitativo: categorización de errores (literalidad, calcos, pérdida de matices pragmáticos, rigidez estilística), con ejemplos representativos para ilustrar tendencias.

Dado el carácter exploratorio del estudio y el diseño metodológico adoptado, la confiabilidad del proceso de evaluación se sustentó en la aplicación de una rúbrica analítica, la calibración previa de los evaluadores y la resolución consensuada de discrepancias durante el proceso de valoración. Estas estrategias permitieron favorecer la consistencia de las evaluaciones y fortalecer la validez de los resultados obtenidos. Asimismo, el análisis estadístico tuvo un propósito eminentemente descriptivo, por lo que las medidas de tendencia central y dispersión se emplearon para caracterizar el comportamiento comparativo de los modelos, mientras que el análisis cualitativo permitió identificar patrones recurrentes de traducción y tipos de error asociados a cada sistema. En consecuencia, los resultados deben interpretarse como evidencia descriptiva circunscrita al corpus analizado y no como inferencias generalizables sobre el desempeño global de los modelos de lenguaje generativo.

4. RESULTADOS

El análisis del corpus de cuarenta expresiones coloquiales permitió evaluar el desempeño de los modelos ChatGPT, DeepSeek, Gemini y Claude en los cuatro criterios definidos: adecuación semántica, fluidez, naturalidad lingüística y manejo de expresiones idiomáticas/culturales. Los resultados revelan tanto diferencias cuantitativas en las puntuaciones promedio como patrones cualitativos en los tipos de errores cometidos por cada sistema.

Para la etapa de evaluación, se diseñó una rúbrica analítica que permitió operacionalizar cada uno de los criterios considerados. La adecuación semántica se evaluó en función del grado de conservación del significado global de la expresión fuente y de la ausencia de distorsiones conceptuales relevantes. La fluidez se valoró atendiendo a la corrección gramatical y sintáctica del texto meta, así como a su coherencia interna.

Por su parte, la naturalidad lingüística se analizó considerando el grado en que la traducción se ajusta a los usos convencionales del español en contextos coloquiales, evitando formulaciones artificiales o marcadamente literales. Finalmente, el manejo de expresiones idiomáticas y culturales se evaluó a partir de la capacidad del modelo para reconocer y resolver adecuadamente unidades fraseológicas, referencias culturales implícitas y marcadores pragmáticos propios del habla coloquial.

La aplicación sistemática de esta rúbrica permitió identificar no solo variaciones en el rendimiento promedio de los modelos, sino también patrones recurrentes de dificultad asociados a cada criterio, tales como la tendencia a preservar la corrección semántica en detrimento de la naturalidad pragmática o la neutralización de rasgos culturales específicos. De este modo, el análisis aporta una visión más detallada de los tipos de intervención de posedición que cada sistema suele requerir, fortaleciendo la interpretación de los resultados y su relevancia para la investigación en traducción asistida por IA.

4.1. Desempeño cuantitativo general

La tabla 7 resume las puntuaciones promedio (escala Likert de 1 a 5) para cada modelo.

Modelo	Adecuación semántica	Fluidez	Naturalidad lingüística	Manejo idiomático/ cultural	Promedio general
ChatGPT	4.5	4.3	4.2	4.0	4.25
DeepSeek	4.6	3.8	3.5	3.4	3.83
Gemini	4.4	4.4	4.5	4.3	4.40
Claude	4.2	4.1	3.7	3.6	3.90

Tabla 7. Promedio de puntuaciones por modelo y criterio.

En términos globales, Gemini obtuvo el mejor desempeño promedio (4.40), seguido de ChatGPT (4.25), *Claude* (3.90) y finalmente DeepSeek (3.83). Las diferencias observadas entre los modelos deben interpretarse como tendencias descriptivas derivadas del corpus analizado. Aunque las puntuaciones promedio permiten identificar comportamientos diferenciados entre los sistemas, el objetivo del estudio no fue establecer diferencias estadísticamente significativas, sino describir

comparativamente su desempeño en distintos criterios de evaluación relacionados con la traducción del habla coloquial.

