¿La convención de quién?

Cómo interpretan los modelos de lenguaje las cifras y las fechas cuando dependen de las convenciones de cada país

Cuaderno de investigación

Autor: Manuel Jesús Gómez Zotano

Director de Sistemas e Innovación, RTVE

Doctor en Informática por la Universidad Complutense de Madrid

Fecha: octubre de 2026

¿La convención de quién?

Índice

Resumen.....3

1. El problema.....4

2. Convenciones numéricas y de fechas.....5

3. Cómo se realizó el estudio.....6

4. Modelos evaluados.....9

5. Resultados.....10

6. El sorprendente comportamiento de las fechas.....14

7. Pruebas con noticias reales y con la Comisión Europea.....15

8. Comparación con personas.....18

9. Implicaciones para los medios.....19

10. Aplicación en RTVE.....19

11. Limitaciones.....21

Conclusiones.....21

Referencias.....21

Resumen

Los modelos de lenguaje ya leen, resumen, traducen y locutan noticias. En ese trabajo, una cifra mal interpretada no es un error de estilo, sino un error de información. Y muchas cifras no tienen un único valor posible: «1.500» son mil quinientos en España y uno y medio en Estados Unidos; «04/07/2027» es el 4 de julio en Madrid y en Londres, y el 7 de abril en Washington.

Este cuaderno presenta un estudio sobre cómo interpretan esas expresiones distintos modelos de lenguaje actuales, comerciales y de pesos abiertos ejecutados en local. El estudio evalúa doce configuraciones en su fase inicial y ocho configuraciones representativas en el banco de pruebas final. Hemos construido un banco de pruebas controlado de 561 casos en español de España, inglés británico e inglés estadounidense, y lo hemos completado con tres validaciones: 547 cifras de noticias publicadas en rtve.es, revisadas por dos personas; 557 cifras de comunicados bilingües de la Comisión Europea; y una prueba a ciegas con cuatro periodistas y documentalistas de RTVE.

Los resultados principales son cinco:

1. **Los modelos conocen las convenciones.** En noticias ya publicadas aciertan entre el 94 % y el 99,5 % de las cifras, y adaptan al español las escalas inglesas (*billion*, *trillion*) con al menos un 99,8 % de acierto.
2. **Fallan al decidir qué convención aplicar.** Cuando una página española cita literalmente una fecha de un documento estadounidense, ningún sistema aplica de forma fiable por defecto la convención de la fuente: aciertan entre el 0 % y el 28 % de los casos. Con números en la misma situación, entre el 89 % y el 100 %.
3. **Fallan de dos maneras.** Los modelos comerciales responden «ambiguo»: ven el problema, pero no lo resuelven. Los locales siguen la convención de la página y dan una fecha equivocada, y razonar lo empeora.
4. **Una instrucción en el prompt corrige a los comerciales** (82–98 %), **pero no a los locales** (26–41 %).
5. **La prueba con personas reproduce el patrón.** Los cuatro profesionales que participaron acertaron el 28 % de las fechas citadas en conflicto y el 95 % de los números, con estrategias de error parecidas a las de los modelos.

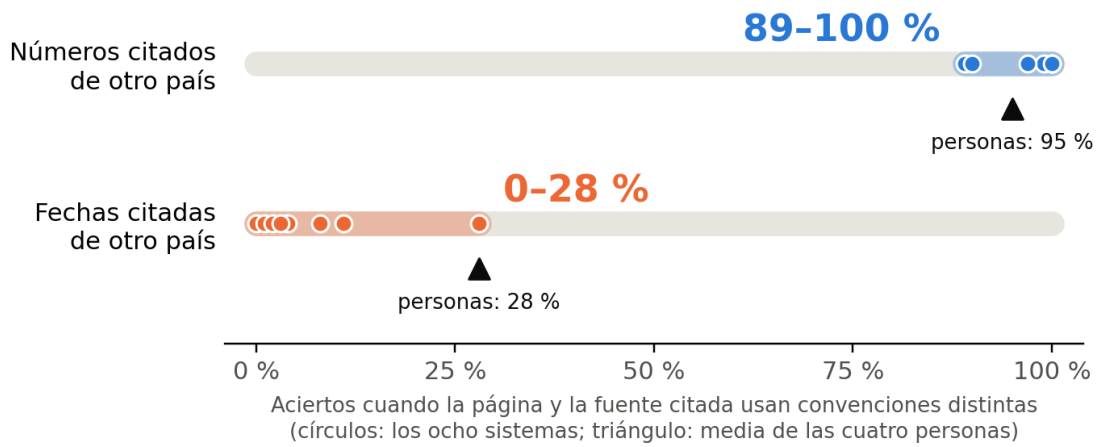


Figura 1. La cifra clave del estudio: cuando la página y la fuente citada usan convenciones distintas, los sistemas aciertan casi siempre con los números y casi nunca con las fechas. Las personas muestran la misma diferencia.

El cuaderno termina con recomendaciones concretas para el uso de estos sistemas en una redacción y, en particular, en RTVE.

1. El problema

1.1 Un ejemplo

Una noticia de una web española dice:

> *El memorando del FBI, con fecha «02/12/2025», recoge las primeras pesquisas.*

La fecha va entre comillas: es una copia literal del documento. En la convención estadounidense, las fechas numéricas escriben primero el mes. La fecha es, por tanto, el **12 de febrero de 2025**, no el 2 de diciembre.

Un periodista español que lea deprisa entenderá «2 de diciembre». Un modelo de lenguaje al que se le pida resumir, traducir o locutar esa noticia también puede hacerlo. Y si la noticia pasa a otro texto, a un rótulo o a una locución automática, el error se propaga.

1.2 Por qué importa en un medio

En una redacción, las cifras no pueden fallar. Una fecha, un importe o una escala equivocados son errores de información, y muchas veces pasan inadvertidos porque el texto sigue siendo correcto y fluido.

Las métricas habituales para evaluar la generación automática de texto miden el parecido entre el texto producido y uno de referencia. Una cifra cambiada altera muy poco ese parecido, porque es una palabra entre muchas. Por eso conviene evaluar directamente si el sistema **entiende el valor** de cada cifra, y no solo si el texto se parece al esperado.

El problema aparece en tareas muy distintas de un medio:

- **Redacción y edición** asistidas, cuando un modelo resume o reescribe material de agencias o de fuentes extranjeras.
- **Traducción y adaptación** de textos entre español e inglés.
- **Subtitulado y accesibilidad**, donde las cifras se transcriben o se adaptan.
- **Locución automática y síntesis de voz**, que deben decidir si «2.30» es una hora o un decimal, o cómo leer «1,5 M€».
- **Documentación y archivo**, donde las fechas y cifras se extraen para indexar y buscar.

