Does New Medical Evidence Change AI Treatment Answers?

A Longitudinal Study of Evidence Citation, Medicine Naming and Selected Treatment Framing Around EASD 2026

Ozue C. Ezeife · AlphaCitation Research · October 2026 · Published by AlphaCitation Research, the research programme of AlphaCitation, a digital product of Akunudo LLC-FZ

Version 1.0 · 12 October 2026 · AlphaCitation Research Series · Paper No. 01

This paper has not undergone independent external peer review. Its protocol, technical supplement, limitations and reproducibility record are published with it, and scholarly and practitioner feedback is invited.

Abstract

Background. When AI-generated treatment answers change after a scientific event, the change is usually read as the event's effect. That reading is rarely tested: answers vary from day to day, and a citation does not show that a source shaped an answer.

Design. Longitudinal surveillance around the 62nd Annual Meeting of the European Association for the Study of Diabetes (EASD 2026; Milan, 28 September – 2 October 2026). Four developer-API model endpoints with web search (gpt-4o-mini, Claude Haiku 4.5, Gemini 2.5 Flash, Perplexity sonar) answered maintained batteries of market-framed question slots about semaglutide (Wegovy) in the United Arab Emirates (UAE) and empagliflozin (Jardiance) in Singapore, both registered arms, and teplizumab (Tzield) in Saudi Arabia, an arm added after registration. Observation ran from 22 August to 11 October 2026 (64 of 67 cycles admitted). Primary endpoints were congress penetration and latency (E1) and naming of the medicine (brand or molecule, including mentions that do not recommend its use) against a pre-congress band corrected for an instrument change (E2), with a pre-specified attribution rule. The protocol was committed on 2 October 2026, before collection of after-window cycles, after the registered readouts had been presented and a during-window UAE cycle had been inspected. Some computational interpretations were specified after preliminary after-window counts were visible; additional robustness and source-tracing analyses were undertaken after the results. These timings limit the confirmatory interpretation of the movement verdicts. Post-registration analyses (D24), including registered source analyses first reported late, are labelled throughout.

Findings. No qualifying citation or result text was verified by the implemented inspection procedure for any registered readout in its own series (not detected within follow-up, day 10.0 to 12.1). In the registered arms, six of 2,408 after-window answers qualified through cited congress-derived sources, all sponsor releases or media. A subsequent check identified one additional UAE Gemini answer reporting the OCTANE result without a qualifying source, giving seven of 2,408 answers under the endpoint's either-route definition; in the post-registration Saudi arm, 5 of 1,789 qualified. One registered movement was declared: Claude named semaglutide or Wegovy in 256 of 378 scored UAE answers (67.725%), 0.395 percentage points beyond its band; a post hoc cycle-level analysis estimates an increase of 16.8 percentage points (95% interval 7.2 to 26.4) under independent-cycle and bridge assumptions. No Claude UAE answer cited a congress-derived source. A local commercial web page displaying an

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

August 2026 date first appeared in Claude's resolved citations on 29 September; after the congress, Claude's answers citing it named the medicine in 143 of 152 cases (94.1%), against 113 of 226 (50.0%) in answers that did not, and from 29 September 152 of 164 citing answers repeated its sentences. Perplexity's naming stepped on 7 October alongside a change in its whole source mix, yet its pooled rate stayed 0.074 points inside its band. Both UAE movement verdicts change under alternative readings of the band rule. For the UAE engines, the attribution test could not detect a congress link at the observed counts of zero to two governed qualifying answers per engine.

Interpretation. The one declared movement sat in answers citing a local page, not congress evidence. This is a descriptive, post hoc association, not a causal estimate. A naming rate read alone would have suggested a congress effect.

Limitations. Between 9.9% and 31.5% of new-URL citations per engine after the congress could not be read, and the source and text screens have incomplete and unquantified recall; coding was model-assisted by one model family, checked by blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader, in which the author agreed with 17/17 codes in the ordinary sampled strata and 30/33 codes in deliberately risk-enriched strata (quality-check results for the selected items, not estimates of overall coding accuracy or independent clinical validation). The efficacy endpoint failed audit and is quarantined; retrieval queries, consumer applications and downstream behaviour were not observed.

Suggested citation

> Ezeife OC. *Does New Medical Evidence Change AI Treatment Answers? A Longitudinal Study of Evidence Citation, Medicine Naming and Selected Treatment Framing Around EASD 2026.* AlphaCitation Research Series, Paper No. 01, version 1.0. Dubai: AlphaCitation Research; 12 October 2026. https://alphacitation.com/blog/does-new-medical-evidence-change-ai-treatment-answers

Licence

AlphaCitation is a digital product of Akunudo LLC-FZ. Akunudo LLC-FZ is the copyright holder and licensor of the original material identified in this publication. That material is licensed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0), subject to the third-party exclusions stated below.

Except where separately identified, the original text, figures and compiled research outputs for which Akunudo LLC-FZ holds the relevant rights are licensed under CC BY 4.0. Quoted third-party text, third-party data fields, API/provider-restricted material and third-party marks are excluded from that grant unless their own licence expressly permits the stated reuse.

The original figures (Figures 1–13) are included in the grant. The evidence package carries an asset-by-asset licence manifest (LICENCES.md). The active question battery, the proprietary measurement instruments, production software, confidential methodology implementations and unreleased datasets are proprietary and not licensed.

Executive synthesis

Why the question matters. Medical congresses are where much trial evidence first becomes public, and AI engines are now one of the channels through which it could reach people asking which treatment to use. Organisations that watch these engines want to know whether a congress "moved" the answers. Two facts make that hard to judge. The answers move between days: the same question asked of the same engine on consecutive days can return different medicines. Before the congress, observed naming rates varied across cycles; this study did not determine whether contemporaneous evidence or service changes contributed to that variation. And a citation shows only that a page was displayed beside an answer, not that its content was represented or that it changed what the answer recommended.

What was investigated. Whether evidence presented at EASD 2026 reached what four AI engine APIs told people about three diabetes and obesity medicines, how quickly, through which sources, and whether it changed how the medicines were named, positioned or described. Maintained batteries of question slots were run before and after the congress; naming was judged against pre-congress bands corrected for a change in the measuring instrument; every newly cited page was subjected to an automated fetch and term screen, with candidate pages classified from retrieved content; and an attribution rule was fixed in advance.

What was found.

1. *The largest movement sat in answers citing a local web page, not the congress.* The share of Claude's UAE answers naming semaglutide or Wegovy rose about 17 percentage points. None of its UAE answers cited congress material. Instead, a local commercial page displaying an August 2026 date first appeared in its resolved citations on 29 September. After the congress, answers citing that page named the medicine in 94.1% of cases, and answers not citing it in 50.0%, close to the pre-congress level; the citing answers repeated the page's wording. This is an association observed after the fact, not proof of cause, but it is the same pattern AlphaCitation documented before the congress, when a clinic's page entered another engine's citations [A1].
2. *The registered trial readouts were not verified in the answers.* For none of the seven was a qualifying citation or result text verified by the implemented inspection procedure in its own series by day 10 to 12. In the registered arms, six of 2,408 after-window answers qualified through cited congress-derived sources, all sponsor releases or media. A subsequent check identified one additional UAE Gemini answer reporting the OCTANE result without a qualifying source, giving seven of 2,408 answers under the endpoint's either-route definition. The frozen and corrected tallies are shown separately (Section 5.2).
3. *The verdicts depend on the reading of the band rule.* The pooled movement verdict is sensitive to the band-rule interpretation, as a check added after the results shows. A post hoc cycle-level analysis estimates a Claude increase of 16.8 percentage points (95% interval 7.2 to 26.4) under independent-cycle and bridge assumptions; it supports a descriptive increase within this design (D24). The evidence linking any change to the congress is absent, and the registered attribution test could not have found such a link at the observed volumes.

What remains uncertain. Up to a third of new-URL citations could not be read. Coding was performed by models of one family. In blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader, the author agreed with 17/17 codes in the ordinary sampled strata and 30/33 codes in deliberately risk-enriched strata. These are quality-check results for the selected items, not estimates of overall coding accuracy or independent clinical validation. The local-page finding comes from analyses added after the results. Nothing in the study observes what users chose or clinicians prescribed.

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

Why decision-makers should care. A brand-naming rate, the most visible metric, moved in the week after a major congress, and the source record located the movement in answers citing a local page. Any team that reports AI "share" gains after an evidence event faces the same inferential gap. The paper proposes a six-stage framework (Figure 2) and shows, in the case of Claude, how measuring the links between stages turns a misleading headline into a traceable account.

---

1. Scientific and industry background

1.1 How congress evidence enters circulation

The 62nd EASD Annual Meeting ran from 28 September to 2 October 2026 in Milan, in hybrid format [1], with late-breaking abstracts and trial sessions at stated times [2]. Major trials were published in journals around their presentation and announced in sponsor releases [3–5]. EASD co-authors, with the American Diabetes Association (ADA), consensus reports that frame clinical practice [6,7].

Congress evidence is an early and partial stream. Five scientific meetings generated 252 news stories about 147 abstracts, a quarter of which remained unpublished three years later [8]; 37.3% of meeting abstracts are later published in full [9]. Intermediaries shape what is carried forward: journal press releases rarely state limitations [10], exaggeration in health news is strongly associated with exaggeration in the press release [11], and spin appeared in 47% of randomised-trial press releases [12]. This study did not assess the releases it observed for spin or accuracy.

"New evidence" is not uniform across a congress. REIMAGINE 1's primary results were published and presented at the ADA meeting on 7 June 2026, about 16 weeks before EASD, where it appeared as two secondary late-breaking abstracts [13,14]. Three ACHIEVE trials were published between February and August 2026 [15–17]. TRIUMPH-2 and ACHIEVE-4 had their full data first presented and published at or around EASD (Crossref record dates 29 and 30 September) [3,4], but each was preceded by a sponsor topline release: TRIUMPH-2's weight and A1C results on 23 July 2026 [18,19] and ACHIEVE-4's primary cardiovascular result on 16 April 2026 [20,21]. The sponsor pages returned HTTP 403, so syndicated copies of both releases were read; the trade reports corroborate the dates. The readout registered for ACHIEVE-4 is a kidney-outcome analysis (LBA 25), whose earlier disclosure was not established. SMARTEST compared dapagliflozin with metformin as first-line therapy; it had a design paper and baseline data, and no primary-results publication was found by 11 October 2026 [22,23]. OCTANE, added after registration, was presented at EASD as a late-breaking abstract [5]. The paper therefore classifies each readout by what was new at EASD: **first disclosed at EASD as far as could be established** (SMARTEST, OCTANE and the registered ACHIEVE-4 kidney-outcome analysis, LBA 25); **full data first disclosed at EASD, topline public earlier** (TRIUMPH-2; and the ACHIEVE-4 primary cardiovascular result, which is not itself a registered readout); and **re-presented or secondary analyses** (REIMAGINE 1 and most of the ACHIEVE programme). The appetite abstract 122 and the proteomics study (LBA 29) are secondary analyses whose novelty at EASD was not assessed. For TRIUMPH-2, latency from the EASD session measures arrival of the full data, not of the headline result. A topline source dated before 28 September fails the date rule for E1 classes (c) and (d), so an answer stating a TRIUMPH-2 or ACHIEVE-4 cardiovascular headline figure without a congress-dated source may reflect the earlier topline rather than the congress. This is interpretation; it changes no endpoint.

1.2 How AI engines produce treatment answers

A language model answers from what it absorbed in training ("parametric" knowledge), from documents retrieved when the question is asked, or from both [24]. A model's accuracy on a fact tracks how many relevant documents it saw in training [25], and its effective knowledge cutoff often differs from the one

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

reported [26]. A result absent from a model's training can appear through information supplied at inference, including retrieval or prompt context, or through subsequent model or service updates. The study observes citations and answer text, not all of those mechanisms. Retrieval depends on what is indexed, ranked and placed in context: a single chatbot response retrieved about 39% of the studies experts would include for a medical question (preprint) [27]; search results improve answers on fast-changing questions, and the number and order of retrieved items matter [28]; in medicine, retrieval changed accuracy and showed position effects [29], and the corpus searched shaped answer quality [30]. In a large study of Chinese-language generative search, cited pages had a half-life of about 39 days for timely queries, and about 13% of brand exposures could not be matched to the cited pool [31]. Commercial engines mix the two routes in undisclosed ways [32].

