FIELD NOTE 11

The Invisible Competitor: How AI Omissions Shape Medicine Comparisons

An AI answer can be factually correct and still present an incomplete comparison. In a study of 287 answers across four AI models, AlphaCitation found that omissions were heavily concentrated on one medicine in obesity treatment comparisons, yet largely absent in Alzheimer's treatment comparisons on the same measure. The findings reveal a dimension of AI answer quality that conventional factual accuracy monitoring can miss: what the answer leaves out.

Ozue C. Ezeife

17 November 2026

Obesity and Alzheimer's disease · United States · 4 engines · 10 min read

When AI models compare competing medicines, what their answers omit can matter as much as what they get wrong.

Most discussions of AI reliability focus on hallucinations, such as invented clinical evidence, incorrect indications or inaccurate safety information. However, a more subtle vulnerability exists. An AI assistant can describe two medicines without contradicting either label, yet omit relevant prescribing information for one while detailing it for the other.

This is **comparative asymmetry**: an uneven representation of the two medicines' prescribing information.

For pharmaceutical executives, Medical Affairs and brand teams, this raises a critical question beyond conventional accuracy monitoring:

Do AI answers give both medicines a complete and appropriately balanced account of their approved profiles?

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

The Zepbound and Wegovy Gap

On 9 November 2026, AlphaCitation tested four AI models (GPT-5.5, Claude Sonnet 5.5, Gemini 3.8 Flash and Sonar Pro) using structured questions comparing two pairs of medicines:

- **Obesity:** Zepbound (tirzepatide) and Wegovy (semaglutide).
- **Early Alzheimer's disease:** Leqembi (lecanemab) and Kisunla (donanemab).

The study collected 287 answers through developer APIs with web search enabled, from the United Arab Emirates. All questions specified the United States, and answers were evaluated against the medicines' US prescribing information.

The obesity comparisons revealed a concentrated pattern of omission.

One-sidedness was defined precisely: an answer discussed a category of prescribing information for one medicine but omitted that category for the other, despite both labels containing relevant information.

METRIC	FIRST CODER	SECOND CODER
Total obesity answers	144	144
One-sided answers	27 (18.8%)	29 (20.1%)
One-sided answers involving testing/monitoring	24	27
Answers omitting Wegovy's testing/monitoring	22	24

The reported figures are descriptive. The testing and monitoring row counts omissions affecting either medicine: for the first coder 22 left out Wegovy's and 2 Zepbound's; for the second, 24 and 3.

A text search of those 22 answers, made after the coding was tallied and not a pre-specified measure, found that all 22 gave Zepbound's advice on oral contraceptives and none mentioned heart rate.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

The contrast is clinically specific.

Zepbound's US label tells prescribers to advise patients using oral hormonal contraceptives to switch to a non-oral method, or add a barrier method, for four weeks after starting treatment and for four weeks after each dose escalation.

Wegovy's label instructs healthcare professionals to monitor heart rate at regular intervals, consistent with usual clinical practice.

While these represent different clinical obligations, both belong to the broader testing and monitoring domain evaluated in the study.

Each of the first coder's 22, 12 from GPT-5.5, 5 from Claude Sonnet 5.5 and 5 from Gemini 3.8 Flash, gave a precaution for Zepbound and no testing or monitoring information for Wegovy.

One Claude Sonnet 5.5 response illustrates the distinction. It discussed the head-to-head trial, the approved indications, the thyroid boxed warning and Zepbound's contraceptive advice. Neither coder found a statement in it that contradicts either label, and neither found any testing or monitoring information for Wegovy.

No statement contradicted either label. The comparison was incomplete.

Brand Preference and Missing Burdens

The study also examined whether models recommended one medicine overall.

Across all 144 obesity answers, neither coder classified any answer as preferring Wegovy overall. Of the 22 answers omitting Wegovy's testing and monitoring, the first coder read 4 as favouring Zepbound and 18 as making the choice depend on the patient, the second coder 11 and 11 (a count of the stored codes made after the tally).

However, the coders differed substantially on how often an answer expressed a definitive preference rather than recommending treatment based on individual patient circumstances.

The first coder classified 18 of 144 answers as favouring Zepbound overall; the second classified 61. Agreement on preference was moderate (Cohen's $\kappa = 0.43$, 287 answers across both pairs), against agreement on one-sidedness of $\kappa = 0.81$.

These findings must not be conflated.

Omitting Wegovy's monitoring requirements could make its treatment profile appear less burdensome. The study does not establish that the omissions caused, contributed to or otherwise influenced the observed treatment preferences.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

It establishes more reliably the **direction and concentration of the missing information**.