4.2. Visualización de tendencias

El gráfico de barras (figura 1) permite observar diferencias claras entre los modelos. Gemini se distingue por alcanzar la puntuación más alta en naturalidad (4.5) y en manejo idiomático-cultural (4.3), lo que lo posiciona como el sistema más competente en la traducción de expresiones coloquiales. ChatGPT presenta un rendimiento equilibrado en todos los criterios, aunque con puntuaciones ligeramente menores en el manejo de expresiones idiomáticas. DeepSeek lidera en adecuación semántica (4.6), confirmando su precisión conceptual, pero muestra la puntuación más baja en naturalidad (3.5), lo que refleja rigidez estilística. Finalmente, Claude ofrece resultados intermedios y consistentes, con una clara tendencia hacia la neutralidad cultural, lo que asegura coherencia, aunque a costa de expresividad.

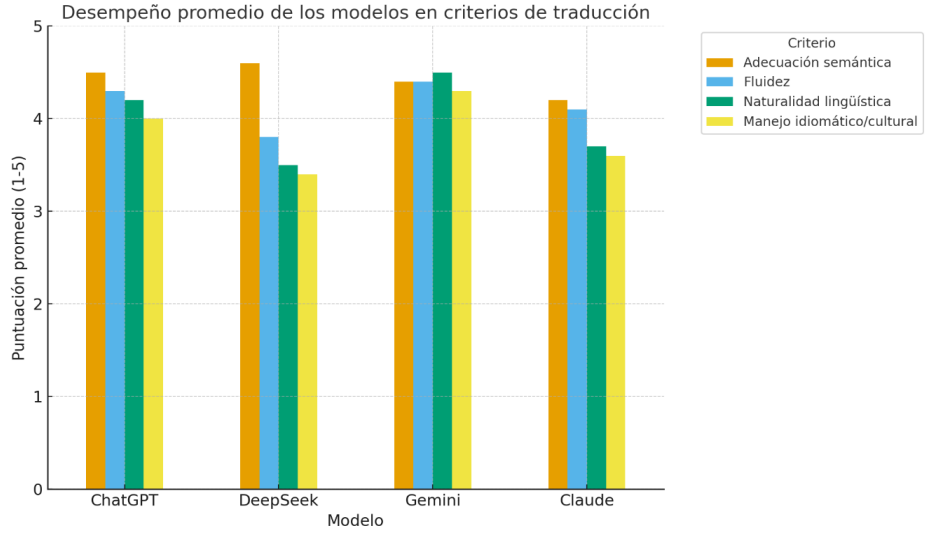


Figura 1. Desempeño promedio de los modelos en criterios de traducción.

Por otra parte, la figura 2 ilustra las diferencias de perfil: *Gemini* y *ChatGPT* aparecen más balanceados, *Claude* más conservador y *DeepSeek* más extremo en su énfasis semántico frente a su debilidad pragmática.