1.3 Why changes in answers are hard to interpret

Answers vary for reasons unrelated to evidence. Hosted models are not deterministic even at settings meant to make them so [33,34], a variation one non-peer-reviewed account attributes to server-side batching [35]. Services drift: the "same" commercial model changed materially within three months in 2023 [36], and one service's handling of health disinformation had changed at a 12-week re-test [37]. The same guideline questions asked three times a day for ten days showed significant day-to-day accuracy variation in two of four engines [38]. Identical prompts almost never return the same brand list, although appearance rates are steadier [39], and visibility is better reported as a distribution over repeated runs [40]. In this study's own pre-congress record, one engine's naming of a UAE medicine stepped 16 points in September as a clinic's page entered its citations, with no congress involved [A1].

Citation does not settle the question. Only about 51.5% of sentences from four generative search engines were fully supported by their citations [41]; the best models in a citation benchmark lacked complete support about half the time on one dataset [42]; in medicine, 50–90% of responses were not fully supported by their cited sources [43]; without retrieval, models fabricate references [44]; and AI search tools often cite syndicated copies rather than the original publisher, or broken links [45]. Answer content and human decisions also differ: in a preregistered randomised study of 1,298 participants, models alone identified the relevant condition in 94.9% of scenarios, but participants using them did so in fewer than 34.5% of cases [46]. Naming can move without any change in evidence: in a preprint covering 258,660 model calls, advertisements in system prompts raised selection of the advertised drug by 12.7 percentage points where two options were guideline-appropriate [47].

1.4 Medical-information research and reporting

Most studies of chatbot health information judge accuracy at one time point [48–50], with a few re-queries over short intervals [51]. Answers that were usually accurate still included some that experts judged potentially harmful [52]; GLP-1 receptor agonist answers were reproducible but "lacked updated coverage of emerging issues" [53]; and an incorrect answer persisted after a model update [54]. Of 137 chatbot health-advice studies, 99.3% did not identify the model version and only 39.4% reported the query date [55]. CHART now sets expectations for model identification, prompts and protocol [56], alongside TRIPOD-LLM [57] and STROBE [58,59].

1.5 Generative engine optimisation and measurement

Generative engine optimisation (GEO) was introduced as editing content to raise its visibility in generative answers, with gains of up to 40% in an experimental setting [60]. Crafted text can raise a product's chance of being recommended [61] and steer production search engines [62]. Evaluations are less favourable: most conversational-SEO methods were largely ineffective in a multi-actor benchmark [63], and a survey of 45 GEO studies found no technique with a stable, longitudinal, cross-platform

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

causal effect [64]. AI search has been reported to favour earned media [65]. Careful practitioner measurement already reports visibility as a distribution over repeated runs [39,40]; this study extends that discipline from visibility to attribution. GEO asks whether content changes can raise visibility; this study asks whether a change in answers can be distinguished from ordinary variation and linked to the sources cited in the answers that changed.

1.6 What is and is not new

The closest design relative queried 12 models almost daily through the 2024 US election season and could not trace most step changes to causes [32]; longitudinal refusal audits detected policy changes [66], and their authors note that coincidence with dated events cannot establish a causal link [67]. Repeated health measurements exist [37,38,51,68]. Within the documented structured search, we did not identify a study meeting the stated criteria (below) for surveillance around a dated evidence event. That answers vary, that citations often fail to support claims, and that retrieval surfaces fresh material are consistent with this prior work, not new.

The criteria were a registered protocol for surveillance of AI-generated treatment answers around a scientific congress, answer-by-answer tracing of congress-derived sources, and a pre-specified rule for attributing answer movement to the event. In a structured (not systematic) search of PubMed, arXiv, ClinicalTrials.gov, the ACL Anthology and the web conducted on 11 October 2026, we did not identify a study meeting these criteria in the documented search.

---

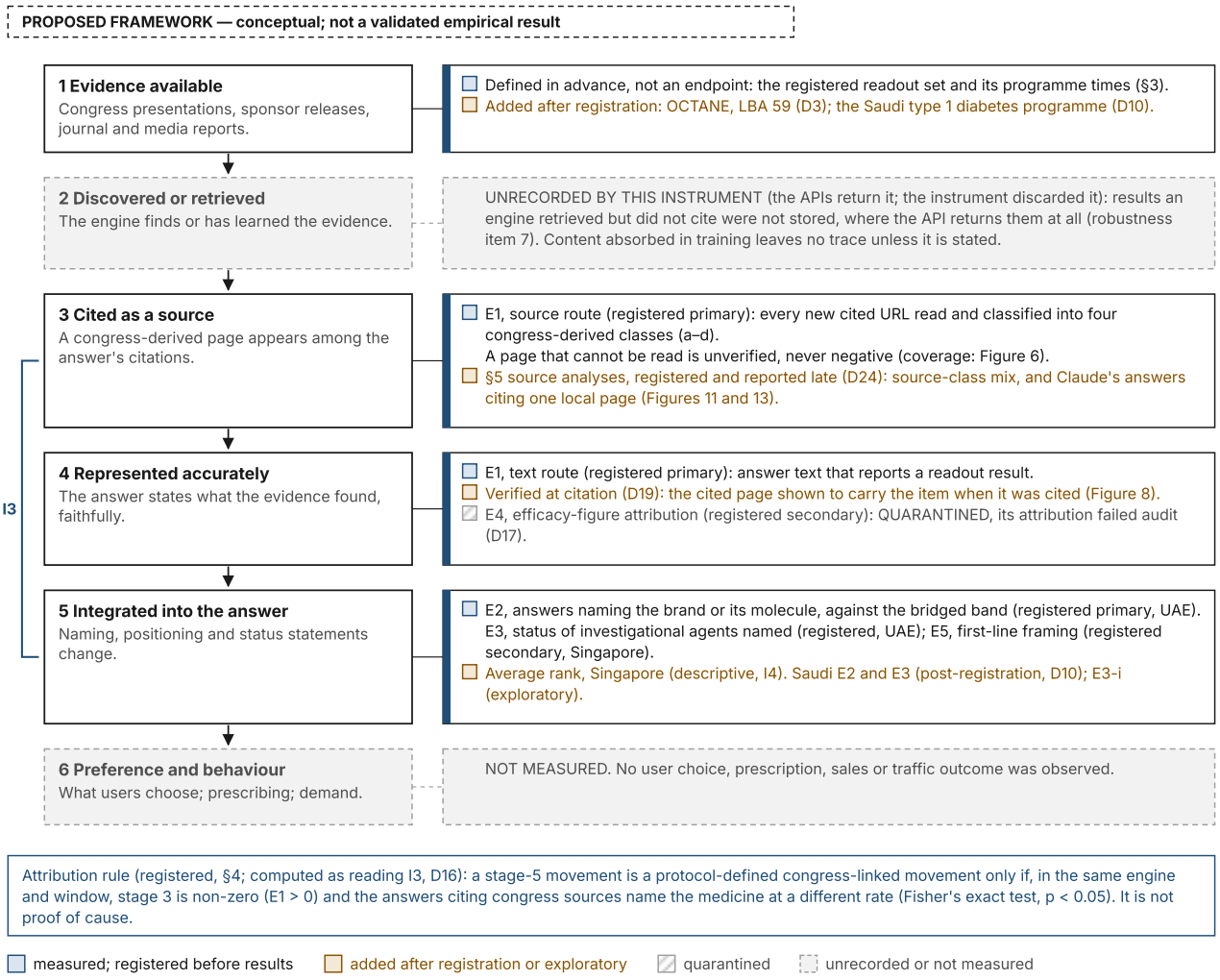
2. A proposed framework: from evidence to answer

The central question is: **when new scientific evidence enters the public domain, what changes in AI-generated treatment guidance, how can those changes be measured, and what evidence permits them to be attributed to the scientific event?**

Figure 2 sets out a proposed six-stage chain. It is a conceptual model, written after the protocol and fitted to its registered attribution rule, not a validated result; the study does not test that the stages are causally ordered or complete.

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

Figure 2 **From evidence to answer: what this study measured**



Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. Endpoint definitions and registration status: easd-2026-evidence-to-answer-prereg.md (§3, §4, §9 D3, D10, D16, D17, D19). The framework is the authors' proposal; the study does not test that the stages are causally ordered or complete.

Figure 2. Measurement architecture (proposed framework). A proposed conceptual chain from the availability of new evidence to downstream behaviour: (1) evidence available; (2) discovered or retrieved; (3) cited as a source; (4) represented accurately; (5) integrated into the answer through naming, positioning and status statements; (6) preference and behaviour. For each stage, the endpoint that observes it and its registration status; stage 2 was not recorded by this instrument, and stage 6 was not measured. The figure's "every new cited URL read and classified" means that new URLs were fetched and screened and candidate retrieved content was classified (Section 4.5). E4 is shown as quarantined (D17). The bracket marks the registered attribution rule (I3), linking citation (stage 3) to integration (stage 5). Conceptual; no data plotted. Source: protocol §3, §4; D3, D10, D16, D17, D19.

Stage	Endpoint in this study (status)	What was not observed
1. Evidence available	Not an endpoint: readout set and times fixed in advance; OCTANE (D3) and the Saudi programme (D10) added later	Whether each result was new or re-presented is interpretation (Section 1.1)
2. Retrieved	None. In part observable through the APIs (issued queries, retrieved but uncited results), but not recorded by this instrument; training-data inclusion is not observable in a ten-day window	Queries issued; results retrieved but not cited

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

Stage	Endpoint in this study (status)	What was not observed
3. Cited	E1 source route (registered primary; new URLs fetched and screened, candidate retrieved content classified); registered §5 source analyses (reported late, D24)	Unreadable pages (unverified, not negative); pages updated after an earlier citation; derivative pages that paraphrase congress material without a qualifying date
4. Represented	E1 text route (registered primary); "verified at citation" (D19 label) by archive or, here, by echo, which shows textual overlap only (D19, D24); E4 (quarantined)	What a page said when cited, unless archived; echo shows overlap, not complete historical content
5. Integrated	E2 naming (registered primary, UAE); E3 status statements; E5 first-line framing; Saudi E2, E3, E3-i (post-registration)	Why integration changed when no source accompanies it
6. Preference and behaviour	None	Everything at this stage

The framework's practical consequence is that **attribution is a property of links, not of stages**: a change at stage 5 can be linked to a source only if stage 3 or 4 shows that source in the same engine, window and answers, and the answers carrying it differ from those that do not. This is the logic of the registered attribution rule. A link test of this kind is neither necessary for an effect (evidence can travel uncited) nor sufficient for a cause (citing answers may come from particular questions). The framework is meant to transfer to other evidence events and industries; this study does not demonstrate that.

3. Research questions and study design

3.1 Questions and endpoint hierarchy

The registered objective asked whether the evidence presented at EASD 2026 reached what four AI engines tell people about established diabetes and obesity medicines, how fast, through which sources, and whether it changed how those medicines are named, positioned or described [A2].

	Question	Endpoint	Status
Q1	Did congress evidence enter the answers, and how long after presentation?	E1 penetration and latency	Registered primary (UAE, Singapore)
Q2	Did the measured medicine's naming move beyond ordinary variation?	E2 naming movement	Registered primary (UAE); descriptive positioning (Singapore)
Q3	Did investigational agents presented at the congress enter answers, and are they described as unapproved?	E3 status statements	Registered (primary UAE, secondary Singapore)
Q4	Did the clinical content change: efficacy figures, and the place of SGLT2 inhibition relative to metformin?	E4 efficacy figures; E5 first-line framing	Registered secondary; E4 quarantined (D17)

The protocol fixed in advance what would count as a reportable result, including "nothing moved, and nothing from the congress arrived" [A2].