1.3 La pregunta del estudio

¿Interpretan correctamente los modelos de lenguaje el valor de una expresión numérica cuando ese valor depende del formato, de las convenciones del país o del contexto?

Separamos tres capacidades:

1. **Conocer las convenciones**: saber que en España «1.500» son mil quinientos y que en Estados Unidos «04/07» es el 7 de abril.
2. **Elegir la convención correcta** a partir del contexto: la página, el texto o la fuente citada.
3. **Reconocer cuándo no se puede decidir**: un «1.500» aislado, sin idioma ni contexto, es ambiguo.

La evaluación es exacta: el valor que da el sistema es correcto o no lo es.

2. Convenciones numéricas y de fechas

2.1 Las diferencias que importan

Las tres variantes estudiadas (español de España, inglés británico e inglés estadounidense) difieren en varios puntos:

Aspecto	España	Reino Unido	Estados Unidos
Separador de miles	punto o espacio: 15.000	coma: 15,000	coma: 15,000
Separador decimal	coma: 2,5	punto: 2.5	punto: 2.5
Fecha numérica	día/mes: 04/07 = 4 de julio	día/mes: 04/07 = 4 de julio	**mes/día**: 04/07 = 7 de abril
Mil millones	mil millones	billion	billion
Un millón de millones	**un billón**	trillion	trillion
Abreviatura de millones	M, M€	m, £45m	M, $3.5M

Dos detalles generan muchos errores:

- **Reino Unido y España escriben la fecha igual** (día/mes); es Estados Unidos quien la escribe al revés. Lo que diferencia la fecha no es el idioma, sino el país.

- **«Billion» no es «billón».** Un *billion* inglés son mil millones; un billón español es un millón de millones. Es el falso amigo clásico de la traducción periodística.

La Real Academia Española recomienda el espacio fino como separador de miles y admite tanto la coma como el punto decimal según la zona. Por eso «1.500» puede ser ambiguo incluso dentro del español.

2.2 Cifras que no son cantidades

Muchas cifras de las noticias no son cantidades: «COVID-19», «Boeing 737», «artículo 100.2», «AP-7», «11-M», «H-1B». Un sistema que las trate como números («once millones» por «11-M») se equivoca.

2.3 Cuándo el contexto decide

La idea central del estudio es que **el mismo fragmento puede tener valores distintos según el contexto**. Distinguimos cinco niveles:

- **Explícito.** El texto declara la convención: «publicación que sigue las convenciones de España».
- **Metadatos.** Solo la información técnica de la página (el idioma declarado y la dirección web) permite decidir; el fragmento es neutro: «Total: 1,375».
- **Implícito.** El texto delata el país: «Ofgem will set the new price cap on 03/08/2026». Ofgem es el regulador británico de la energía, así que es el 3 de agosto.
- **Conflicto.** La página es de un país y la expresión se cita literalmente de una fuente de otro, como el memorando del FBI.
- **Sin contexto.** Un fragmento aislado («Ratio: 4.915»). Lo correcto es responder que es ambiguo y dar las dos lecturas posibles.

3. Cómo se realizó el estudio

3.1 Por qué un banco de pruebas sintético

Si se toman frases reales al azar, salen muchos casos fáciles, pocos difíciles y ninguna garantía sobre el valor correcto. Un generador de frases a partir de plantillas permite:

- cubrir de forma equilibrada cada fenómeno y cada país;
- conocer con certeza el valor correcto de cada caso;
- cambiar una sola cosa cada vez: el contexto, la pista sobre el origen o el idioma de la cita.

El coste es que las frases no son reales. Por eso el estudio se completa con noticias publicadas, comunicados oficiales y una prueba con personas (secciones 7 y 8).

3.2 Los fenómenos estudiados

Partimos de una clasificación de 44 fenómenos en siete familias: cantidades y separadores; escalas; porcentajes, monedas y unidades; fechas y horas; rangos, ordinales y números

romanos; números en palabras; y cifras que no son cantidades. Elegimos 15 para cubrir seis de las siete familias (todas salvo los números en palabras) con las trampas más relevantes para la prensa. Unos dependen de la convención o del contexto; otros sirven de control porque su valor no depende de ella, como los romanos o las cifras que no son cantidades.

Código	Fenómeno	Ejemplo	Trampa
A2	Separador de miles	15.000 · 15,000	Leerlo como decimal
A4	Separador ambiguo	1.500 · 1,500	Mil quinientos o uno coma cinco
A5	Miles y decimales juntos	1.234,56 · 1,234.56	Invertir punto y coma
B1	Escala en palabras	1,5 millones · 1.5 million	Orden de magnitud
B2	Escala abreviada	340 M€ · £45m · $3.5B	Expandir mal
B3	Falsos amigos de escala	un billón · a billion	10^{12} frente a 10^{9}
B5	«m» ambigua	£300m · 300 m	Millones o metros
C2	Puntos frente a porcentaje	sube 2 puntos · sube un 2 %	No son lo mismo
C3	Moneda con símbolo	85,20 $ · C$20 · MX$1,500	«$» puede ser dólar, peso…
D1	Fecha numérica	04/07/2027	Día y mes intercambiables
D5	Hora	2.30 · 2:30 p.m. · a las 8	Decimal u hora; 12 o 24 h
D6	Fechas-evento	11-M · 23-F · 9/11	«11-M» no son 11 millones
E1	Rangos	18-30 años · £3-5bn	La escala afecta a los dos extremos
E3	Números romanos	siglo XXI · Henry VIII	Reconocer el ordinal
G2	Cifras que no son cantidades	COVID-19 · Boeing 737	No tratarlas como cantidad

3.3 El valor correcto de cada caso

Cada caso guarda su valor en un formato independiente de cómo esté escrito: el tipo (cantidad, porcentaje, puntos porcentuales, fecha, hora, rango, ordinal, cifra que no es cantidad o ambigua), el valor exacto y la unidad. Así, «1,5 M€» y «1.5 million euros» tienen el mismo valor: un millón quinientos mil euros. Un caso ambiguo guarda todas sus lecturas posibles.

3.4 Los conflictos entre página y fuente

Es la parte central del diseño. En un caso de conflicto, la página es de un país y la expresión se cita **literalmente, entre comillas**, de una fuente de otro. La respuesta correcta es la convención de la fuente.

Es una **definición de la tarea**, no una afirmación sobre cómo trabaja siempre el periodismo: un periodista puede adaptar una fecha al citarla. Precisamente por eso, todas las cifras en conflicto añadidas en la versión final van entre comillas: así queda claro que son copia literal de la fuente.