Detecting an incomplete comparison is not equivalent to proving competitive bias or demonstrating an impact on prescribing decisions.

Context Shifts in Generic Naming

Answers differed in the evidence they brought in according to how the question named the medicines.

Tirzepatide is also sold as Mounjaro, and semaglutide as Ozempic and Rybelsus; all three are approved in the United States for type 2 diabetes.

When prompts used generic names for obesity comparisons, answers more often introduced diabetes-related evidence.

- **Generic-name prompts:** 37 of 72 answers (51.4%) introduced a diabetes product, trial, dose or outcome.
- **Brand-name prompts:** Nine of 72 answers (12.5%) introduced similar material on the first coder's reading. The second coder identified 11.

Not every reference to diabetes evidence is inappropriate. Some evidence may be relevant to understanding a medicine's pharmacology, clinical development or therapeutic profile.

Nevertheless, the disparity is substantial.

Named by molecule, the obesity comparison brought in diabetes evidence in half the answers; named by brand, in about one in eight.

For pharmaceutical teams, brand-name questions alone did not show what generic-name questions drew into the same comparison.

That evidence concerned the same molecules' diabetes products, trials and doses, which sit outside the two obesity labels being compared.

The finding does not establish that such evidence is incorrect or inapplicable, or that the naming caused the difference: the design is descriptive.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

The Alzheimer’s Comparator: A Different Pattern

The study applied the same methodology to Leqembi and Kisunla, two medicines for early Alzheimer’s disease.

Both carry boxed warnings for amyloid-related imaging abnormalities (ARIA), including brain swelling and bleeding detectable on MRI. Their labels also differ in administration, treatment schedules, monitoring requirements and other clinical considerations.

Of 143 collected answers, the first coder identified just one one-sided comparison (0.7%).

The second coder also identified one, although in a different repetition of the same question. The two coders therefore did not agree on which individual answer was one-sided.

One planned answer was not collected. Even if that answer had been one-sided, the count would have increased to at most two of 144.

The same four models, on the same day and under the same coding, gave one-sided obesity comparisons in 27 of 144 answers on the first coder’s reading (29 on the second’s) and one-sided Alzheimer’s comparisons in 1 of 143 on both; Sonar Pro gave none in either pair, on both coders’ readings.

This does not mean the Alzheimer’s comparisons were error-free. The coders also found statements contradicting the labels: the first coder in 35 answers for Leqembi, 30 of them on administration, and 24 for Kisunla, 16 on administration; the second coder in 12 and 10, 5 and 5 on administration.

The pattern of one-sided omission was therefore not a feature of the models or the method wherever they were applied: it appeared in one of the two pairs tested.

The asymmetry was concentrated in a specific competitive comparison, not consistently reproduced wherever the study looked.

The study does not establish the underlying cause of this difference.

Study Design and Limitations

AlphaCitation used 24 structured questions, each repeated three times at least two hours apart across four models. The questions varied audience perspective, medicine naming and presentation order.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

Two AI coders from different developers coded every response independently. Neither was told the model or the repeat; the question each coded showed the asker and the order, and an answer's style or citations can hint at the model. They evaluated the responses against seven predefined domains of US prescribing information:

- Indication and population
- Primary efficacy
- Boxed warning
- Contraindications
- Testing and monitoring
- Administration
- Other safety

Both coders passed a qualification test on answers with known defects before collection: each found every planted omission and preference, and neither flagged a balanced answer.

They subsequently agreed on whether an answer was one-sided in 277 of 287 cases (Cohen's κ = 0.81). Agreement was weaker for treatment preference (κ 0.43) and cross-trial comparisons (κ 0.32). Taken one part of the label at a time, agreement on administration and other safety was 0.00, because the coders never flagged the same answer there; the agreement condition was met on the answer-level reading, which no record made before collection fixes.

These results represent agreement between two AI-based coders, not independent clinical adjudication.

Protocol deviations

The study recorded several departures from its pre-specified protocol.

Seven Gemini 3.8 Flash calls, for six planned answers, returned no answer. Five of those answers were collected on a second attempt; one failed twice and was never collected.

The protocol required collection to stop following a collection fault and specified that rates should not be published if the condition failed. Collection nevertheless continued under a founder-authorised decision recorded before the answers were reviewed.

The second coder also required an additional pass to complete 67 answers. This was recorded before the codes were analysed. Agreement on one-sidedness was lower for these 67 answers (κ = 0.63) than for the remaining 220 (κ = 0.87).