3.2 Markets and medicines

Series	Medicine	Market	Questions per cycle	Scored per engine per cycle	Registration
UAE	semaglutide (Wegovy)	United Arab Emirates	60	54	Registered
Singapore	empagliflozin (Jardiance)	Singapore	26	19	Registered
Saudi Arabia	teplizumab (Tzield)	Saudi Arabia	56	55	Post-registration (D10)

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

The UAE and Singapore series were the two with a pre-congress baseline published in Signal Note 04 on 24 September 2026 [A1]. Their endpoints differ, so the markets are not compared. In the UAE, single pre-congress cycles named semaglutide (Wegovy) in between about 20% and 74% of scored answers, depending on the engine, so naming could move either way. Singapore naming of empagliflozin (Jardiance) was near ceiling: the combined-engine cycle rate ranged from 98.7% to 100%, while individual-engine cycle rates ranged from 94.7% to 100%, so its registered E2 measure is positioning, reported descriptively. The Saudi arm was added on 5 October to test the design in a type 1 diabetes market with an existing pre-congress record (D10); everything in it is exploratory. Teplizumab is not a glucose-lowering treatment: its US label indicates it to delay the onset of Stage 3 type 1 diabetes in patients aged 1 year and older with Stage 2 disease, with a further accelerated-approval indication in patients aged 8–17 recently diagnosed with Stage 3 disease [69]. Its Saudi registration status was not established in the study record. Its naming rate is therefore not comparable with the UAE rates.

The three products are named consistently, whatever their results, generic name first. Naming implies no endorsement, sponsorship or commercially meaningful effect. The decision to name them (D22) was made after preliminary and final results were available; it is not pre-registered and is not neutral in perception. Every finding describes the engines' answers, not the medicines.

3.3 Engines, cycles and windows

The four engines are developer API endpoints, not the vendors' consumer applications, queried programmatically by AlphaCitation's instrument:

- **ChatGPT:** gpt-4o-mini on the OpenAI Responses API with the web search tool; the model decides per request whether to search.
- **Claude:** Claude Haiku 4.5 (claude-haiku-4-5) on the Anthropic Messages API with the web search server tool (at most five searches per request); the model decides whether to search.
- **Gemini:** Gemini 2.5 Flash with Google Search grounding; grounding is model-decided.
- **Perplexity:** sonar on chat completions; retrieval on every request.

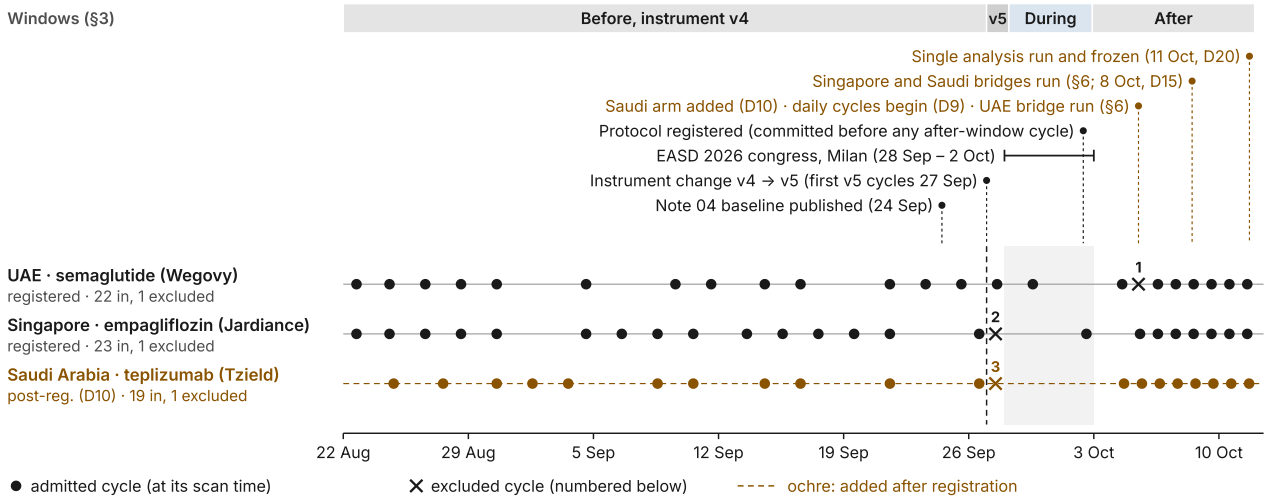
The engine names below refer to these endpoints. No user location is sent with any request; the market exists only in a framing instruction carried by every question, whose wording is withheld (Section 4.9). A **cycle** is one complete run of a market's battery through all four engines; an **answer** is one engine's response to one question in one cycle. A cycle was admitted only if it used the same four models, captured at least 95% of its questions, matched the series' first question set at the rate already recorded, and had no cached answers; failing cycles were excluded, never repaired. The study repeatedly sampled a maintained battery of question slots. Some slot texts differed across cycles; analyses requiring identical text used an explicitly matched subset. Prompt matching, exclusions and their denominators are reported separately from slot identity.

Windows were fixed in UTC: **before** (on two instrument versions, v4 and v5), **during** the congress (28 September 06:00 to 2 October 22:00) and **after** (cycles starting after 2 October 22:00 and completing by 11 October 23:59). UAE and Singapore cycles ran at about 14:05 UTC; Saudi cycles at about 14:05 UTC before the congress and 16:30 UTC from 4 October; all three series ran daily from 5 to 11 October (D9).

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

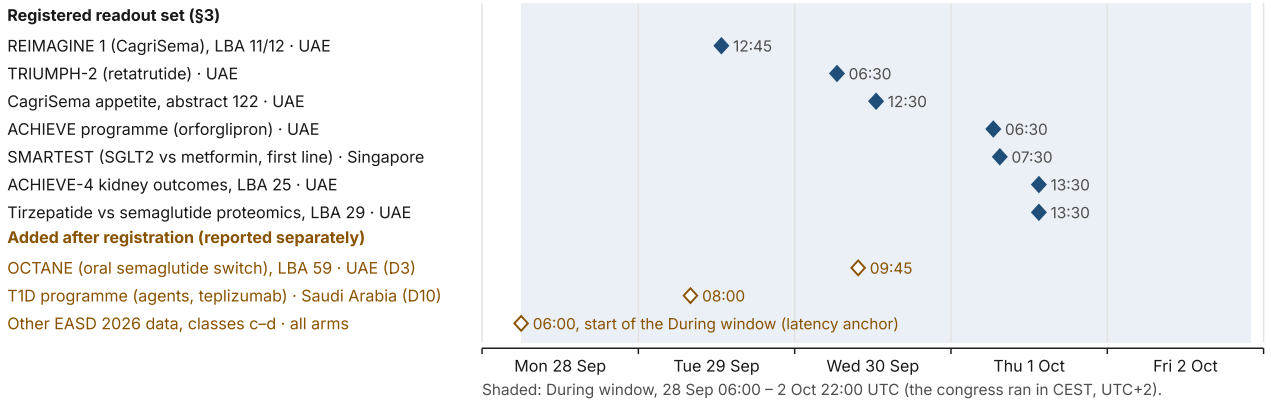
Figure 1 **Study timeline, 22 Aug – 11 Oct 2026 (UTC)**

A **Whole observation period**



1 UAE, 5 Oct (UAE-C17): captures for 15 of 60 prompts; served from cache (D14).
2 Singapore, 27 Sep (SG-C16): captures for 17 of 26 prompts; fails the §2 capture rule (§3).
3 Saudi Arabia, 27 Sep (SA-C12): captures for 9 of 56 prompts; fails the §2 capture rule (D13).

B **Congress week: readout presentation times (UTC); latency in Figure 7 is counted from these**



Shaded: During window, 28 Sep 06:00 – 2 Oct 22:00 UTC (the congress ran in CEST, UTC+2).

Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. Cycles: easd-2026-final-coding.json and easd-2026-saudi-final-coding.json (admissions). Readout times: easd-2026-e1.json (readouts; EASD 2026 Final Programme, verified 2 Oct; OCTANE verified 11 Oct, D21). Registration and deviation dates: easd-2026-evidence-to-answer-prereg.md (§3, §9); bridge dates: easd-2026-bridge.json (ranOn).

Figure 1. Study timeline. Panel A: the four observation windows (before on v4; before on v5, from 27 September to 28 September 06:00 UTC; during, 28 September 06:00 – 2 October 22:00 UTC; after, 2 October 22:00 – 11 October 23:59 UTC); the instrument change v4 → v5 (27 September); the congress (28 September – 2 October); the protocol registration (2 October, 09:50 UTC) and the dated deviations that followed; and every cycle of the three series, admitted (dot) or excluded (cross, with the reason). Panel B: presentation times of the seven registered readouts and of those added after registration (OCTANE, LBA 59, D3, 30 September 09:45 UTC; the Saudi type 1 programme, D10; other EASD 2026 data, the latency anchor for E1 classes c and d). Period 22 August – 11 October 2026. Units: UTC dates and times. N: 67 cycles, 64 admitted (UAE 22, Singapore 23, Saudi Arabia 19); 10 readouts (7 registered, 3 post-registration). Source: final coding files (admissions); E1 output (readouts); protocol §3, §9.

Every after-window observation sits on instrument v5. One admitted v5 cycle precedes the congress in the UAE series; none does in Singapore (its 27 September cycle failed the capture rule) or in Saudi Arabia. The v4 bands therefore had to be translated into v5 terms before they could judge the after window.

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

3.4 The readout set

Seven readouts were registered with their programme times: REIMAGINE 1 (LBA 11/12; 29 September; NCT06323174 [70]), TRIUMPH-2 (30 September; NCT05929079 [71]), the cagrilintide–semaglutide (CagriSema) appetite abstract 122 (30 September), the orforglipron ACHIEVE programme (1 October), SMARTEST (dapagliflozin versus metformin; 1 October; NCT03982381 [72]), ACHIEVE-4 kidney outcomes (LBA 25; 1 October; NCT05803421 [73]) and a tirzepatide–semaglutide proteomics study (LBA 29; 1 October) [2]. SMARTEST's SGLT2 inhibitor is dapagliflozin [22], so for the Singapore series it is class evidence, not trial evidence about empagliflozin. OCTANE (LBA 59), a retrospective real-world cohort of patients switching from injectable obesity medicines to oral semaglutide, was added after registration (D3) [5]; its programme slot (30 September, 11:45–12:45 CEST) was verified on 11 October (D21). Latency is measured from the presentation time. TRIUMPH-2's sponsor release was datelined 29 September, ahead of its session [74]; at day resolution this changes no latency.

3.5 Attribution rule

The registered rule calls a movement a *congress effect* only if, in the same engine and window, E1 is non-zero **and** the answers carrying congress-derived sources name the medicine at a rate different from that engine's other answers. Every other combination is reported by name: *movement without congress sources*, *congress sources without movement*, or *neither* [A2]. The difference is tested with a two-sided Fisher's exact test at $\alpha = 0.05$ (reading I3, D16); a movement meeting the rule would be a "protocol-defined congress-linked movement", not proof of cause. The test was registered without a feasibility estimate: with one qualifying answer among an engine's scored answers it cannot reach $p < 0.05$ whatever that answer says, with two it could reach no lower than 0.206 in Gemini's UAE table, and with none it cannot run (Section 5.1).