Los conflictos cubren las seis combinaciones posibles entre España, Reino Unido y Estados Unidos, con fechas y con números. El origen de la cita se indica de tres maneras: con el país («el Departamento de Trabajo de Estados Unidos»), con una institución («el FBI») o con un lugar («un juzgado de Texas»).

Hay también **casos de control**: una fuente extranjera que usa la misma convención que la página, como una fecha británica citada en una web española. Sirven para comprobar que el sistema no cambia de lectura solo porque aparece un país extranjero.

Un control añadido durante el estudio. En la primera versión, los conflictos de números citaban una frase entera en el idioma de la fuente («*«8,750 weapons were seized»*»), mientras que en los de fechas solo iba entre comillas la fecha. Las palabras en inglés son una pista muy fuerte de que la cita sigue las convenciones inglesas, así que la diferencia entre fechas y números podía deberse a esa pista. Añadimos las dos combinaciones que faltaban, con fechas citadas junto con texto de la fuente y números citados solos, para separar los dos efectos.

3.5 Tareas y puntuación

Cada caso se plantea como una pregunta: «¿Qué valor representa la expresión marcada en este texto?». El sistema responde en un formato estructurado con el tipo, el valor y la unidad, o con «ambiguo» y sus lecturas. En la fase piloto evaluamos además otras dos tareas sobre las mismas expresiones:

- **Verbalizar**: escribir cómo se leería en voz alta. Un programa propio, con 134 pruebas, convierte esa lectura de nuevo en número para comprobarla.
- **Equivalencia**: decidir si una versión escrita en el otro idioma expresa el mismo valor. Las versiones falsas contienen errores típicos: escala multiplicada por mil, separador leído al revés, un dígito cambiado o día y mes intercambiados.

La puntuación es exacta y automática:

- Una respuesta es correcta si coinciden el tipo, el valor y la unidad.
- En los casos ambiguos, lo correcto es responder «ambiguo». En los demás, responder «ambiguo» cuenta como error (una falsa alarma).
- En las fechas-evento sin año («23-J») basta con el día y el mes.

Cada caso se pregunta dos veces, con las instrucciones en español y en inglés; el texto evaluado va siempre en su idioma original.

Intervalos de confianza. Las dos respuestas a un mismo caso no son independientes, y en la prueba humana cada persona responde a los mismos casos. Por eso los intervalos de confianza al 95 % se calculan remuestreando casos (y, en la prueba humana, personas y casos), en lugar de tratar cada respuesta como independiente.

3.6 Versiones del banco de pruebas

Versión	Casos	Qué añade
Piloto	309 (+179 pares de equivalencia)	Los 15 fenómenos, con todos los niveles de contexto en los que el contexto decide
Final	513	Conflictos en todas las direcciones, con fechas y números, tres tipos de pista y controles; más contexto

Versión	Casos	Qué añade
		implícito, metadatos y ambigüedad
Final + control de idioma	561	Fechas citadas con texto de la fuente y números citados solos

Cada versión conserva intactos los casos anteriores, así que todas las respuestas son comparables.

4. Modelos evaluados

Nombre en este cuaderno	Modelo	Dónde se ejecuta	Razonamiento
Qwen / Qwen (razona)	Qwen3.8-27B	En local	Desactivado / activado
Gemma / Gemma (razona)	Gemma 4 31B	En local	Desactivado / activado
GPT Luna / GPT Luna (razona)	GPT-6 Luna	API de OpenAI	Ninguno / medio
GPT Astra	GPT-6 Astra	API de OpenAI	Mínimo (no se puede desactivar)
Claude Sonnet / Sonnet (razona)	Claude Sonnet 5	API de Anthropic	Desactivado / adaptativo medio
Claude Opus	Claude Opus 5.5	API de Anthropic	Mínimo (no se puede desactivar)
Gemini Flash	Gemini 3.8 Flash	API de Google	Mínimo / medio

Los modelos **locales** (Gemma y Qwen) son de pesos abiertos y se ejecutaron en un único ordenador de trabajo, sin enviar datos fuera. Es la opción que una organización elegiría por coste o por confidencialidad. Los **comerciales** se usan a través de los servicios de cada empresa.

«Razonar» indica que se utilizó el modo de razonamiento ofrecido por el modelo antes de producir la respuesta final. En los modelos que lo permiten se probó con y sin él.

La fase piloto usó las doce configuraciones. Para la versión final se eligieron ocho representativas: Claude Opus, GPT Astra, Claude Sonnet sin razonar, Gemini Flash con el razonamiento mínimo, y Gemma y Qwen con y sin razonamiento.

En total se han puntuado casi 30.000 respuestas: unas 22.700 del banco de pruebas, 4.376 de noticias reales y unas 2.770 de comunicados de la Comisión. El coste en servicios comerciales fue modesto: entre 1 y 6 dólares por configuración en la fase piloto y unos 8 dólares para la versión final. Los modelos locales no tienen coste por uso, pero sí de tiempo: Qwen con razonamiento necesitó unas cinco horas para los 504 casos nuevos.

5. Resultados

5.1 Las convenciones se conocen

Sistema	Normalizar	Verbalizar	Equivalencia	Ambigüedad detectada	Falsas alarmas
Qwen	73	92	81	15	8,3
Gemma	84	91	88	36	3,0
GPT Luna	81	92	81	56	6,4
Claude Sonnet	85	95	91	53	6,8
Qwen (razona)	86	—	—	48	0,4
Gemma (razona)	89	—	—	55	1,9
GPT Luna (razona)	89	94	91	70	1,3
Claude Sonnet (razona)	87	95	93	49	3,6
Gemini Flash (mínimo)	88	94	85	81	5,5
Gemini Flash (medio)	91	95	84	85	5,8
GPT Astra	91	94	92	91	4,1
Claude Opus	**93**	**97**	**93**	**91**	5,5

Tabla 1. Fase piloto (309 casos): porcentaje de acierto en las tres tareas, ambigüedades reales detectadas y falsas alarmas en casos no ambiguos. Los modelos locales con razonamiento solo se evaluaron en normalizar.

Los fenómenos que solo exigen **conocer** una convención se resuelven casi siempre. Los separadores de miles, los miles con decimales y los números romanos alcanzan el 100 % en las doce configuraciones. En separadores y escalas solo hay cuatro errores de valor en total, todos de escala: tres de Qwen y uno de GPT Luna. Los demás fallos de esas familias son unidades omitidas, sobre todo de Gemini Flash.

Un resultado negativo. Esperábamos muchos errores con «billion» y «billón», y apenas los hay. Los falsos amigos de escala, que esperábamos problemáticos, aparecen ya como un error poco frecuente en los modelos actuales.

La dificultad se concentra donde hay que **elegir** una convención: las fechas numéricas (entre el 36 % y el 77 % de acierto según el sistema), el separador ambiguo, la «m» de millones o metros y las horas.