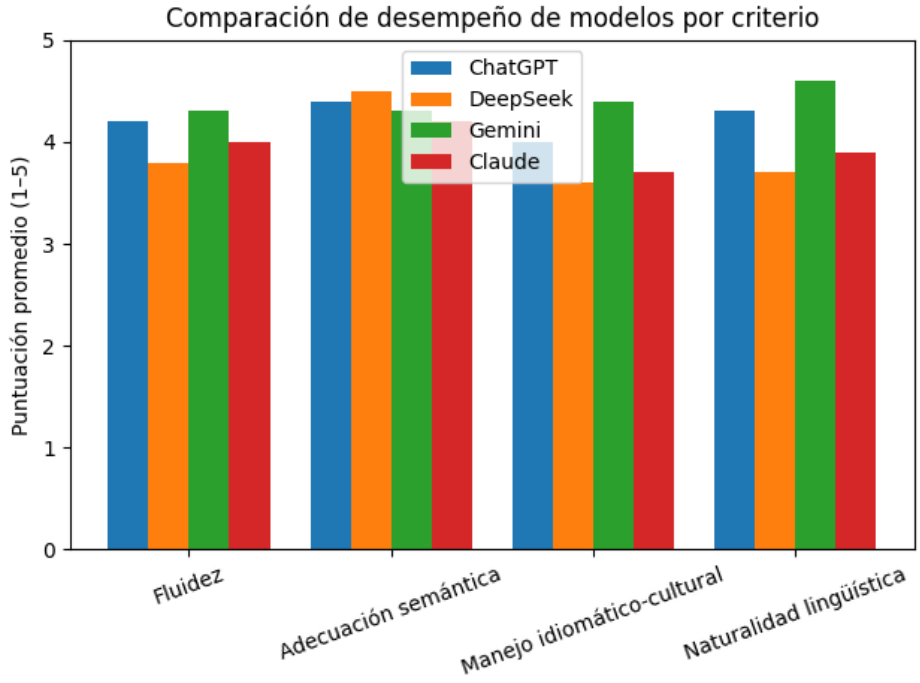


Figura 2. Perfil comparativo de desempeño de los modelos.

4.3. Resultados por criterio

En cuanto a la adecuación semántica, todos los modelos alcanzaron un desempeño alto (≥4.2), lo que confirma su capacidad general para preservar el contenido del texto original. DeepSeek destacó con la puntuación más elevada (4.6), lo que evidencia su fortaleza en la precisión conceptual. Sin embargo, en varios casos se identificaron fenómenos de sobreliteralidad: aunque el sentido denotativo se mantenía, la expresividad del enunciado en la lengua meta se veía empobrecida, lo que limitaba su naturalidad.

Respecto a la fluidez, los modelos Gemini (4.4) y ChatGPT (4.3) obtuvieron los resultados más altos, produciendo traducciones gramaticalmente correctas, cohesionadas y cercanas al uso natural del español. En contraste, DeepSeek (3.8) mostró oraciones menos cohesivas y con cierta rigidez, mientras que Claude (4.1) se mantuvo en un nivel

aceptable, aunque con un estilo más estructurado y previsible, lo que redujo la variedad expresiva.

En el criterio de naturalidad lingüística, se observaron las diferencias más notorias. Gemini (4.5) produjo traducciones altamente cercanas al español coloquial, incluyendo expresiones propias de la oralidad, como «estuvo padrísima», que reflejan sensibilidad cultural y pragmática. ChatGPT (4.2) también logró un desempeño sólido, aunque sus soluciones tendieron a ser más neutras. Por el contrario, DeepSeek (3.5) y Claude (3.7) presentaron limitaciones notables: sus traducciones, si bien correctas desde el punto de vista gramatical, resultaron poco representativas del español coloquial auténtico, con estructuras más propias del registro formal.

Finalmente, en el manejo idiomático y cultural se identificaron diferencias significativas entre los modelos. Gemini (4.3) y ChatGPT (4.0) lograron trasladar con mayor éxito expresiones figurativas y referencias culturales, ofreciendo traducciones más cercanas al uso nativo. En contraste, DeepSeek (3.4) y Claude (3.6) tendieron a recurrir a soluciones literales o excesivamente conservadoras, lo que redujo la adecuación cultural de sus resultados y los distanció del registro coloquial.

4.4. Análisis cualitativo

El análisis cualitativo permitió identificar patrones diferenciados de comportamiento entre los modelos, tanto en la resolución de expresiones idiomáticas como en el tratamiento de elementos pragmáticos y culturales. Los ejemplos presentados a continuación ilustran las tendencias más representativas observadas durante la evaluación.