3.6 Registration and its timing

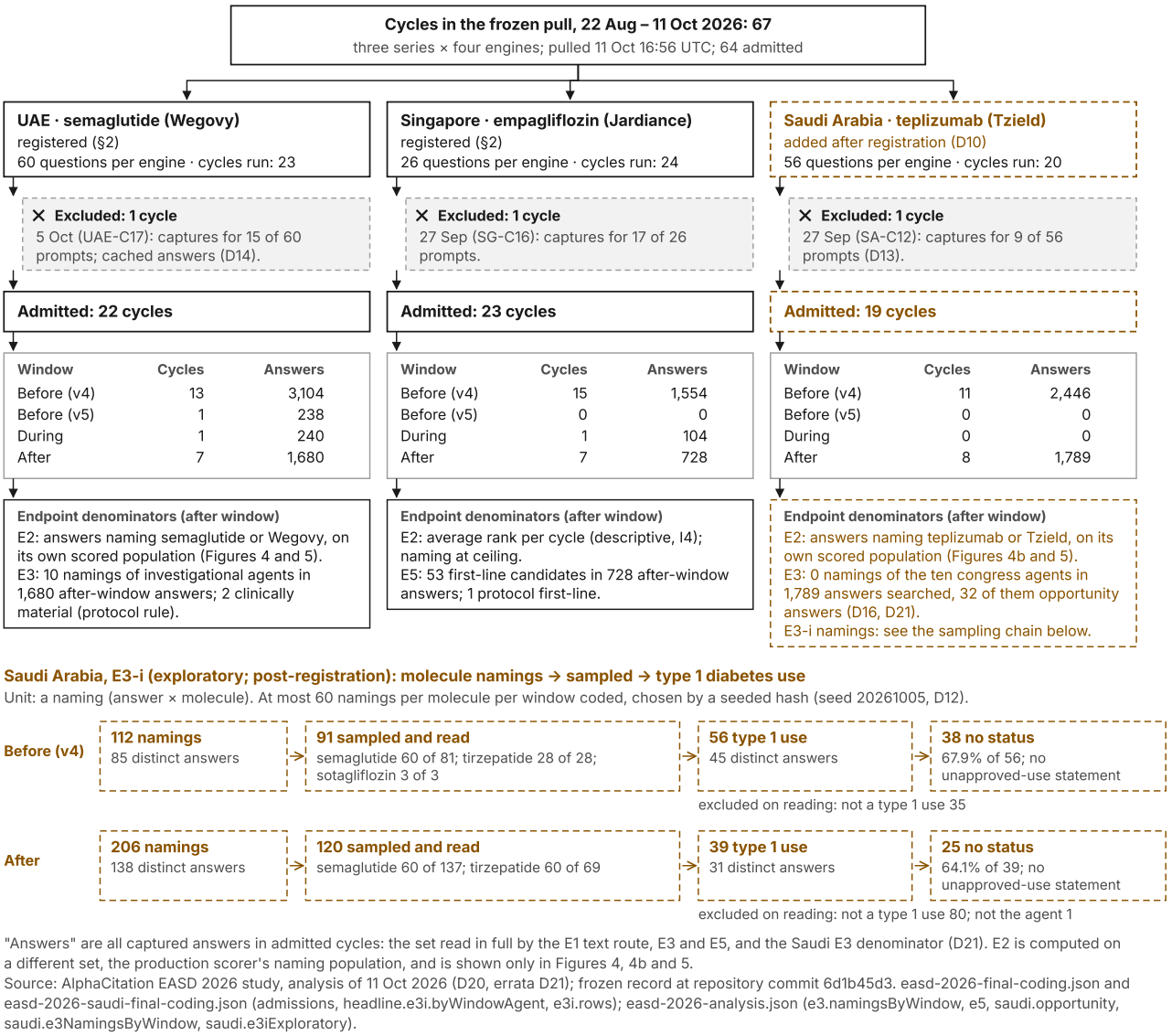
The study protocol was committed on 2 October 2026, following initial congress-period observations but before collection of post-congress observations. The registration chronology, prior data exposure, subsequent amendments, and associated methodological limitations are documented in the technical supplement. The protocol was committed to AlphaCitation's repository at 09:50:05 UTC on 2 October 2026; that timestamp is the repository's own commit record, not a third-party registry. It preceded every after-window cycle. It did not precede the congress (all seven registered readouts had been presented) or the UAE congress-window cycle of 29 September, whose naming counts were in the record it rests on; in that cycle Claude named the medicine in 31 of 54 scored answers, more than in any v4 cycle. The bands and rules were therefore fixed before any after-window data existed, but not before all congress-window data had been seen (Supplement §1.3).

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

4. Methods

4.1 Observation and admission

Figure 3 **Observation flow: cycles, answers and endpoint denominators**



"Answers" are all captured answers in admitted cycles: the set read in full by the E1 text route, E3 and E5, and the Saudi E3 denominator (D21). E2 is computed on a different set, the production scorer's naming population, and is shown only in Figures 4, 4b and 5.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-final-coding.json and easd-2026-saudi-final-coding.json (admissions, headline.e3i.byWindowAgent, e3i.rows); easd-2026-analysis.json (e3.namingsByWindow, e5, saudi.opportunity, saudi.e3NamingsByWindow, saudi.e3iExploratory).

Figure 3. Observation flow. Observation units in the frozen pull (22 August – 11 October 2026): cycles run, excluded (with the failed rule and its deviation number) and admitted per series; admitted cycles and captured answers per window; the after-window denominators of each endpoint; and, for the Saudi exploratory endpoint E3-i, the chain from molecule namings to the seeded sample (at most 60 per molecule per window) to namings coded as a type 1 diabetes use and those lacking qualifying indication-status wording under the study's rules. E2's scored population is a different set and appears in Figures 4, 4b and 5. Units: cycles, answers, or namings (answer × molecule). N: 67 cycles, 64 admitted. Source: final coding files; analysis output.

Three cycles were excluded under rules fixed before results: two pre-congress capture failures on 27 September (Singapore, 17 of 26 questions captured; Saudi Arabia, 9 of 56) and the UAE cycle of 5 October, in which 45 of 60 questions per engine were served from a 24-hour cache (D14). The UAE after window therefore has seven admitted cycles. Admitted cycles contain 5,262 UAE, 2,386 Singapore and 4,235 Saudi answers.

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

4.2 Populations, units and the naming measure

E2 uses a **scored (governed) population** that excludes questions asking directly about the medicine or setting it head to head with a named competitor, which name the medicine by construction: 54, 19 and 55 scored answers per engine per cycle. E1, E3, E4 and E5 read **every captured answer**. In the Saudi endpoint E3-i, the unit is a **naming** (one molecule named in one answer); at most 60 namings per molecule per window were coded, chosen by a seeded hash. Rates on different units are never combined.

The primary naming endpoint records whether an answer mentions the measured brand or molecule, including mentions that do not recommend its use. For the UAE series, molecule-only semaglutide mentions can refer to products other than Wegovy. The registered E2 measure is whether an answer **names the brand or its molecule**: the production scorer credits the brand, its configured aliases and, where the scan grants molecule credit, the molecule. For the UAE series that means Wegovy or semaglutide; for Singapore and Saudi Arabia molecule credit was off and the alias lists named the molecule (Jardiance or empagliflozin; Tzield or teplizumab). Brand-only results, computed after the results (D24), are reported separately. Re-scoring all 64 admitted cycles reproduced the stored naming counts with no mismatch. This verifies deterministic consistency of the pipeline. Naming accuracy was not independently assessed against human-coded answers.

4.3 Bands, the bridge and its self-check

For each engine, the registered **band** is the range of single-cycle naming rates across the pre-congress v4 cycles (for Perplexity, only the cycles from 9 September, after the step documented in Signal Note 04 [A1]). A band is a decision rule in the spirit of statistical process control [75], not a control chart with derived limits. A movement is declared only if the pooled after-window rate lies outside it.

On 27 September the instrument changed (v4 to v5): a sampling setting was pinned on three engine paths; ChatGPT's path, already pinned, did not change. A registered **bridge** ran each series' battery six times on one day (5 October for the UAE, 8 October for Singapore and Saudi Arabia), interleaving the two settings three runs each, and estimated each engine's offset (v5 minus v4) with a bootstrap 95% interval over questions. ChatGPT, whose path did not change, served as a **negative control**; the bridge script recorded it as holding ("VALID", a script decision) in all three series. The unchanged-path control showed no detected shift within its estimated interval. No equivalence margin was prespecified. Because same-day pairing removes day effects for every engine, the control can detect only run-order or request-path artefacts; with an interval of −6.2 to +2.5 points in the UAE, it sets a low bar. The bridge estimates a same-day setting difference and assumes that this offset applies to the before/after comparison; transport across dates and prompt contexts was not independently established. The **bridged band** shifts the band's lower end by the interval's lower bound and its upper end by the upper bound (reading I1, D16). For Claude in the UAE the offset was +3.1 points (95% interval −6.2 to +11.7), widening the band from 40.7–55.6% to 34.53–67.33%. The protocol's words ("shifted by the offset … with each end widened by the offset's 95% interval") admit other readings; their verdicts are reported in Section 5.1. A house scale labels the size of a change against the bridged pre-congress reference (SMALL < 5 points; POTENTIALLY RELEVANT 5–10; DECISION-RELEVANT 10–15; MAJOR ≥ 15); it describes size only, and is neither a significance test nor a judgement of commercial relevance.

The protocol also registered a **self-check** on the bridge: the spread of the three same-day v4-setting runs "should reproduce the v4 band's width", with the v5 spread beside it. No threshold was registered, and the analysis first omitted the check; it is reported under D23 and corrected under D24 (Section 5.1).

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

4.4 Uncertainty

The registered decision compares a pooled rate with a band and carries no p-value. The protocol stated that 378 answers per engine resolve about ±5.0 points; that is the binomial half-width of the after-window rate, not the change the rule could detect. As executed, each engine's upper band edge sat 9.2 (four engines) to 23.4 points (Gemini) above its bridged pre-congress reference: 16.4 for Claude and 12.3 for Perplexity. Answers to the same question are correlated across cycles, so a post-registration analysis (D16) resampled whole questions 2,000 times for a **question-clustered 95% interval** of each after-window rate [76,77]. Under D19, a verdict whose clustered interval contains the nearer band edge is **borderline**; it is never reversed. That interval describes the after-window level, not the change, and it holds the cycles fixed; a cycle-level change estimate with an interval was therefore added under D24.

4.5 Source tracing (E1)

E1 counts answers that **cite** a congress-derived source or whose **text reports a result** of a readout. A source is congress-derived if reading the page places it in one of four classes: (a) the congress's own pages; (b) a journal publication of a readout trial; (c) a sponsor release dated 28 September or later reporting congress data; (d) news or medical media dated 28 September or later reporting congress data. Every URL first cited in the congress or after window, and not cited by the same series before, formed a new series–URL pair. All new series–URL pairs were subjected to an automated fetch and term screen. Candidate pages were classified from retrieved content, with the qualifying sentence recorded: 5,268 new series–URL pairs were screened, 114 candidates read by hand, and the ledger holds 122 entries. Of 5,268 pairs, 947 remained unreadable after retry. Both the source and text screens have incomplete and unquantified recall, including known misses of unnamed readout results. A page that could not be fetched (failure or HTTP status 400 or above) is **unverified**, never negative; "read" is defined by HTTP status, so a consent wall or bot page served with status 200 would count as read, and coverage is an upper bound. A readout carried by a page today but not shown to have been there at citation (no archived copy, not echoed by the answer) is **unverified at citation** (D19). Translational time lags are measured too heterogeneously to compare without an agreed definition [78]. Here **latency** is the number of days from the specified presentation or programme anchor to first qualifying detection in an admitted sampled cycle. The previous negative and first positive cycles bracket sampled detections, not the engine's true evidence-uptake time. Detection is imperfect and may be non-persistent. With no qualifying answer a cell is "not detected by day N", where N runs to the series' last admitted cycle (11 October, about 14:05 UTC; Saudi Arabia about 16:34 UTC); this refers to this battery and observation schedule. Programme-anchored exploratory rows do not give readout-specific uptake times.