5.2 Resultado central: fechas frente a números en conflicto

	Opus	Astra	Sonnet	Gemini	Gemma	Gemma (r.)	Qwen	Qwen (r.)
Conflicto, fechas	**28**	0	4	8	11	1	2	3
↳ intervalo al 95 %	18–40	0 de 48 casos	1–8	3–15	4–20	0–3	0–5	0–7
↳ errores: sigue la página / «ambiguo» (de 96)	2 / 67	21 / 75	18 / 74	7 / 81	40 / 45	56 / 39	31 / 63	**85** / 8
↳ con la instrucción	**98**	96	82	**98**	41	—	26	—
Conflicto, números	100	100	89	100	99	100	90	97
Controles (misma convención)	100	100	33	100	83	100	67	100
Contexto implícito, fechas	64	8	15	38	42	**85**	56	78
↳ con la instrucción	71	11	41	86	70	—	69	—
Metadatos	92	98	92	97	78	100	40	100

	Opus	Astra	Sonnet	Gemini	Gemma	Gemma (r.)	Qwen	Qwen (r.)
Ambigüedad real detectada	86	**100**	18	80	0	29	0	14

Tabla 2. Versión final, porcentaje de acierto. Conflicto de fechas: 96 respuestas por sistema (48 casos × 2 idiomas de instrucciones). «(r.)»: con razonamiento. La instrucción solo se probó en las configuraciones sin razonamiento opcional.

Ningún sistema aplica de forma fiable por defecto la convención de la fuente cuando una fecha se cita desde otro país. El mejor, Claude Opus, lo hace en el 28 % de los casos; GPT Astra, en ninguno de los 48. Con los números, en la misma situación, seis de los ocho sistemas aciertan entre el 97 % y el 100 %, y los otros dos, el 89 % (Sonnet) y el 90 % (Qwen).

Volviendo al ejemplo del FBI en una página española:

- **Claude Opus, Claude Sonnet y Gemini Flash** responden «ambiguo: 2 de diciembre o 12 de febrero».
- **GPT Astra, Gemma y Qwen**, con y sin razonamiento, responden «2 de diciembre», siguiendo la convención de la página.

El tipo de pista importa poco. Nombrar el país ayuda algo más que nombrar una institución o un lugar, pero las diferencias son pequeñas.

5.3 Dos maneras de fallar

Casi todos los fallos con fechas en conflicto son de dos tipos:

- **«Ambiguo».** El sistema reconoce que hay dos convenciones en juego, pero no decide. Es el fallo típico de los comerciales: entre 67 y 81 de cada 96 casos.
- **«Sigue la página».** El sistema aplica la convención de la página e ignora la fuente. Los locales lo hacen con mucha más frecuencia: entre 31 y 85 de cada 96 casos, frente a entre 2 y 21 de los comerciales.

Para una redacción, el segundo fallo es más difícil de detectar. Una respuesta «ambigua» invita a comprobar el dato; una fecha equivocada dada con seguridad puede pasar inadvertida.

Razonar no lo arregla, y puede empeorarlo. Con razonamiento, Gemma sigue la página en 56 casos en lugar de 40, y Qwen, en 85 en lugar de 31. Pensar más no basta cuando el sistema da prioridad a la pista equivocada. En cambio, el razonamiento sí ayuda en otros puntos: lleva al 100 % la lectura de los metadatos en los dos modelos locales (desde el 78 % y el 40 %) y sube el acierto de Gemma con el contexto implícito del 42 % al 85 %. Parece reforzar la señal más visible, los metadatos de la página, en lugar de sopesar todas las pistas.

¿Es «ambiguo» un error? Es una respuesta prudente y defendible. Pero, con la tarea tal como está definida, la cita se atribuye a una fuente extranjera concreta, y quien conozca su convención no necesita dudar. Lo contamos como error, y presentamos siempre por separado los dos tipos de fallo.

5.4 Una instrucción ayuda… pero no a todos

Añadimos al final de las instrucciones una sola frase:

> *Si la expresión objetivo está citada o copiada de un documento o una fuente de otro país, interprétala con las convenciones de esa fuente, aunque sean distintas de las de la página o de la lengua del texto.*

Con ella, los modelos comerciales pasan del 0–28 % al 82–98 % en los conflictos de fechas. Esto indica que el conocimiento está ahí: los modelos saben que el FBI escribe primero el mes, pero por defecto no lo aplican. El resultado es compatible con un fallo de **prioridad entre pistas** más que con una simple falta de conocimiento. El coste es pequeño: en los controles, Opus y Astra pasan de 24 a 22 aciertos de 24.

Los modelos locales mejoran mucho menos: Gemma pasa del 11 % al 41 % y Qwen, del 2 % al 26 %, y Gemma sigue aplicando la convención de la página en 35 de los fallos restantes. **En un modelo local, la instrucción no basta.**

5.5 Detectar la ambigüedad o aprovechar el contexto

Las dos capacidades aparecen enfrentadas:

- **GPT Astra** detecta el 100 % de las ambigüedades reales, pero resuelve solo el 8 % de las fechas con contexto implícito. Declara ambigua incluso «Ofgem will set the new price cap on 03/08/2026», que es claramente británica.
- **Gemma y Qwen con razonamiento** hacen lo contrario: resuelven el 78–85 % del contexto implícito, pero detectan poca ambigüedad real (14–29 %).
- **Claude Opus** es el más equilibrado: 86 % de ambigüedad detectada y 64 % de contexto implícito resuelto.
- **Claude Sonnet** sin razonar queda bajo en ambas y duda incluso sin conflicto: 33 % en los controles.