El examen detallado de las traducciones permitió identificar tendencias específicas en el comportamiento de los modelos. En el caso de ChatGPT, se observó una capacidad destacada para traducir con sensibilidad cultural, adaptando expresiones idiomáticas al español coloquial. Por ejemplo, la frase «spill the beans”» fue traducida como «soltar la sopa», lo que refleja una solución idiomática adecuada. Sin embargo, en otros contextos el sistema mantuvo estructuras demasiado cercanas al inglés, como al traducir «under the weather» de manera literal como «debajo del clima», lo que resulta inadecuado en el español coloquial. DeepSeek, por su parte, demostró un alto nivel de precisión conceptual, pero una limitada flexibilidad pragmática. Su tendencia a la literalidad se hizo evidente en expresiones de jerga juvenil. Un ejemplo

claro fue «That party was lit», traducido como «esa fiesta estuvo encendida». Aunque el sentido denotativo es correcto, la frase pierde la carga cultural y expresiva del original.

En contraste, Gemini mostró la mayor capacidad para integrar el contexto y adaptar el registro. En el mismo ejemplo, «That party was lit» fue traducido como «esa fiesta estuvo padrísima», una solución natural, idiomática y culturalmente adecuada, que reproduce fielmente el tono juvenil del original. Finalmente, Claude se inclinó hacia un estilo más conservador y seguro, priorizando la coherencia semántica sobre la expresividad. Esto se evidenció en la traducción de «kick the bucket», resuelta como «perder la vida». Aunque correcta y clara desde el punto de vista semántico, esta opción es más formal y menos idiomática que alternativas como «estirar la pata».

4.5. Patrones de error

El análisis cualitativo permitió identificar tres patrones de error recurrentes en los modelos evaluados. El primero fue la literalidad idiomática, observable en casos como «let the cat out of the bag», traducido como «dejar salir al gato de la bolsa». Este error refleja la dificultad de algunos sistemas para reconocer y adaptar expresiones figurativas. El segundo patrón fue la rigidez estilística, particularmente visible en la traducción de marcadores pragmáticos como «you know». En varios casos, se tradujo como «tú sabes» sin un ajuste pragmático al español, lo que genera una expresión poco natural y distante del registro coloquial.

Los errores identificados fueron clasificados de acuerdo con su efecto predominante sobre la calidad traductora: (a) literalidad idiomática, cuando la expresión original fue traducida palabra por palabra sin conservar su significado figurado; (b) rigidez estilística, cuando la traducción preservó el significado básico, pero produjo construcciones poco naturales en español; y (c) neutralización cultural, cuando se eliminaron o atenuaron rasgos propios del contexto sociocultural de la expresión original. Esta clasificación permitió sistematizar el análisis cualitativo y comparar los comportamientos observados entre los distintos modelos.

El tercer patrón correspondió a la neutralización cultural, evidenciado en la traducción de referencias socioculturales como «The American

dream», que en ocasiones apareció como «el sueño estadounidense» en lugar de «el sueño americano».

Esta neutralización, aunque semánticamente válida, pierde la fuerza cultural que caracteriza la expresión original. Estos errores fueron más frecuentes en DeepSeek y Claude, mientras que ChatGPT y Gemini mostraron un mejor manejo de estrategias de adaptación, logrando resultados más idiomáticos y cercanos al español coloquial.

4.6. Síntesis de hallazgos

En conjunto, los resultados obtenidos sugieren que Gemin*i* presentó el desempeño promedio más alto dentro del corpus analizado, seguido por ChatGPT, mientras que DeepSeek y Claude mostraron fortalezas y limitaciones diferenciadas según el criterio de evaluación considerado. ChatGPT también mostró un rendimiento sólido y flexible, aunque todavía vulnerable a producciones demasiado literales en ciertos contextos. DeepSeek destacó en precisión semántica, pero presentó debilidades notorias en registros informales, donde su rigidez estilística afectó la naturalidad de las traducciones. Finalmente, Claude se caracterizó por garantizar coherencia y seguridad en sus salidas, pero con una clara tendencia a la neutralidad cultural y la pérdida de expresividad idiomática.