4.6 Status, framing and efficacy coding (E3, E5, E4)

E3 read every answer naming an agent from the readout set that held no approval in the market on the cycle date. Status came from the regulator's own record: the Emirates Drug Establishment (EDE) directory for the UAE [79] and the Health Sciences Authority register of therapeutic products for Singapore [80]. US and EU status, recorded beside them for context and never scored, came from Drugs@FDA through openFDA [81] and the EMA medicines data [82]. In Saudi Arabia, status could not be established from SFDA records: the searchable register is behind a reCAPTCHA form, and the readable listing covers new approvals from 1 February 2025 to 2 February 2026, so a date in it is a listing date, not a first approval (D12) [83]; every Saudi code is provisional. Orforglipron was EDE-registered from 2 April 2026, so UAE answers naming it (1,190) were out of scope. *Status correct* required an explicit statement that the agent is investigational, in trials or not approved; *status absent*, naming without such a statement; *status wrong*, presenting it as available. The registered "clinically material" count is status absent or wrong in answers to a question asking which medicine to choose.

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

After 2 October the EDE directory stopped returning semaglutide, a product known to be registered, so later status could not be trusted; UAE status is established only to 2 October and after-window codes are provisional (D15, D20).

E5 read every Singapore answer containing a first-line term and coded whether SGLT2 inhibition was framed as first-line instead of or ahead of metformin, for patients generally or for high-risk patients only; "first-line together with metformin" was reported but not counted (D2). The registered reference frame is the 2022 ADA/EASD type 2 consensus [6], fixed before results; answers were not coded against later standards of care or national guidance.

E4 coded every numeric efficacy figure stated for semaglutide (Wegovy) to its source, against reference values fixed before coding. Section 7.4 explains why its scripted categories were quarantined.

4.7 Coding, adjudication and the second pass

Hand coding, ledger classification and adjudication were model-assisted: performed by Anthropic Claude models under the author's direction, with every code, note and adjudication recorded. One of the engines studied is a Claude model. A blind second pass by a separate instance of the same model family re-coded every E3 code, the non-trivial E5 codes plus a seeded sample of 20 not-first-line codes, and every sampled E3-i type 1 use or not-the-agent code plus a seeded sample of 20 not-a-type-1-use codes, with a top-up on 11 October; 14 E3-i items later displaced from the capped sample remain in the agreement statistics (D16). The adjudicator was the coordinator that produced the first codes. Agreement is reported raw and as Cohen's kappa [84–86]. Model annotation can match crowd-worker accuracy [87], but evaluators favour outputs of their own family [88]; the second pass measures **consistency within a model family, not independent validation**. The interpretive coding was also checked by blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader: a seeded representative sample and a risk-enriched sample built from hard cases by fixed rules, coded without access to the original codes, the engine or the date (Section 7.5). Because the reader is the author, this is not independent or external validation.

4.8 Registered protocol and deviations

Registering the protocol is used here as disclosure of analytic flexibility, not a guarantee of validity [89–91]. Every change after registration is an appended, dated deviation (D0–D24). Table 1 groups them; the supplement gives each with its commit and what had been observed when it was written (Supplement §2.2).

Table 1. Deviations from the registered protocol, grouped.

Group	Deviations	What changed	Observed when written	Effect on results
Timing and coding clarifications	D0, D1, D2, D5, D8	Baseline coding time; E5 reading unit; "first-line with metformin" reported, not counted; option-listing convention; an example corrected	Before any after-window answer was coded	None on any registered comparison
Added readouts and arms	D3, D10, D12, D13	OCTANE added; Saudi arm added with its own agent set, rules and admission threshold	Before after-window coding	Every OCTANE and Saudi figure is post-registration and exploratory
Exploratory additions	D6	Status errors about approved medicines (E3b)	Before after-window data	No E3b figure in the frozen outputs
Instrument and operations	D7, D9, D11, D14, D15	Coder byte defect fixed; daily cadence from 5 Oct; bridge rerun after a tooling failure; cache cycle excluded; bridges on 8 Oct; regulator control check	Before or during after-window collection; some naming counts displayed during bridge validation	Seven, not eight, UAE after-window cycles; UAE status established only to 2 Oct

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

Group	Deviations	What changed	Observed when written	Effect on results
Analysis readings and robustness	D16	Readings I1–I5 of the registered rules; clustered intervals; strict scenarios; Saudi denominators; blind second pass	After the hand-coded E3, E5 and E3-i tallies and the UAE per-engine naming counts through 10 Oct had been displayed	I1 determines how the bands are bridged, and therefore the UAE verdicts (Section 5.1)
Quarantine	D4, D17	E4 hand check (D4); audit of scripted attributions (D17)	D4 before after-window data; D17 with preliminary results visible	E4 scripted categories not published
Source adjudication and safeguards	D18, D19	Whole-article reading of sources; cross-arm reporting; full-precision E2; "unverified at citation"; freeze record	Preliminary results visible	No registered cell changed
Date move	D20	Analysis ran on 11 Oct, not 12 Oct, after the window's last scheduled cycle	Final run	Date only
Errata after the freeze	D21	Saudi E3 denominator; quoted question removed; OCTANE origin verified; times and counts corrected; class (b) limitation recorded	After the final analysis	No registered verdict or result changed; OCTANE latencies about 0.4 day shorter
Naming the medicines	D22	The three medicines named in the publication	After preliminary and final results	None
Omitted registered check reported	D23	Registered bridge self-check reported; it does not hold as worded	After the final analysis	No registered verdict changed
Post-registration robustness	D24	Registered §5 source analyses reported late; band readings; cycle-level change; per-cycle view; corrected self-check; brand-only naming; search rate; attribution sensitivity; Saudi use context; per-answer outcome table	After the results and the internal reviews	No registered verdict changed; Sections 5–7

Reporting follows STROBE principles [58,59] and CHART [56], with one departure, set out in Section 4.9.

4.9 Question-instrument availability, and what is disclosed instead

> **Question-instrument availability.** The exact question battery is withheld to preserve the integrity of AlphaCitation's longitudinal measurement instrument. Public disclosure of the active questions could enable content to be optimized specifically against the instrument, potentially influencing future observations and compromising comparability over time.
>
> To support scientific scrutiny, the study reports its question taxonomy, sampling design, endpoint definitions, coding procedures, analytical methods and reproducibility controls. The complete instrument is preserved in the frozen study record.
>
> This restriction represents a departure from CHART Item 5b and limits independent replication of the exact data-collection procedure.

The optimisation risk is a possibility, not a demonstrated effect; studies have shown that crafted content can steer what generative engines recommend [61,62]. The public disclosure boundary has been amended to release battery sizes and taxonomy, selected calibration results, version/deviation history, endpoint rules and approved coding diagnostics. Exact active question wording, the v5 parameter and its values, production scoring internals, unreleased implementation and full raw answers remain withheld. The amended boundary supersedes the conflicting descriptions in the earlier protocol; the original text remains available as provenance. What is disclosed:

- **Question taxonomy.** Batteries of 60 (UAE), 26 (Singapore) and 56 (Saudi Arabia) questions. *Focused* questions ask about the measured medicine directly or set it head to head with a named

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

competitor and are excluded from the naming population (6, 7 and 1). The rest are *brand-neutral* category questions. *Which-medicine* questions (asking the engine to choose, rank, recommend, compare or list alternatives) are identified from the question text by one fixed rule; 15 of the 21 in-scope E3 namings came from them. In Saudi Arabia, questions inviting new or emerging treatments are identified by a published word-list rule. The UAE and Singapore batteries use the instrument's general persona; every question carries a market-localised framing instruction, whose wording is withheld.

- **Sampling design.** One answer per engine per question per cycle; fixed schedules and admission rules (Section 3.3). Every sample was drawn by a recorded seed: the E3-i cap (20261005), the D17 audit of 150 of 1,082 efficacy figures (20261010) and the second pass (20261011). Bootstraps use 2,000 resamples with recorded seeds.
- **Coding procedures.** E3 searched every answer for each agent's names and development codes, and every hit was read. E5 searched a fixed list of first-line terms (for example "first-line", "initial", "instead of metformin", "monotherapy"), read at paragraph level with headings (D1), and applied codes in a fixed order. Scheme decisions (S1–S11; Saudi T1–T9), ambiguity labels and reason codes are recorded with every code, its answer span and its note. E1 fetched each new page once (20-second timeout, 1.5 MB cap), retried unreadable pages once with an ordinary browser agent, and resolved redirect links without following them.
- **Challenge paths.** A per-answer outcome table without question or answer text (D24) allows E2, the band readings, the attribution tests and the clustered and per-cycle analyses to be recomputed (in that table, `e1Qualifying` is the frozen E1 classification; the corrected attribution calculation (Section 5.1) additionally uses the E1 correction record (`robustness/easd-2026-e1-corrections.csv`), which identifies the corrected answer by the table's public identifiers); each code, ledger entry and adjudication can be contested from its recorded note and span; each seeded sample can be checked against its seed. Re-scoring answers and re-running the E1 screen require the frozen pull, which is held because it contains the question text and is identified by its SHA-256 hash. These files are in the evidence package published with the paper (Section 14).

4.10 Analyses added after the results (D24)

After four internal reviews, a separate script computed post-registration robustness analyses into separate outputs; it first reproduced every registered count and verdict. They include three descriptive source analyses that the protocol registered in §5 and the frozen analysis omitted: (a) Perplexity's citations of the clinic page from Signal Note 04, per cycle; (b) new local or regional pages drawing at least 10% of one engine's answers in a cycle; (c) the source-class mix per engine and window, by a rule-based host classifier (5.4% of links unclassified). Sources are identified under one documented policy, applied in the same way across engines and windows and never according to the direction of a finding: major international sources (scientific publishers, medical databases, research organisations, international retailers and eligible hospitals) are named, and clinics, pharmacies and local businesses protected by the registered disclosure boundary are typed and given opaque tokens, with the full mapping kept in a restricted archive. Throughout, the paper distinguishes three claims about a source: that it was **cited** (it appears among an answer's citations), that it **supports** a claim (its content is shown by a contemporaneous archive, or, more weakly, its wording overlaps with the answer's by echo), and that it **changed** the answer (which no analysis here establishes). Overlap with a page was tested **by echo**: an answer echoes a page sentence when its normalised text shares a contiguous run of at least 40 characters with it. Shared wording between a captured answer and a later-fetched page corroborates textual overlap. Without a contemporaneous archive, it does not establish the page's complete historical content, the direction of copying or exclusive source provenance. Everything in

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

these analyses was computed after the results were seen, is labelled post-registration, and never replaces a registered verdict. They are exploratory: the contrasts (band readings, cycle-level estimates, per-cycle and question-group splits, source-mix comparisons) were chosen after the results, several were examined, and no adjustment for multiplicity was made, so their intervals and z-values are descriptive.

5. Primary findings (registered)

5.1 E2: naming movement against the bridged band (UAE)

Table 2. UAE, semaglutide (Wegovy): naming in the after window (4–11 October 2026; 7 cycles) against the bridged band. Registered primary endpoint.

Engine	Pre-congress v4 pooled (unbridged)	After window, K/N (exact rate)	Brand-only after (D24)	Bridged band	Distance to nearer edge (pp)	Clustered 95% (robustness)	Size (house scale)	Registered verdict	Attribution verdict (frozen E1 classification)
Perplexity	193/371 (52.0%)	243/378 (64.286%)	196/378 (51.9%)	41.98–64.36%	−0.074 (inside)	56.6–72.0%	+12.3 DECISION-RELEVANT	No movement (borderline)	Congress sources without movement
Gemini	416/701 (59.3%)	206/378 (54.497%)	199/378 (52.6%)	35.14–78.42%	−19.357 (inside)	45.5–63.8%	−0.5 SMALL	No movement	Congress sources without movement
ChatGPT	212/702 (30.2%)	110/378 (29.101%)	103/378 (27.2%)	14.23–46.87%	−14.871 (inside)	18.8–39.1%	+0.8 SMALL	No movement	Neither
Claude	336/702 (47.9%)	256/378 (67.725%)	254/378 (67.2%)	34.53–67.33%	**+0.395 (beyond)**	57.1–78.0%	+16.8 MAJOR	**Movement** (borderline)	**Movement without congress sources**
Four engines	1,273/2,794 (45.6%)	815/1,512 (53.902%)	752/1,512 (49.7%)	36.88–53.98%	−0.078 (inside)	47.2–60.9%	+9.1 POTENTIALLY RELEVANT	No movement (borderline)	Not applicable

Registered measure: brand or molecule. Distances signed (positive beyond the nearer edge). Size is measured against the bridged reference (v4 pooled plus the offset; I2), not the unbridged column. Borderline: the clustered interval contains the nearer edge (D19). Brand-only rates are post-registration and unbridged. Source: analysis output, E2 rows; robustness output, item 6.