On 11 November 2026, after the results had been seen, a further recorded founder decision (deviation 8) authorised publication of the descriptive rates despite the failed collection condition.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

The missing answer could increase the number of one-sided Leqembi/Kisunla comparisons by at most one and cannot change the Zepbound/Wegovy counts. However, that bound does not resolve the broader limitations introduced by the protocol deviations.

Accordingly, **the results should be interpreted as exploratory, descriptive findings rather than fully protocol-conforming confirmatory evidence**.

The study measures what the models produced, not why. It cannot determine whether omissions arose from training data, web retrieval, model behaviour, safety tuning or other mechanisms.

It does not measure actual consumer-app exposure, clinical decision-making or patient outcomes.

The classification of one-sidedness measures coverage of prescribing-information domains, not the clinical equivalence or materiality of every omitted detail.

The study also does not assess whether the answers satisfy regulatory fair-balance requirements applicable to manufacturer promotional communications. These were third-party AI-generated responses, and the applicable regulatory obligations are distinct.

Every pre-specified measure, both coders' results and the deviations that touch them are in the accompanying annex, ***S7, The Invisible Competitor: every pre-specified measure*** (https://alphacitation.com/insights/the-invisible-competitor/evidence).

Strategic Imperatives for Pharmaceutical Teams

The findings highlight a dimension of AI answer quality that conventional factual accuracy monitoring can miss.

Traditional audits ask whether a model made an incorrect claim. Comparative evidence assurance must ask an additional question:

What relevant information was represented for one medicine but omitted for its competitor?

This creates three priorities for pharmaceutical leadership.

1. Medical and Regulatory: Look beyond factual contradictions

An answer may contain no explicitly false statement yet still provide an incomplete account of the relevant prescribing information.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

Medical and Regulatory teams need to distinguish factual contradictions from omissions, assess their potential clinical significance and verify findings against current, jurisdiction-specific authoritative sources.

Not every omission is material. The critical question is whether the information left out changes the completeness or clinical interpretation of the comparison.

2. Brand and Competitive Intelligence: Monitor the full comparison

Brand representation cannot be assessed solely by counting mentions, citations or favourable recommendations.

Teams need to understand how their medicines are described relative to competitors, including which benefits, limitations and practical treatment requirements are represented or omitted.

The generic-name findings reinforce the importance of evaluating different audience types, question formulations and naming conventions.

A brand may be accurately mentioned while relevant aspects of its approved profile remain underrepresented.

3. Digital and Medical Information: Investigate recurring evidence gaps

Repeated omissions provide a basis for further investigation.

Digital and Medical Information teams can examine whether authoritative product information is accessible, clearly communicated and appropriately represented in sources that AI systems may retrieve.

Where gaps are identified, teams can evaluate evidence-based communication improvements and test whether subsequent AI responses change.

The study does not establish that additional content or digital optimisation will correct the omissions. Such interventions require their own prospective evaluation.

The objective is not to make AI models recommend a particular medicine. Treatment decisions must remain grounded in clinical evidence and individual patient circumstances.

It is to understand whether AI answers present the relevant information accurately and with sufficient comparative completeness.

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

The Missing Information Matters

The Zepbound and Wegovy finding is not that the answers ignored Wegovy, invented its safety profile or consistently recommended an inappropriate treatment.

It is more precise.

Across repeated comparisons, AI models sometimes discussed prescribing-information requirements for one medicine while omitting the corresponding category for its competitor.

In the obesity comparisons, 22 of the first coder's 27 one-sided answers left out Wegovy's testing and monitoring and 5 left out a part of Zepbound's label; for the second coder, 24 and 7 of 29, two answers leaving out a part of each.

The Alzheimer's comparison showed that the same pattern did not occur to a comparable extent across every medicine pair tested.

This is why factual accuracy alone is an incomplete measure of AI answer quality.

An AI answer can be correct in what it says and materially incomplete in what it leaves out.

For pharmaceutical companies seeking to understand how their medicines are represented in AI-generated answers, both dimensions deserve scrutiny.

About the research

This study forms part of AlphaCitation's research into AI evidence assurance and competitive intelligence for pharmaceutical brands. The analysis examines how AI systems represent medicines, competitors and prescribing information across structured questions, with findings evaluated against authoritative evidence.

Answers collected: 9 November 2026 | United States | Four AI models | 287 collected answers.

Full methodology and results: ***S7, The Invisible Competitor: every pre-specified measure*** *(https://alphacitation.com/insights/the-invisible-competitor/evidence).*

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

CITE THIS FIELD NOTE

Ezeife OC. The Invisible Competitor: How AI Omissions Shape Medicine Comparisons. AlphaCitation Field Note 11, 17 November 2026. https://alphacitation.com/insights/the-invisible-competitor

© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0