En conjunto, los resultados respaldan la existencia de diferencias en el desempeño de los modelos evaluados dentro del corpus analizado. Gemini y ChatGPT obtuvieron las puntuaciones promedio más altas en naturalidad lingüística y manejo idiomático-cultural, mientras que DeepSeek destacó en adecuación semántica y Claude mostró un comportamiento más conservador.

Estas diferencias deben interpretarse como tendencias observadas en las condiciones específicas del estudio y constituyen evidencia descriptiva útil para comprender el comportamiento comparativo de los modelos, sin que ello implique establecer relaciones causales entre dicho desempeño y las características técnicas de sus arquitecturas.

5. Discusión

Los resultados obtenidos evidencian diferencias en el comportamiento de los modelos de lenguaje generativo frente a la traducción del habla

coloquial, particularmente en aspectos relacionados con la naturalidad lingüística, el manejo de expresiones idiomáticas y la adecuación pragmática. En términos descriptivos, Gemini y ChatGPT obtuvieron las puntuaciones promedio más altas en varios de los criterios evaluados, mientras que DeepSeek y Claude presentaron fortalezas y limitaciones diferenciadas según el fenómeno lingüístico analizado.

Estos hallazgos deben interpretarse como tendencias observadas dentro del corpus estudiado y no como evidencia suficiente para establecer relaciones causales entre el desempeño de los modelos y características específicas de sus arquitecturas o procesos de entrenamiento. En consecuencia, los resultados permiten describir diferencias comparativas entre los sistemas evaluados, pero no explicar los mecanismos internos que las originan.

Estos hallazgos deben interpretarse como tendencias observadas dentro del corpus estudiado y no como evidencia suficiente para establecer relaciones causales entre el desempeño de los modelos y características específicas de sus arquitecturas o procesos de entrenamiento. En consecuencia, los resultados permiten describir diferencias comparativas entre los sistemas evaluados, pero no explicar los mecanismos internos que las originan.

En términos de investigación en traducción y estudios del lenguaje, los hallazgos del presente estudio evidencian la necesidad de profundizar en el análisis empírico de la competencia pragmática e intercultural en contextos de traducción asistida por modelos de lenguaje generativo, especialmente en registros coloquiales y no normativos. Asimismo, los resultados invitan a revisar los modelos tradicionales de evaluación de la calidad traductora, incorporando dimensiones como la naturalidad percibida, la adecuación sociocultural y el esfuerzo cognitivo implicado en los procesos de posedición, aspectos centrales en la investigación reciente sobre traducción asistida por IA y estudios procesuales.

A modo de ejemplo, futuras investigaciones podrían comparar el desempeño de traductores en formación y profesionales al poseditar traducciones de habla coloquial generadas por distintos MEL, analizando variables como el tiempo de revisión, el tipo de intervenciones pragmáticas y la percepción de naturalidad por parte de evaluadores expertos. Desde esta perspectiva, los MEL se conciben no como sustitutos del traductor humano, sino como artefactos mediadores cuya eficacia depende de la competencia traductora del usuario, lo que refuerza la pertinencia de

enfoques descriptivos, procesuales y centrados en el usuario en la investigación traductológica contemporánea.

De la misma forma, estos modelos pueden ser utilizados como herramientas de apoyo en la traducción de textos formales o semiformales; sin embargo, su aplicación directa al habla coloquial requiere necesariamente de procesos de posedición experta para garantizar la adecuación cultural, la naturalidad discursiva y la funcionalidad pragmática del texto meta. En este tipo de registros, la posedición no se limita a la corrección léxico-gramatical, sino que implica intervenciones profundas de carácter pragmático y sociocultural, tales como la reformulación de coloquialismos, la adaptación de marcadores discursivos, el ajuste del grado de informalidad o la sustitución de referencias culturales implícitas que los modelos de lenguaje tienden a resolver de manera literal o neutralizada.