Claude named semaglutide or Wegovy in 256 of 378 scored UAE answers (67.725%), 0.395 points beyond its bridged band (0.441 on a band rebuilt from exact single-cycle rates): **movement**, labelled MAJOR (+16.8 points). It named Wegovy in 254 of the 378 answers (67.2%; 45.4% in v4 cycles), so the rise is in brand naming, not molecule credit. None of Claude's 420 after-window UAE answers cited a congress-derived source, so the attribution test could not run, and the registered verdict is **movement without congress sources** (that is, without verified congress sources). **Perplexity** named it in 243 of 378 (64.286%), 0.074 points inside its upper edge (0.093 on the exact-cycle band): **no movement**, although its size label is DECISION-RELEVANT. Under the frozen E1 classification, one scored Perplexity answer qualified under E1 and named the medicine, as did 242 of its other 377 (Fisher p = 1); the minimum p any split could have reached was 0.357, so this verdict classifies a combination and is not evidence that congress sources leave naming unchanged. The same holds for Gemini under the frozen E1 classification (one qualifying answer, as did 205 of its other 377; minimum attainable p 0.455). The answer added by the E1 correction (Section 5.2) is in the governed attribution population and named the

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

medicine. Under the corrected classification, Gemini's UAE attribution table has two qualifying answers, both naming it, against 204 of 376 other scored answers (two-sided Fisher p = 0.503; minimum attainable p = 0.206): still structurally uninformative, and no verdict changes. The frozen calculation (one qualifying answer; minimum attainable p = 0.455) remains available as provenance. Both calculations can be reproduced from the public files: the per-answer table's `e1Qualifying` gives the frozen one, and the E1 correction record adds the corrected answer (Section 14). Gemini (206/378) and ChatGPT (110/378) sat well inside their bands; the four-engine rate (815/1,512) lay 0.078 points inside.

How robust are these verdicts? Table 2b and Figure 12 set the registered verdicts beside readings added after the results (D24), with each after-window cycle against the bridged band.

Table 2b. UAE verdicts under other readings of the band rule, and a cycle-level change estimate (all post-registration, D24, except the I1 column).

Engine	I1 (registered): verdict (distance, pp)	Shift by offset, widen by half-width	Widen by full width	Shift only (no widening)	Cycle-level change net of offset (95% CI; z)	After cycles outside bridged band (above / below)
Perplexity	No movement (−0.074)	**Movement** (+0.546)	No movement (−4.394)	Movement (+5.486)	+12.2 (−3.0 to +27.4; 1.57)	6 of 7 (5 / 1)
Gemini	No movement (−19.357)	No movement	No movement	No movement	−0.5 (−11.0 to +10.0; −0.10)	0 of 7
ChatGPT	No movement (−14.871)	No movement	No movement	No movement	+0.8 (−5.7 to +7.2; 0.23)	0 of 7
Claude	Movement (+0.395)	Movement (+0.085)	**No movement** (−8.865)	Movement (+9.035)	+16.8 (+7.2 to +26.4; 3.43)	3 of 7 (3 / 0)
Four engines	No movement (−0.078)	No movement (−0.078)	No movement (−3.628)	Movement (+3.472)	+9.1 (+3.9 to +14.3; 3.42)	4 of 7 (4 / 0)

Cycle-level change: mean after-window cycle rate minus mean v4 cycle rate minus the bridge point offset, cycles as units, normal approximation. Source: robustness output, items 2–4.

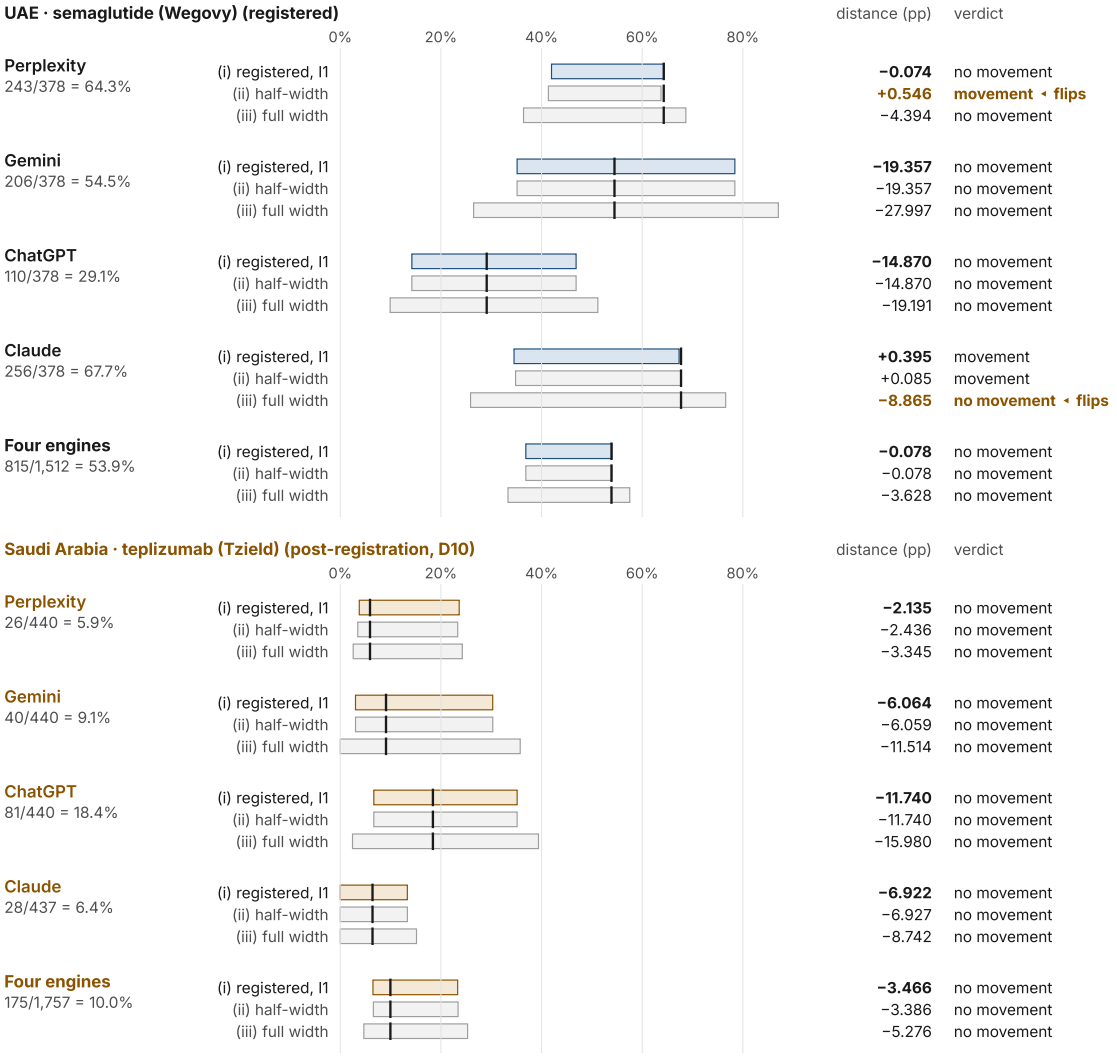
© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

Figure 12 E2 under three readings of the band rule, and a cycle-level change estimate

Post-registration robustness (D24). The registered verdicts are those of reading (i), I1, and stand; readings (ii) and (iii) never replace them.

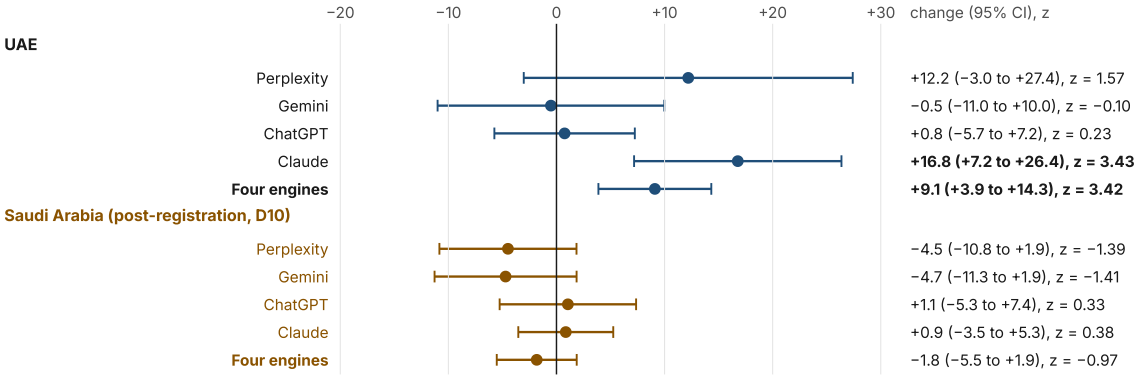
A After-window rate of answers naming the brand or its molecule against each reading's band

(i) band ends plus the offset CI bounds (registered reading I1, D16); (ii) shift by the point offset, then widen each end by half the CI width; (iii) shift by the point offset, then widen each end by the full CI width.



B Cycle-level change net of the offset, pp (a different estimand from the band rule)

Mean after-window cycle rate minus mean v4 reference cycle rate minus the bridge point offset; cycles are the unit. A different estimand from the band rule. Normal approximation; cycles treated as independent.



Panel A: thin vertical line = after-window pooled rate; shaded bar = the band under each reading; distance signed (positive = beyond the nearer edge); "◂ flips" marks a verdict that differs from the registered one. The two verdicts that depend on the reading: UAE Perplexity under (ii) half-width (+0.546 pp, movement); UAE Claude under (iii) full width (−8.865 pp, no movement). Reading I1 was fixed in D16 after per-cycle counts had been displayed. Bands below 0% are clipped at 0. Panel B asks whether the mean cycle level changed, not whether the pooled rate left a band; bold where the 95% interval excludes 0. A shift-only band, which omits the registered widening, is in the CSV for reference.

Source: AlphaCitation EASD 2026 study, post-registration robustness analyses (D24; easd-2026-robustness.json in the frozen record at repository commit 6d1b45d3); they never replace a registered verdict. Fields: item2BandReadings (readings, rows), item3CycleLevelChange (estimand, rows); reading (i) reconciled with easd-2026-analysis.json e2.