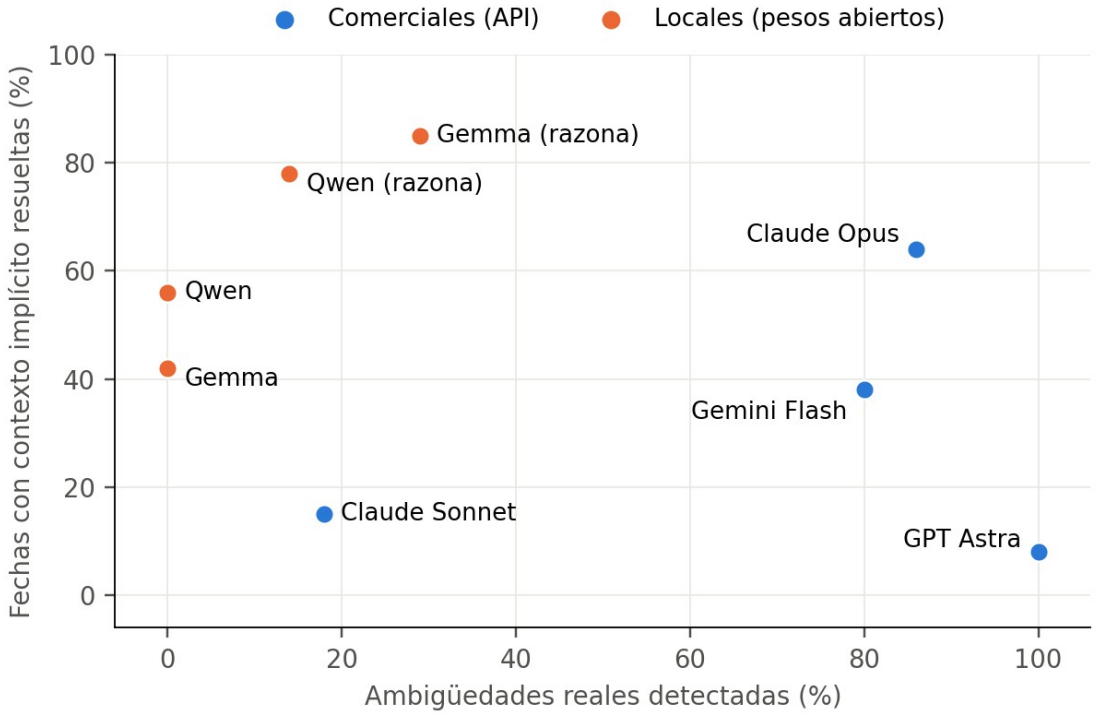


Figura 2. Ambigüedad real detectada frente a fechas con contexto implícito resueltas, por sistema (versión final).

5.6 Otros resultados

- **Idioma de las instrucciones.** Darlas en español o en inglés cambia el acierto entre 0 y 5 puntos, sin una dirección constante.
- **Normalizar frente a verbalizar.** Esperábamos encontrar sistemas que normalizaran bien y verbalizaran mal, o al revés. Casi todas las discrepancias son del tipo «normaliza mal, verbaliza bien», probablemente porque normalizar exige tipo y unidad y verbalizar no. No sacamos conclusiones de este punto.

6. El sorprendente comportamiento de las fechas

6.1 La pista del idioma explica una parte, pero no todo

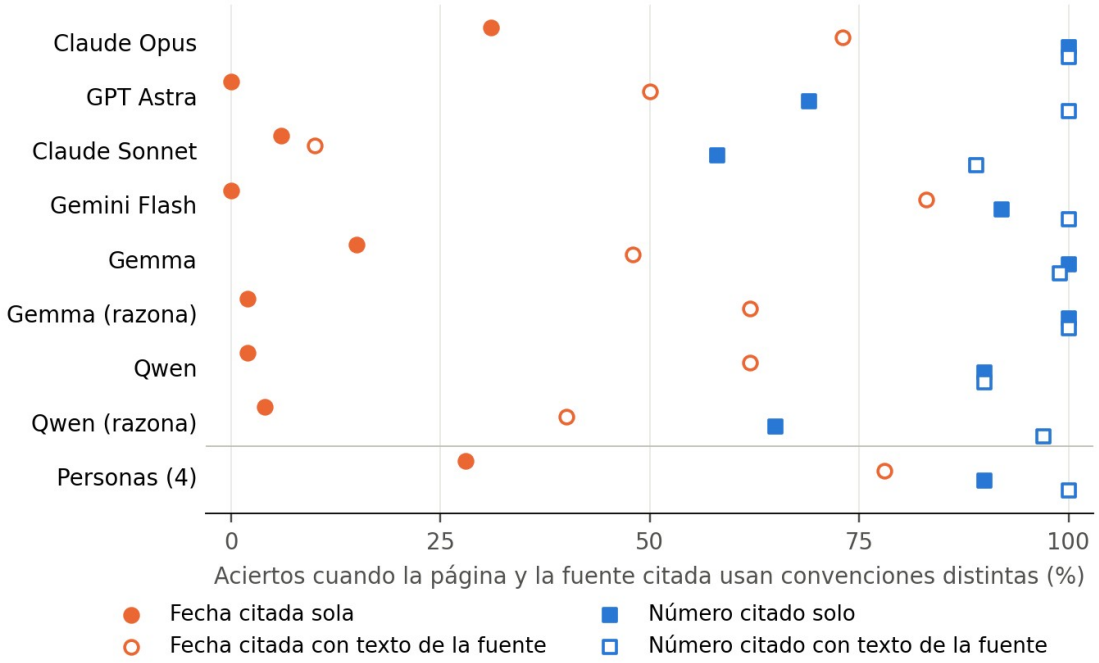


Figura 3. Acierto en los conflictos según el tipo de expresión y según la cita incluya o no texto en el idioma de la fuente. Modelos: casos español–inglés del banco de pruebas. Personas: media de cuatro participantes en los bloques correspondientes de la prueba a ciegas.

	Opus	Astra	Sonnet	Gemini	Gemma	Gemma (r.)	Qwen	Qwen (r.)
Fecha citada sola	31	0	6	0	15	2	2	4
Fecha citada con texto de la fuente	73	50	10	83	48	62	62	40
Número citado solo	100	69	58	92	100	100	90	65
Número citado con texto de la fuente	100	100	89	100	99	100	90	97

Tabla 3. Control de la pista de idioma, casos español–inglés, porcentaje de acierto.

El texto en el idioma de la fuente es una pista muy fuerte. Con él, las fechas suben del 0–31 % al 10–83 %. Sin él, los números bajan claramente en tres sistemas (Sonnet, 58 %; Qwen con razonamiento, 65 %; Astra, 69 %) y algo en Gemini (92 %).

Pero, en igualdad de condiciones, **la diferencia entre fechas y números se mantiene en los ocho sistemas**:

- con la expresión sola, fechas al 0–31 % frente a números al 58–100 %;
- con texto de la fuente, fechas al 10–83 % frente a números al 89–100 %.

La formulación precisa del hallazgo es, por tanto: **los modelos aprovechan mucho mejor la convención de la fuente cuando la cita la delata también mediante el idioma, pero,**

incluso en igualdad de condiciones, les cuesta sustancialmente más aplicarla a las fechas que a los números.

6.2 ¿Por qué las fechas?

No lo sabemos con certeza, pero hay tres explicaciones compatibles con los datos:

1. **La plausibilidad.** «1,375 personas» leído como «uno coma tres personas» no tiene sentido, así que el significado ayuda a descartar la lectura errónea. En cambio, «02/12/2025» es una fecha válida en las dos convenciones: nada en el sentido de la frase ayuda. Es la explicación más coherente con la prueba humana, en la que las personas aciertan casi todos los números y muy pocas fechas.
2. **Cómo se trocean las fechas.** Los modelos procesan el texto en fragmentos, y las fechas numéricas se fragmentan de forma irregular, lo que puede dificultar asociar la cadena con su convención.
3. **La prudencia aprendida.** Los modelos comerciales parecen entrenados para señalar la ambigüedad de las fechas, un caso de manual, y eso les lleva a dudar incluso cuando el contexto basta.