A modo de ejemplo, una salida generada por un MEL que traduzca expresiones coloquiales con equivalentes semánticamente correctos pero pragmáticamente inadecuados puede requerir que el traductor reconfigure la expresión para mantener efectos como la cercanía, la ironía o la atenuación propias del contexto comunicativo original. Desde esta perspectiva, la posedición se consolida como una competencia clave en la traducción asistida por IA, cuyo estudio resulta fundamental tanto para la práctica profesional como para la investigación en traducción, al permitir analizar el esfuerzo cognitivo, las estrategias de intervención y la toma de decisiones pragmáticas que el traductor despliega al interactuar con modelos de lenguaje generativo. Además, la comparación evidencia que la práctica traductora no puede desligarse de la mediación intercultural: los sistemas automatizados aún no son capaces de sustituir la creatividad y sensibilidad cultural que aporta el traductor humano. En este sentido, los resultados refuerzan la importancia de formar a traductores capaces de integrar críticamente la IA en su quehacer profesional.

El estudio, sin embargo, presenta limitaciones. El corpus, aunque representativo, estuvo restringido a cuarenta expresiones y se centró exclusivamente en la traducción del inglés al español, lo que limita la generalización de los resultados a otros pares de lenguas. Si bien el análisis se centra en expresiones aisladas y no en fragmentos discursivos extensos, esta elección metodológica no contradice el enfoque de la traducción como mediación intercultural. Por el contrario, parte del supuesto de que muchas de las tensiones interculturales más relevantes en el habla coloquial se manifiestan en unidades microdiscursivas altamente convencionales, tales

como coloquialismos, fórmulas rutinarias o marcadores pragmáticos, cuyo significado y función dependen en gran medida del contexto sociocultural implícito. El análisis de estas unidades permite observar con mayor precisión hasta qué punto los modelos de lenguaje generativo son capaces de activar inferencias pragmáticas y culturales sin apoyarse en un contexto discursivo amplio, lo que resulta particularmente pertinente para evaluar sus límites como herramientas de mediación intercultural.

Otra limitación del estudio radica en el tamaño del corpus y en la ausencia de análisis estadísticos inferenciales, decisiones metodológicas coherentes con el carácter exploratorio de la investigación. En consecuencia, los resultados deben entenderse como evidencia descriptiva útil para generar hipótesis y orientar futuras investigaciones, más que como conclusiones generalizables sobre el comportamiento de los modelos de lenguaje generativo.

En este sentido, el estudio no busca describir el desempeño global de los modelos en la traducción de textos completos, sino aislar y examinar microdecisiones traductoras que suelen requerir una intervención experta durante la posedición. Este enfoque analítico permite identificar patrones de literalidad, neutralización o pérdida pragmática que, en textos más extensos, podrían quedar diluidos por la coherencia discursiva general, pero que son cruciales para comprender el papel del traductor humano como mediador intercultural. Finalmente, la naturaleza en constante evolución de los MEL implica que los resultados representan un corte temporal, por lo que futuras actualizaciones de los modelos podrían modificar sustancialmente el panorama observado.

6. CONCLUSIONES

En síntesis, los resultados obtenidos sugieren que, dentro del corpus analizado, Gemini presentó el desempeño promedio más alto en la traducción del habla coloquial, seguido de ChatGPT, mientras que DeepSeek y Claude mostraron fortalezas y limitaciones diferenciadas según el criterio de evaluación considerado. Estos hallazgos permiten caracterizar con mayor claridad el comportamiento comparativo de los modelos evaluados, así como identificar fortalezas y limitaciones diferenciadas en la traducción del habla coloquial.

La principal contribución de este estudio consiste en ofrecer evidencia empírica sobre el comportamiento comparativo de cuatro modelos de lenguaje generativo en la traducción del habla coloquial desde una

perspectiva traductológica sustentada en evaluación humana. A diferencia de investigaciones centradas exclusivamente en métricas automáticas o en textos especializados, este trabajo incorpora criterios relacionados con la naturalidad lingüística, la adecuación pragmática y la mediación intercultural, aportando evidencia útil tanto para la investigación como para la práctica profesional de la traducción asistida por IA .