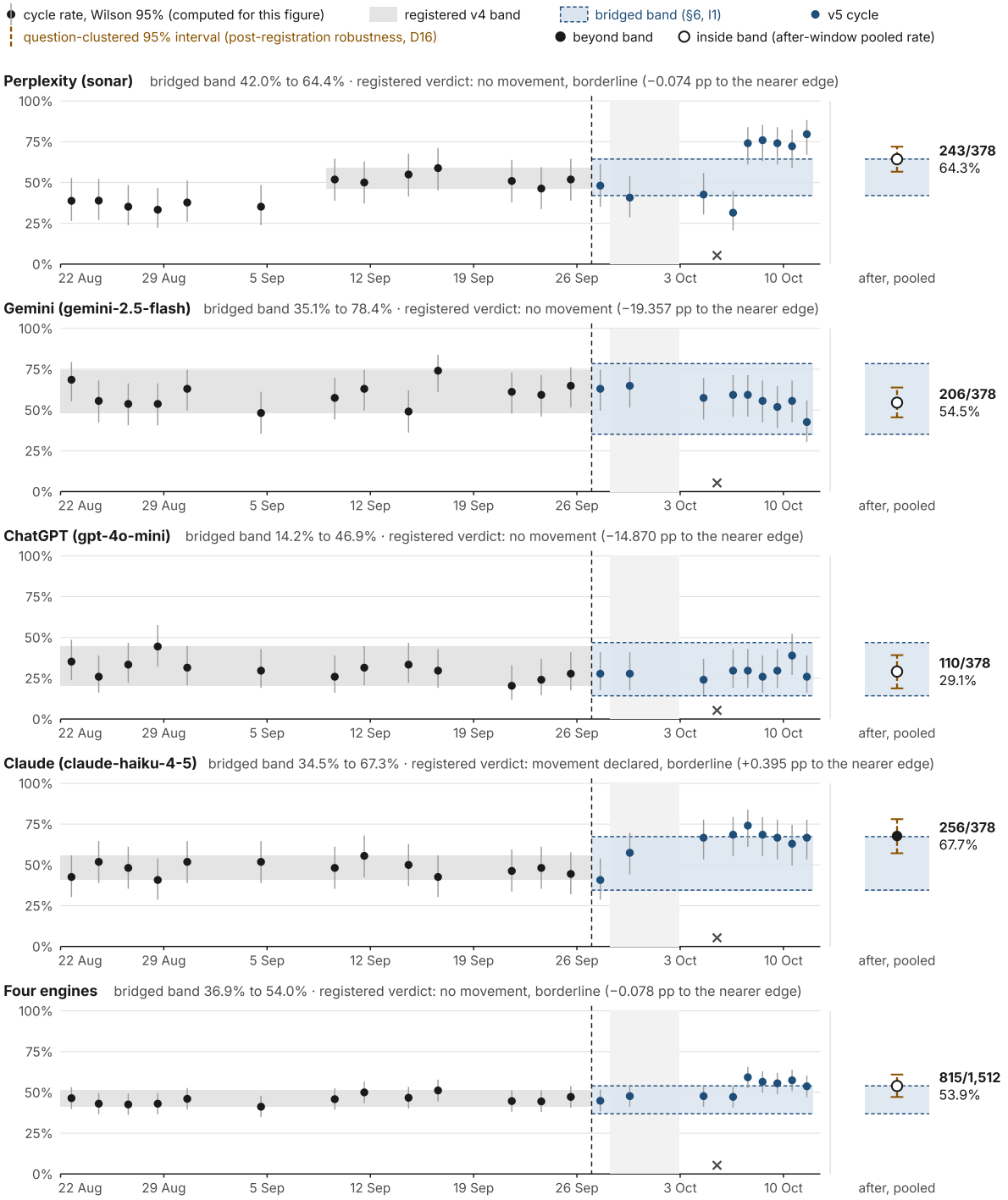
© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

Figure 12. E2 under three readings of the band rule, and the cycle-level change. Post-registration robustness (D24). Panel A: for each engine, the after-window pooled rate (registered measure: brand or molecule) against the bridged band under three readings of "widened by the offset's 95% interval": (i) the registered reading I1 (D16); (ii) shift by the point offset and widen each end by half the interval width; (iii) shift and widen by the full width; signed distances in percentage points, with verdicts that differ from the registered one marked. UAE is the registered endpoint; Saudi Arabia is the post-registration arm (D10). Panel B: the cycle-level change (mean after-window cycle rate minus mean v4 reference cycle rate minus the bridge point offset) with its 95% interval and z, a different estimand from the band rule. Period: after window, 4–11 October 2026, against the v4 reference cycles. Units: per cent; percentage points; z. N: UAE 378 scored answers per engine (1,512 combined); Saudi 437–440 per engine (1,757 combined). Source: robustness output, items 2 and 3.

Three conclusions follow. First, **the verdicts depend on the reading of the rule**: under the reading "shift by the offset, widen by its half-width", Perplexity becomes a movement; under "widen by the full width", Claude does not. The registered verdicts are those of I1, which was written after per-engine counts through 10 October had been displayed, and they stand. Second, **the pooled movement verdict is sensitive to the band-rule interpretation**. A post hoc cycle-level analysis estimates a Claude increase of 16.8 percentage points (95% interval 7.2 to 26.4; z = 3.43) under independent-cycle and bridge assumptions; it supports a descriptive increase within this design. Perplexity's estimate is imprecise (z = 1.57) because its after window holds two regimes. Third, **the pooled rule hides Perplexity's step**: Perplexity named the medicine in 23 of 54 answers on 4 October and 17 of 54 on 6 October (the second below its bridged band), then 40 of 54 on 7 October and 39 to 43 of 54 to 11 October, above it. Split at that step, chosen after seeing the data, the window holds 40 of 108 (37.0%) and 203 of 270 (75.2%). Claude's rise began earlier and on the new instrument: 22 of 54 on 27 September (v5), 31 of 54 on 29 September, then 34 to 40 of 54 in every after-window cycle. At the old v4 setting on 5 October, Claude's three bridge runs named the medicine in 33 to 37 of 54 answers (61.1–68.5%; 106 of 162 pooled), above its v4 band before any offset; the rise is not explained by the v4 → v5 setting change as measured that day.

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

Figure 4 **UAE: answers naming semaglutide or Wegovy, per engine and cycle**



Shaded vertical strip: During window (28 Sep 06:00 – 2 Oct 22:00 UTC). Dashed vertical line: instrument change v4 → v5 (27 Sep). Dark dots: v4 cycles; blue dots: v5 cycles; each dot sits at its cycle's scan time. Cross: the excluded cycle (5 Oct, D14). Perplexity's registered band is its post-step half (cycles from 9 Sep, Note 04).

Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-analysis.json (perCycleNaming: governed counts per engine and cycle, validated against the production scorer; e2: band, bridgedBand, after, clusterCI95, precision). Every pooled window drawn reproduces the analysis exactly.

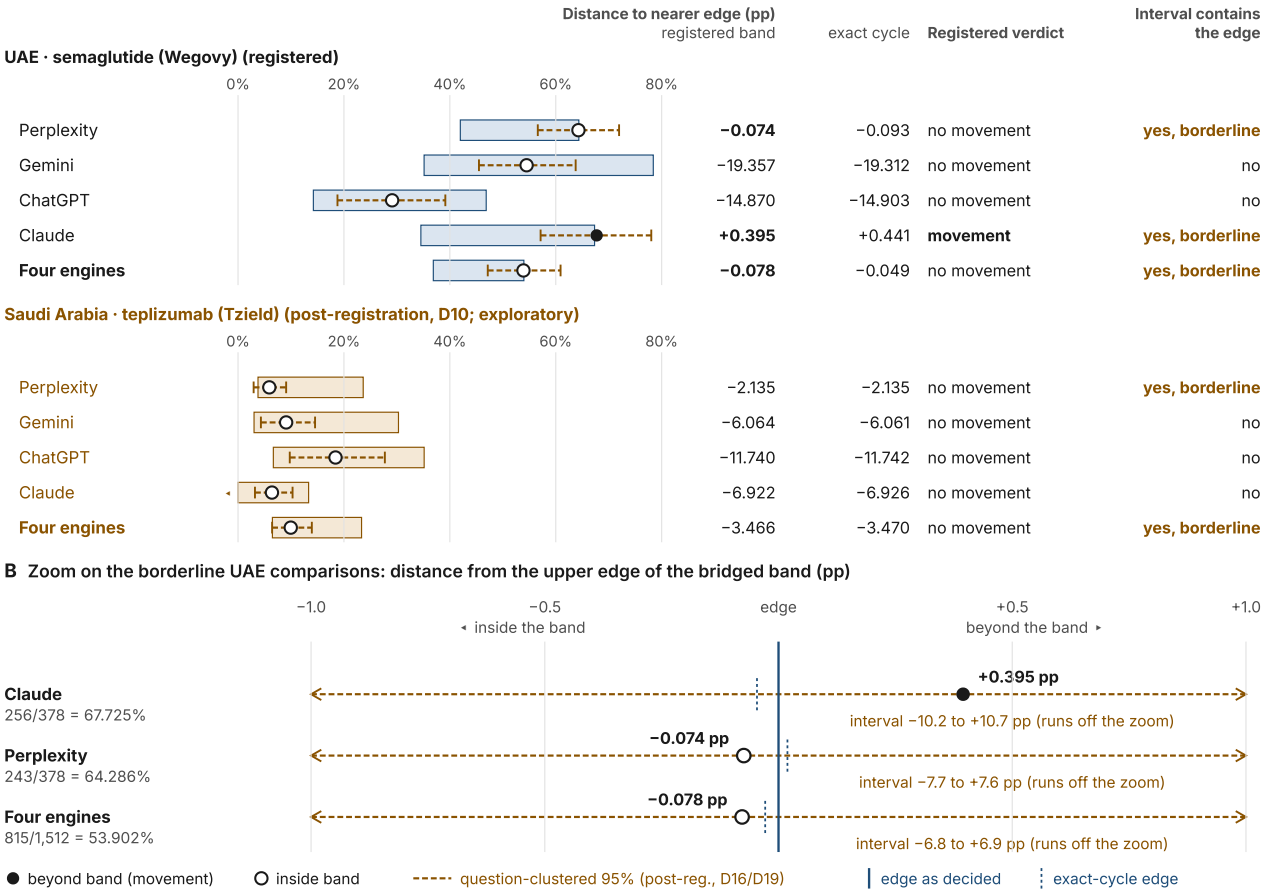
Figure 4. UAE naming trajectories. Share of scored UAE answers naming semaglutide (Wegovy) (brand or molecule; E2, registered primary) for each engine and the four engines combined, in every admitted cycle (13 v4; 2 v5 before the after window; 7 after). Error bars are single-cycle Wilson 95% intervals computed for this figure [92]. Single-cycle Wilson intervals are illustrative binomial intervals. Their interpretation assumes an appropriate independent Bernoulli sampling model, which this fixed prompt battery does not establish. Repeated-question dependence additionally affects pooled and longitudinal

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

comparisons. These intervals are not used to validate the registered movement decision. Grey band: registered v4 single-cycle range. Tinted band: bridged band. Right column: after-window pooled rate (K/N) with its question-clustered 95% interval (post-registration robustness; never a replacement for the registered rule); filled marker = movement declared. Period 22 August – 11 October 2026. Units: per cent of scored answers. N: 49–54 scored answers per engine per cycle; after window 378 per engine, 1,512 combined. Source: analysis output.

Figure 5 **E2: answers naming the brand or its molecule, after window, against the bridged band**

A After-window pooled rate, bridged band and clustered interval



Signed distance: positive = beyond the nearer edge of the bridged band, negative = inside it. "Registered band" uses the band as decided (§4 limits quoted to 0.1%, offset bounds as stored); "exact cycle" rebuilds it from the exact single-cycle rates with exact offset bounds (D19 check). The registered verdict is never reversed by either check. Saudi Claude's bridged band starts below 0% (−1.2%), clipped at 0.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-analysis.json (e2[].after, bridgedBand, movementDeclared, precision.*).

Figure 5. E2 band distances at full precision. Panel A: for each engine and the four engines combined, the after-window pooled rate (filled = movement declared), the bridged band and the question-clustered 95% interval (post-registration robustness); right, the signed distance to the nearer edge against the band as decided and against the band rebuilt from exact single-cycle rates (D19), the registered verdict, and whether the clustered interval contains the edge. UAE rows registered; Saudi rows post-registration (D10). Panel B: the three borderline UAE comparisons on a ±1-point scale. Period: after window, 4–11 October 2026. Units: per cent; percentage points. N: 378 per UAE engine (1,512 combined); Saudi 437–440 per engine (1,757 combined); 54 (UAE) or 55 (Saudi) question clusters. The clustered intervals extend at least 7.6 points either side of the rate for Claude and Perplexity, and about 7 points for the four-engine total. Source: analysis output.