La primera explicación se puede comprobar con un experimento sencillo: números en conflicto cuyas dos lecturas sean igual de razonables, como una dosis de «2,500» mg (2.500 mg o 2,5 mg). Si con esos casos los números caen al nivel de las fechas, el problema no sería de las fechas, sino de cualquier cifra cuyo valor no se pueda comprobar por el sentido. Es uno de los próximos pasos.

7. Pruebas con noticias reales y con la Comisión Europea

7.1 Noticias publicadas en rtve.es

La muestra. Descargamos 1.200 noticias de la web de RTVE, publicadas entre el 24 de agosto y el 25 de septiembre de 2026. De ellas, 884 contienen al menos una expresión numérica, y un extractor automático encontró 5.508 expresiones candidatas.

Fenómeno	Expresiones	Noticias
Porcentajes	1.757	340
Punto de miles («80.000»)	1.546	486
Escala en palabras («36,6 millones»)	617	245
Monedas (172 extranjeras)	436	145
Cifras en frases sobre otro país	349	145
Horas	322	168
Números romanos	252	96
Fechas-evento (11-S, 7-O, 29-O…)	38	11
Fechas numéricas	14	7
«Billón» / «billones»	11	11
Coma con tres cifras («1,899»)	10	7

Tabla 4. Frecuencia de algunos fenómenos en 1.200 noticias de rtve.es (detección automática, orientativa).

En el texto publicado no hay conflictos. No aparece ninguna cifra ni fecha citada con la convención de otro país: todo llega ya adaptado. Las 11 apariciones de «billones» corresponden a *trillions* estadounidenses bien convertidos («1 billón de dólares»), y las horas extranjeras aparecen convertidas («las 15:00 hora local, 18:00 horas GMT»). Lo que sí es frecuente es la **adaptación**: 145 noticias hablan de cifras de otro país.

Esto sugiere que el punto de riesgo está **antes de publicar**: en el tratamiento y la adaptación del material de origen (teletipos, comunicados, documentos), justo donde empiezan a usarse los modelos de lenguaje. No hemos podido observar directamente ese paso.

Los fenómenos del banco de pruebas sí aparecen en la realidad:

- precios del carburante con tres decimales («1,899» €/litro), la misma trampa del separador ambiguo;
- números con punto que no son cantidades («artículo 100.2 del Reglamento Penitenciario»);
- fechas-evento («11-S», «29-O», «30-J»);
- códigos («AB-1709», «H-1B»);
- e incluso erratas de la propia noticia («5,347 gramos» por 5.347 gramos).

El valor correcto, revisado por dos personas. Seleccionamos 200 noticias que cubrían los fenómenos menos frecuentes, con hasta tres expresiones por noticia: 547 expresiones. El valor se propuso de forma automática y se corrigió a mano en el 19 % de los casos. Después lo revisaron dos personas:

- **En una primera revisión** se comprobaron las 547 viendo la propuesta. Se confirmó el 98 % y se corrigió una decena de casos: el signo de «−6,4 mm», la errata de los gramos, que «30-J» era el 30 de julio de 2026, títulos de episodio que parecían horas y expresiones recortadas.
- **En una segunda revisión, a ciegas,** otra persona escribió el valor de 152 expresiones sin ver la propuesta. Su acuerdo con el valor final es del **98,7 % en valor y escala** y del **95,4 % contando tipo y unidad**. Esa persona llegó por su cuenta a las nueve correcciones hechas tras la primera revisión.

Sistema	Correcto (valor y unidad)	Solo el valor
Claude Opus	99,5 %	99,8 %
Gemma (razona)	99,3 %	99,6 %
GPT Astra	99,1 %	99,6 %
Qwen (razona)	98,4 %	98,7 %
Gemma	98,2 %	98,9 %
Qwen	96,9 %	97,6 %
Claude Sonnet	96,2 %	96,5 %
Gemini Flash	94,1 %	99,8 %

Tabla 5. Noticias publicadas en rtve.es: 547 expresiones.

Lectura. En un texto ya adaptado a las convenciones españolas, todos los sistemas aciertan casi siempre, y los locales están al nivel de los comerciales. Gemini acierta la cifra, pero omite a menudo la unidad. Los fallos que quedan son de **tipo**, no de lectura: títulos o marcas de tiempo tomados por horas («12:00 P.M.» como título de un episodio, «2:16:13» como minuto de un vídeo), el año de las fechas-evento y la coma decimal con tres cifras.

7.2 Comunicados de la Comisión Europea

Para tener texto real con dos convenciones a la vez, usamos los comunicados de prensa de la Comisión Europea, que se publican en inglés y en español. De 1.500 referencias de 2025 y 2026, 238 tienen versión española: 165 son traducciones automáticas de la propia Comisión, avisadas como tales, y 73 son humanas. El inglés de la Comisión ya usa convenciones europeas (fechas con el día primero, miles con espacio), así que aquí no hay conflicto de fechas: el interés está en las escalas con decimales.

La traducción oficial. Entre el 92 % y el 93 % de las cifras inglesas aparecen con el mismo valor en la versión española; las demás suelen estar escritas en letra. En 353 cifras con *billion* o *trillion* no hay ni un falso amigo. Sí encontramos un error de valor en la traducción automática: «about €849.27 billion» se convirtió en «unos 849 27 000 millones EUR», un número mal formado. También hallamos una errata en un original inglés («€133,5 million»), que la traducción resolvió bien.

Tarea de adaptación. Con 557 cifras con escala o importe (258 combinan escala y decimales), dimos a los modelos la frase inglesa y les pedimos la cifra tal como iría en un texto español.

Sistema	Acierto
Traducción oficial humana	100 % (132 cifras)
Traducción oficial automática	99,3 % (420 cifras)
GPT Astra	100 %
Claude Opus	100 % (538 respuestas)
Gemma	100 %
Gemini Flash	99,8 %
Qwen	99,8 %

Tabla 6. Adaptación de cifras de comunicados de la Comisión Europea del inglés al español.

Los únicos errores caen donde predice el banco de pruebas. Gemini convirtió «€5.44 billion» en «5,44 billones de euros»: es el falso amigo *billion* → «billón», y el único caso en todos nuestros datos reales. Qwen interpretó una elipsis («€28.5 billion, of which €14.9 in loans») como 14,9 millones. **Cuando la convención de destino está clara, la adaptación de escalas alcanza en nuestros datos una precisión prácticamente perfecta**, al nivel de la traducción oficial o por encima.