De cara al futuro, se recomienda ampliar el corpus de análisis para incluir interacciones discursivas más extensas y contextos multimodales (subtitulación, doblaje, chats en línea), así como explorar otros pares de lenguas que presenten mayores retos de divergencia cultural. Igualmente, será relevante diseñar métricas de evaluación híbridas que integren no solo fidelidad semántica, sino también adecuación cultural y percepción de usuarios nativos. Finalmente, se sugiere dar continuidad a estudios longitudinales que permitan observar cómo las constantes actualizaciones de los modelos de IA influyen en su desempeño traductológico.

En conjunto, los resultados respaldan la necesidad de continuar desarrollando investigaciones que integren evaluación humana, análisis traductológico y metodologías comparativas para comprender con mayor precisión el papel que desempeñan los modelos de lenguaje generativo en la traducción profesional. En este contexto, la IA debe entenderse como una herramienta de apoyo al traductor humano y no como un sustituto de los procesos de mediación lingüística, pragmática e intercultural que caracterizan la práctica profesional de la traducción.

REFERENCIAS BIBLIOGRÁFICAS

Achiam, Josh, Adler, Steven, Agarwal, Sandhini, Ahmad, Lama, Akkaya, Ilge, Leoni Aleman, Florencia. OpenAI. (2023). *GPT-4 Technical Report*. arXiv. OpenAI. (2023). *GPT-4 Technical Report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Andrade Preciado, Jahiro Samar (2024). *La traducción de coloquialismos en textos audiovisuales: Estudio de la adquisición de la competencia traductora* (1.ª ed.). Editorial Fontamara.

Anthropic (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku*. Anthropic. https://www.anthropic.com

Bai, Yuntao, Kadavath, Saurav, Kundu, Sandipan, Askell, Amanda, Kernion, Jackson, Jones, Andy, Amodei, Dario (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv.. https://doi.org/10.48550/arXiv.2212.08073

Beeching, Kate (2016). *Pragmatic Markers in British English: Meaning in Social Interaction*. Cambridge University Press.

Bowker, Lynne y Buitrago Ciro, Jairo (2021). *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing. https://doi.org/10.1108/9781800435263

Castilho, Sheila, Guerreiro, Nuno M., Marinho, Zita y Way, Andy (2023). Is neural machine translation the new state of the art? A critical analysis of quality, fluency and acceptability. *Language Resources and Evaluation,* 57 (2), 513-543. https://doi.org/10.1007/s10579-022-09596-4

Castilho, Sheila, Moorkens, Joss, Gaspari, Federico, Sennrich, Rico, Way, Andy y Georgakopoulou, Panayota (2018). Evaluating machine translation in a user-centered paradigm: Reconciling usability and translation quality. *The International Journal of Translation and Interpreting Research*, 10 (1), 1-20. https://doi.org/10.12807/ti.110201.2018.a01

European Commission (2022). *European Master's in Translation (EMT) Competence Framework*. Publications Office of the European Union. https://commission.europa.eu/education/skills-and-qualifications/education-and-training/emt_en

Gaspari, Federico, Toral, Antonio, Naskar, Sudip Kumar, Groves, Declan y Way, Andy (2022). Perception vs. reality: Measuring machine translation post-editing effort. *Machine Translation*, 36 (1), 1-29. https://doi.org/10.1007/s10590-021-09266-6

Gemini Team. (2024). *Gemini 1.5: Unlocking Multimodal Understanding and Long-Context Reasoning*. DeepMind.

Hansen-Schirra, Silvia, Hofmann, Sascha, Nitzke, Jean y Neumann, Stella (2021). *Post-Editing of Machine Translation: Processes and Applications*. Routledge. https://doi.org/10.4324/9781003038179

House, Juliane (2021). Translation as communication across languages and cultures. En Sue-Ann Harding y Ovidi Carbonell Cortés (Ed.), *The Routledge Handbook of Translation and Culture* (pp. 11-25). Routledge. https://doi.org/10.4324/9780367820739

Kenny, Dorothy y Doherty, Stephen (2023). Ethics and governance of artificial intelligence in translation. *Translation Spaces*, 12 (1), 1-26. https://doi.org/10.1075/ts.22018.ken

Liu, Aixin, Dong, Zhihong, Zheng, Zihui, Ma, Chenggang, Li, Xingkai, He, Jiawei y Liu, Zhiwei (2024). *DeepSeek-V2: A strong, Economical, and Efficient Mixture-of-Experts Language Model*. arXiv. https://doi.org/10.48550/arXiv.2405.04434

Liu, Emmy, Chaudhary, Aditi y Neubig, Graham (2023). Crossing the threshold: Idiomatic machine translation and the role of large language models. En Bouamor, Houda, Pino, Juan y Bali, Kalika (Eds.), *Proceedings of EMNLP 2023* (pp. 15095–15111). Association for Computational Linguistics.

Michel, Paul y Neubig, Graham (2018). MTNT: A testbed for machine translation of noisy text. En Riloff, Ellen, Chiang, David, Hockenmaier, Julia y Tsujii, Jun'ichi (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 543–553). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/D18-1050

Moorkens, Joss (2020). A usability study of post-editing productivity. *Machine Translation*, 34 (3), 257-275. https://doi.org/10.1007/s10590-020-09246-5

Naveen, Palanichamy y Trojovský, Pavel (2024). Overview and challenges of machine translation for contextually appropriate translations. *Heliyon*, 10 (2)

O'Brien, Sharon (2022). Cognitive effort in translation and post-editing. En Alves, Fabio y Jakobsen, Arnt Lykke (Eds.), *The Routledge Handbook of Translation and Cognition* (pp. 191-205). Routledge. https://doi.org/10.4324/9781003177984-14

Ouyang, Long, Wu, Jeffrey, Jiang, Xu, Almeida, Diogo, Wainwright, Carroll L., Mishkin, Pamela, Zhang, Chong, Agarwal, Sandhini, Slama, Katarina, Ray, Alex, Schulman, John, Hilton, Jacob, Kelton, Fraser, Miller, Luke, Simens, Maddie, Askell, Amanda, Welinder, Peter, Christiano, Paul F., Leike, Jan y Lowe, Ryan (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. Neural Information Processing Systems Foundation, Inc.PACTE Group. (2020). *Competence Levels in Translation: Working Towards a European Framework*. John Benjamins.

Pramodya, Ashmari, Sakai, Yusuke, Vasselli, Justin, Kamigaito, Hidetaka y Watanabe, Taro (2025). Translating Movie Subtitles by Large Language Models using Movie-meta Information. En Zhao, Jin, Wang, Mingyang y Liu, Zhu (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 315–330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-srw.22

Pym, Anthony (2021). *Exploring Translation Theories* (3.ª ed.). Routledge.

Toury, Gideon (2012). *Descriptive Translation Studies – and Beyond*. John Benjamins. (Primera edición de esta obra publicada en 1995).

Vieira, Lucas Nunes, O'Brien, Sharon y Teixeira, Carlos S. C. (2024). User perception of quality and trust in machine translation and large language models. *Translation Spaces*, 13 (1), 1-25. https://doi.org/10.1075/ts.22023.vie

Vieira, Lucas Nunes, O'Brien, Sharon y Teixeira, Carlos S. C. (2024). User perception of quality and trust in machine translation and large

language models. *Translation Spaces*, 13 (1), 1-25. https://doi.org/10.1075/ts.22023.vie

Zhang, Li, Wang, Chang y Zhao, Harry (2024). *Evaluating the Translation Performance of Large Language Models in Specialized Domains*. arXiv. https://doi.org/10.48550/arXiv.2402.12345