8. Comparación con personas

8.1 Diseño de la prueba

Elegimos 65 casos del banco de pruebas en bloques: 16 fechas en conflicto citadas solas, 8 con texto de la fuente, 10 números en conflicto, 5 controles, 10 fechas con contexto implícito, 6 de metadatos y 10 realmente ambiguos. Cuatro periodistas y documentalistas de RTVE los respondieron por separado en una hoja de cálculo, sin ver la solución y sin volver atrás. Cada pregunta mostraba el país y la dirección de la página, la frase y las lecturas posibles, más la opción «No se puede saber», y pedía el grado de seguridad.

8.2 Resultados

Bloque (casos)	P1	P2	P3	P4	Mejor sistema
Fecha en conflicto, sola (16)	3	8	4	3	4 (Opus, Gemini)
Fecha en conflicto, con texto (8)	6	8	8	3	8 (Opus, Gemini)
Número en conflicto (10)	10	9	10	9	10 (cuatro sistemas)
Controles (5)	5	5	3	1	5 (cinco sistemas)
Contexto implícito (10)	0	5	5	4	9 (Qwen razonando)
Ambigüedad real (10)	6	3	3	6	10 (Astra, Gemini)

Tabla 7. Aciertos por persona y del mejor sistema en los mismos casos.

En esta prueba pequeña, las personas resolvieron las fechas en conflicto más o menos igual que los mejores modelos, es decir, mal: 18 de 64 respuestas correctas (28 %; intervalo al 95 % entre el 9 % y el 50 %) con fechas citadas solas, frente a 38 de 40 (95 %; entre el 82 % y el 100 %) con números. Los intervalos son anchos porque solo hay cuatro participantes, pero no se solapan: **la diferencia entre fechas y números se reproduce en las personas.**

8.3 Tres estrategias de error

Recuadro. «Inglés = formato estadounidense»

Dos de las cuatro personas leyeron mes/día en casi todas las fechas escritas en inglés, y día/mes en las escritas en español. Con esa regla aciertan todo lo que viene de Estados Unidos (la Reserva Federal, la Casa Blanca, el Pentágono) y fallan todo lo que viene del Reino Unido (Leeds, Manchester, el NHS), porque los británicos escriben la fecha como los españoles. Lo que decide el formato de una fecha no es el idioma, sino el país.

1. **Seguir el idioma de la frase** (P2, P3 y en gran parte P4). En una frase en inglés leen casi siempre mes/día («inglés = formato estadounidense»), y en español, día/mes. Aciertan todo lo estadounidense (Reserva Federal, Casa Blanca, Pentágono, Houston) y fallan todo lo británico (Leeds, Manchester, el NHS, el Tesoro británico).
2. **Aplicar la convención propia** (P1). Lee día/mes siempre y, sin página, responde «No se puede saber» en todas las fechas con contexto implícito.

3. **Dudar cuando la página y la fuente no coinciden** (P4). Responde «No se puede saber» en 14 casos de conflicto o control, con comentarios como «¿traducida?» o «puede haber mala conversión». Es exactamente el «ambiguo» de los modelos comerciales.

Nadie aplicó la convención británica en una frase en inglés: las únicas fechas británicas acertadas son de P1, que lee día/mes en todas partes. Además, las personas señalaron poco la ambigüedad real (entre 3 y 6 de 10) y declararon seguridad alta en casi todo (entre 57 y 64 de 65 respuestas en tres de ellas). Coincidieron entre sí en solo 27–53 de las 65 preguntas.

Los mejores modelos superan a las personas en el contexto implícito y en la detección de ambigüedad. Y la confusión «inglés = formato estadounidense» es justo el riesgo de quien adapta cifras extranjeras en una redacción española.

9. Implicaciones para los medios

1. **Las fechas numéricas citadas de fuentes extranjeras son un punto ciego** cuando la propia cita no aporta una pista lingüística clara sobre su procedencia. Cuando la fecha va sola, ningún sistema supera el 31 %.
2. **El tipo de fallo depende del modelo.** Los comerciales avisan con «ambiguo»; los locales se equivocan con seguridad, y razonar no lo corrige.
3. **Con los modelos comerciales probados, una instrucción explícita resuelve en gran parte el problema** (82–98 %). Con los locales hace falta un control externo.
4. **La revisión humana, por sí sola, puede compartir el mismo punto ciego**, sobre todo con las fechas británicas. Nuestra prueba es pequeña, pero el patrón es claro.
5. **Leer cifras de un texto ya adaptado al español es fiable** con cualquiera de los sistemas probados (94–99,5 %). Los errores que quedan son de tipo: títulos, códigos o marcas de tiempo tomados por cantidades u horas.
6. **Adaptar escalas del inglés al español también es fiable** cuando el destino está claro. El falso amigo *billion*/«billón» es hoy raro, aunque no imposible.
7. **Para detectar la ambigüedad real**, algunos modelos comerciales son claramente mejores (Astra, Opus, Gemini). **Para aprovechar el contexto**, el tipo de modelo importa menos: Opus es el más equilibrado, y los locales con razonamiento resuelven muy bien las fechas con contexto implícito.
8. **El riesgo está antes de publicar**, en el tratamiento del material de origen. Es ahí donde conviene poner los controles.

10. Aplicación en RTVE

10.1 Dónde aplicar las conclusiones

Proceso	Riesgo principal	Medida
Resumen o reescritura de teletipos y comunicados	Fechas numéricas citadas de fuentes estadounidenses	Instrucción de convención en el prompt y alerta automática

Proceso	Riesgo principal	Medida
Traducción y adaptación inglés–español	Escalas con decimales, falsos amigos, elipsis	Comprobación automática del valor de cada cifra
Subtitulado y accesibilidad	Fechas y horas extranjeras	Escribir el mes en letra
Locución automática y síntesis de voz	«2.30», «1,5 M€», «11-M», romanos	Normalizar antes de locutar, con revisión de tipo
Documentación y archivo	Fechas ambiguas al indexar	Guardar las fechas en formato AAAA-MM-DD
Verificación de datos	Cifras de documentos extranjeros	Comprobar siempre la convención del documento original

10.2 Recomendaciones concretas

1. **Tratar como alerta cualquier fecha numérica con día y mes menores o iguales que 12** que proceda de una fuente extranjera, sobre todo estadounidense. Es fácil de detectar automáticamente y cubre exactamente el caso en el que fallan modelos y personas.
2. **Escribir el mes en letra** en los textos publicados («12 de febrero de 2025», no «12/02/2025»). Elimina la ambigüedad en origen.
3. **Usar el formato AAAA-MM-DD en los datos internos** (metadatos, bases de datos, archivo).
4. **Si se usa un modelo comercial, incluir siempre en el prompt la instrucción sobre la convención de la fuente.** Es una sola frase y en nuestras pruebas cambia el acierto del 0–28 % al 82–98 %.
5. **Si se usa un modelo local** (por confidencialidad o por coste), **no confiar en la instrucción**: añadir la alerta del punto 1 y una revisión específica de esas fechas.
6. **No desactivar el «ambiguo».** Una respuesta ambigua es una señal útil para revisar, no un fallo que haya que eliminar.
7. **Formar a la redacción en un dato sencillo:** en las tres variantes estudiadas, España y Reino Unido escriben primero el día; Estados Unidos, primero el mes. En la prueba humana, este fue el error más repetido.
8. **Evaluar antes de adoptar.** El banco de pruebas de este estudio permite comparar cualquier modelo nuevo en un día y con un coste mínimo. Conviene pasarlo antes de incorporar un sistema a un flujo de trabajo.

10.3 Próximos pasos

- **Material de origen.** Medir la frecuencia real del conflicto en teletipos y documentos de origen, donde el estudio indica que está el riesgo.
- **Números igual de plausibles.** Comprobar si la plausibilidad explica la diferencia entre fechas y números (sección 6.2).
- **Repeticiones.** Repetir los casos de conflicto en los modelos con variabilidad para afinar los intervalos.

- **Español de América.** Ampliar el estudio a variantes donde conviven varias convenciones.
- **Síntesis de voz.** Llevar la tarea de verbalizar a un sistema de locución real.

11. Limitaciones

- **El resultado central es sintético.** Las noticias reales confirman que los fenómenos existen y que las convenciones se conocen, pero no miden el conflicto, que se produce en el material de origen.
- **La prueba humana es pequeña:** cuatro personas, todas hispanohablantes, sin nativos británicos ni estadounidenses. Describe a quien adapta cifras en una redacción española, no a un lector nativo.
- **Una sola ejecución por caso** e idioma de instrucciones. Los modelos con variabilidad necesitarían repeticiones.
- **Tres variantes y un solo prompt base**, aunque en dos idiomas.
- **Los modelos comerciales cambian con el tiempo.** Hemos registrado la versión y la fecha de cada ejecución.
- **En la tarea de la Comisión** no se evaluaron Claude Sonnet ni las variantes con razonamiento, y Claude Opus tiene 538 de 557 respuestas.

Conclusiones

Los modelos de lenguaje actuales saben cómo se escriben los números y las fechas en español y en inglés. Lo que no hacen bien es decidir qué convención aplicar cuando la página y la fuente citada no coinciden, y con las fechas fallan de forma sustancial todos los sistemas probados. Fallan de dos maneras: los comerciales dudan y los locales se equivocan con seguridad. Una instrucción sencilla corrige en gran parte a los comerciales, pero no a los locales.

En nuestra prueba, las personas que adaptan cifras en una redacción reprodujeron el patrón: fallaron en los mismos casos y por razones parecidas. En las noticias publicadas no hay conflictos sin resolver, lo que sugiere que el riesgo está antes de publicar, donde los modelos de lenguaje se usan cada vez más.

La lección general va más allá de las fechas: **conocer una convención no es lo mismo que saber elegirla**. Saber casi a la perfección cómo se escribe una expresión no garantiza recuperar lo que significa en su contexto. Para un medio, eso se traduce en una recomendación concreta: las fechas numéricas citadas de fuentes extranjeras necesitan una atención específica, y el tipo de modelo determina qué salvaguarda hace falta.

Referencias

- Amrhein, C. y Sennrich, R. (2022). Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET. *Proceedings of AACL-IJCNLP*.

- Bhatia, G., Peyrard, M. y Zhao, W. (2025). Date Fragments: A Hidden Bottleneck of Tokenization for Temporal Reasoning. *Proceedings of EMNLP*.
- Ebden, P. y Sproat, R. (2015). The Kestrel TTS text normalization system. *Natural Language Engineering*, 21(3).
- Gupta, D. y Latorre, J. (2024). Positional Description for Numerical Normalization. *Proceedings of Interspeech*.
- Kaye, P. (2026). Analysis of Numerical Localisation in LLM Translations. arXiv:2608.05232.
- Li, H., Chen, X., Xu, Z., Li, D., Hu, N., Teng, F., Li, Y., Qiu, L., Zhang, C. J., Qing, L. y Chen, L. (2025). Exposing Numeracy Gaps: A Benchmark to Evaluate Fundamental Numerical Abilities in Large Language Models. *Findings of ACL 2025*.
- Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. y Choi, Y. (2023). We're Afraid Language Models Aren't Modeling Ambiguity. *Proceedings of EMNLP*.
- Min, S., Michael, J., Hajishirzi, H. y Zettlemoyer, L. (2020). AmbigQA: Answering Ambiguous Open-domain Questions. *Proceedings of EMNLP*.
- Real Academia Española y Asociación de Academias de la Lengua Española (2010). *Ortografía de la lengua española*.
- Reddy, V., Schmidt, C. W., Ebner, S., Wiemerslage, A., Pinter, Y. y Tanner, C. (2026). 1,729 vs. 1729: The Effect of Scripts and Formats on LLM Numeracy. *Findings of ACL*.
- Singh, A. K. y Strouse, D. J. (2024). Tokenization counts: the impact of tokenization on arithmetic in frontier LLMs. arXiv:2402.14903.
- Tang, W., Yu, J., Li, Y., Zhao, Y., Zhang, W., Feng, W., Zhang, M. y Yang, H. (2025). Investigating Numerical Translation with Large Language Models. *Proceedings of ICASSP*.
- Unicode Consortium. *UTS #35: Unicode Locale Data Markup Language, Part 3: Numbers*.
- Wallace, E., Wang, Y., Li, S., Singh, S. y Gardner, M. (2019). Do NLP Models Know Numbers? Probing Numeracy in Embeddings. *Proceedings of EMNLP-IJCNLP*.
- Wang, J., Xu, C., Guzmán, F., El-Kishky, A., Rubinstein, B. I. P. y Cohn, T. (2021). As Easy as 1, 2, 3: Behavioural Testing of NMT Systems for Numerical Translation. *Findings of ACL-IJCNLP*.
- Wong, M., Alshehri, A., Kao, S. y He, H. (2025). PolyNorm: Few-Shot LLM-Based Text Normalization for Text-to-Speech. *Proceedings of EMNLP 2025: Industry Track*, 77–85.
- Zhang, Y., Bartley, T. M., Graterol-Fuenmayor, M., Lavrukhin, V., Bakhturina, E. y Ginsburg, B. (2023). A Chat About Boring Problems: Studying GPT-based text normalization. arXiv:2309.13426.