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

The registered bridge self-check did not hold, and was ill-posed as registered. In every UAE and Saudi engine the same-day v4-setting spread covered 10–50% of the band's width (Table 2a): raw three-run same-day ranges were narrower than the multi-cycle ranges. But a range of three draws is expected to be narrower than a range of 7 to 13 draws even with no day effect, so the original range comparison alone does not separate within-day and between-day variation. After a post hoc adjustment for the different numbers of runs (D24), four of eight UAE/Saudi engine–market combinations (UAE Gemini and ChatGPT, Saudi Perplexity and Gemini) exceeded the chosen between-day-variation threshold (an adjusted ratio below 0.5, chosen after the ratios were seen); UAE Claude did not (adjusted ratio 0.99), and UAE Perplexity showed little excess (0.71). At the v5 setting, same-day spreads ran up to 12.66 points (Saudi Gemini), not 7.41.

Table 2a. Bridge self-check: registered comparison (D23) and adjusted comparison (D24).

Series	Engine	v4-setting runs	v5-setting runs	Same-day spread v4 / v5 (pp)	Band width (pp)	Registered ratio (v4 spread / width)	Adjusted ratio, v4 runs (≈ 1: no excess between-day variation)	SD between cycles / within day, pooled (pp)
UAE	Perplexity	19, 16, 16 of 54	18, 15, 18	5.56 / 5.56	12.5	0.444	0.708	3.93 / 3.21
UAE	Gemini	31, 32, 30 of 54	31, 26, 29	3.70 / 9.26	26	0.142	0.281	7.49 / 3.55
UAE	ChatGPT	17, 16, 15 of 54	16, 15, 14	3.70 / 3.70	24	0.154	0.303	5.97 / 1.85
UAE	Claude	37, 36, 33 of 54	36, 38, 37	7.41 / 3.70	14.9	0.497	0.985	4.40 / 3.02
Saudi Arabia (post-reg.)	Perplexity	0, 1, 0 of 55	0, 1, 1	1.82 / 1.82	18.04	0.101	0.189	5.59 / 1.05
Saudi Arabia (post-reg.)	Gemini	5, 3, 3 of 55	9, 2, 8 (of 55, 54, 55)	3.64 / 12.66	16.36	0.222	0.416	5.17 / 5.06
Saudi Arabia (post-reg.)	ChatGPT	11, 9, 8 of 55	10, 7, 11	5.45 / 7.27	20	0.273	0.511	6.52 / 3.32
Saudi Arabia (post-reg.)	Claude	0, 1, 2 of 55	1, 2, 1	3.64 / 1.82	10.91	0.333	0.625	4.23 / 1.48

Registered ratio: exact same-day v4 range over the registered band width (no threshold registered). Adjusted ratio: each range divided by the expected range of that many independent draws (d2 constants); computed after the results. Within-day SDs rest on 2 to 4 degrees of freedom and are imprecise. Singapore is at ceiling. Source: analysis output, bridge self-check; robustness output, item 5.

5.2 E1: congress penetration and latency

No qualifying citation or result text was verified by the implemented inspection procedure for any registered readout in its own series, battery and window. All 28 registered engine-by-readout cells are not detected within follow-up: not detected by day 10.0 (ACHIEVE-4 LBA 25; LBA 29), 10.3 (ACHIEVE programme; SMARTEST), 11.1 (appetite abstract 122), 11.3 (TRIUMPH-2) and 12.1 (REIMAGINE 1). On the text route, every during- and after-window answer was screened for congress and readout terms; the 30 candidates were read and none reported the result of a registered readout. Two Gemini UAE answers

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

(6 and 7 October) stated the OCTANE result (post-registration, D3) without naming the trial; the term screen cannot see an unnamed trial, and they were found by the E4 hand check (Section 7.4). This is a finding about what the implemented procedure verified, not about whether the readouts reached the engines.

Congress material reached few answers, through a narrow channel. In the registered arms, seven sources qualified (two sponsor releases, five news or medical-media articles). In the registered arms, six of 2,408 after-window answers qualified through cited congress-derived sources, all sponsor releases or media. A subsequent check identified one additional UAE Gemini answer reporting the OCTANE result without a qualifying source, giving seven of 2,408 answers under the endpoint's either-route definition. The frozen tally and corrected tally are shown separately (Table 2c). In the Saudi arm, three sources were cited in 5 of 1,789 answers. No answer cited a congress page (class a) or a journal publication of a readout trial (class b).

Series	Perplexity	Gemini	ChatGPT	Claude	All
UAE (registered)	5/420	1/420	0/420	0/420	6/1,680
Singapore (registered)	0/182	0/182	0/182	0/182	0/728
Saudi Arabia (post-registration)	3/448	1/448	0/448	1/445	5/1,789

Frozen tally: after-window answers citing a congress-derived source or reporting a readout result as detected by the registered term screen. No before-window answer qualified. Source: E1 output.

Table 2c. E1 qualifying answers after the congress: frozen and corrected tallies.

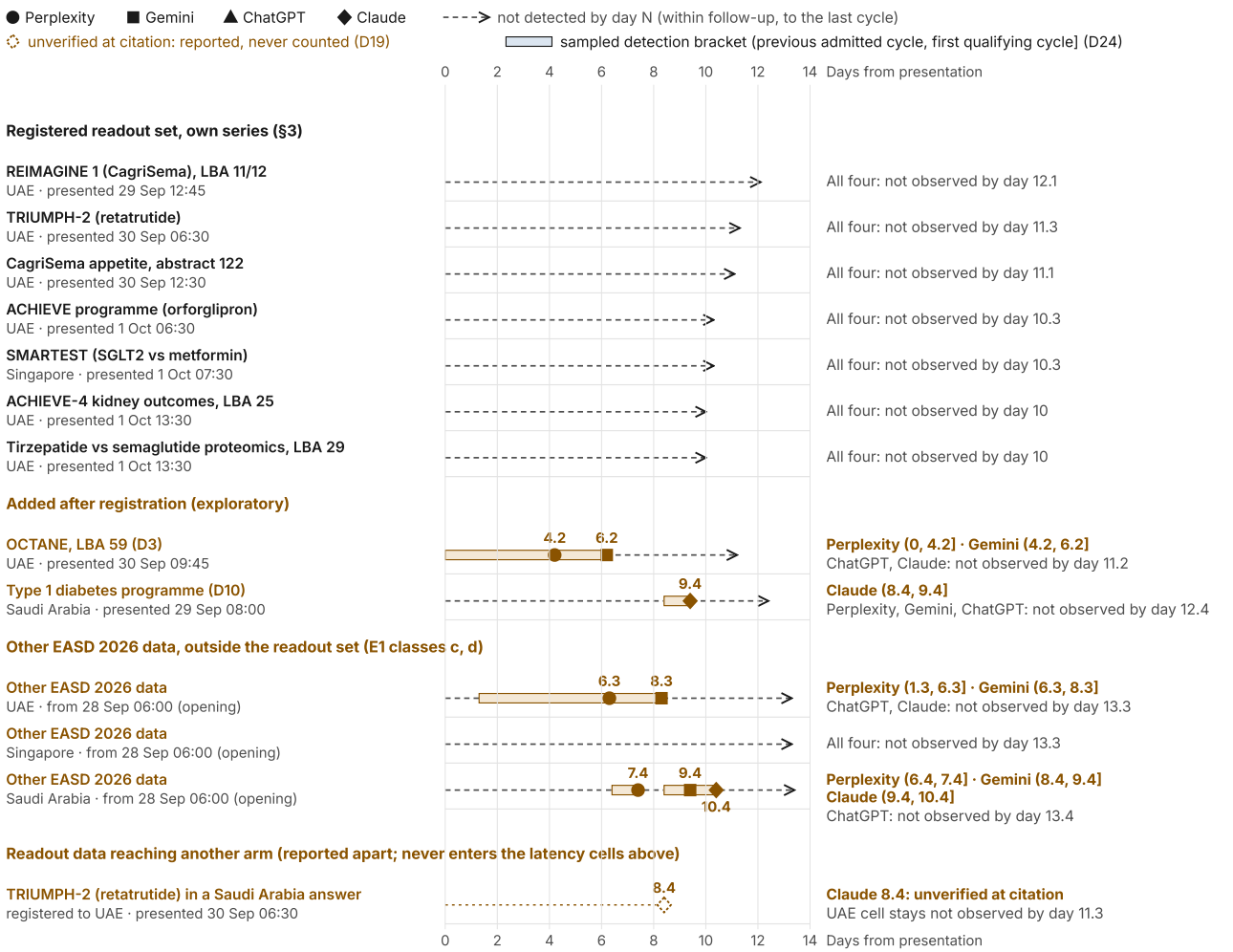
Count	Frozen tally (E1 output)	Corrected tally, either route
Registered arms (UAE and Singapore)	6 / 2,408 (0.249%)	7 / 2,408 (0.291%)
UAE Gemini	1 / 420	2 / 420
Saudi Arabia (post-registration)	5 / 1,789	5 / 1,789 (unchanged)
All arms, registered and post-registration (descriptive)	11 / 4,197 (0.262%)	12 / 4,197 (0.286%)

The corrected tally adds one text-only UAE Gemini answer of 7 October that reports the OCTANE result without a qualifying source; the E4 hand check found it, and it was identified as an E1 miss after the freeze (D24). OCTANE was added after registration (D3), so the added answer concerns a post-registration readout and leaves every registered readout cell unchanged; Gemini's OCTANE latency is unchanged. Frozen values are not altered. Under the corrected classification, Gemini's UAE attribution test gives two-sided Fisher p = 0.503 (minimum attainable 0.206), against p = 1.000 (minimum attainable 0.455) under the frozen E1 classification; no verdict changes (Section 5.1). Source: E1 output; robustness output, item 11(b).

In Perplexity's first after-window UAE cycle (4 October), four sources were first observed together: a syndicated copy of the sponsor's OCTANE release, a trade-press article carrying OCTANE and other EASD data, a syndicated sponsor release of an indirect treatment comparison, and a news report of other EASD data. Of the ten sources, four carried OCTANE and two reported indirect treatment comparisons. Four of Perplexity's five qualifying UAE answers came from focused questions (42 of its 420 after-window answers), so they sit outside the registered attribution test; penetration depends on whether the battery asks questions that send engines to news.

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

Figure 7 **E1: days from presentation to the first qualifying answer, by readout and engine**



Bars show sampled detection brackets (previous admitted cycle, first qualifying cycle]; arrows: no detection within follow-up; neither is uptake time. Each is drawn as a bar and listed on the right (post-registration correction, D24 item 11a); the marker sits at the first positive cycle. The first qualifying cycle is the first admitted cycle with at least one qualifying answer (a cited congress-derived source, or answer text reporting a readout result); a cell with none is not detected within follow-up, to the completion of the series' last admitted cycle (UAE 11 Oct, 14:05 UTC; Singapore 11 Oct, 14:02 UTC; Saudi Arabia 11 Oct, 16:34 UTC), measured from presentation. A readout counts only in its registered series. ChatGPT has no congress-linked answer in any arm. Gemini also cited a PR Newswire copy of the OCTANE release that could not be read: it is unverified, not counted and not drawn; its first citation (5 Oct 14:02 UTC) fell in the excluded 5 Oct cycle (D14), so no latency can rest on it.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-e1.json (latency, readouts, crossArm); easd-2026-e1-ledger.json (adjudication of the cross-arm source); latency intervals from easd-2026-robustness.json item11E1Corrections.a_latencyIntervals (D24).

Figure 7. E1 penetration and latency. Days from the specified presentation or programme anchor to first qualifying detection in an admitted sampled cycle: an engine's answer citing a congress-derived source carrying that readout or reporting its result (E1, registered primary). Each detection is drawn at the first positive sampled cycle; its sampled detection bracket runs from the last negative sampled cycle; dashed arrows mark cells not detected within follow-up (N measured to the series' last admitted cycle). The published figure's labels were corrected for display ("sampled detection bracket"; "not detected by day N within follow-up"); the generated source graphic, which reads "not observed by day N", remains as provenance. Groups: the seven registered readouts in their own series; readouts added after registration (OCTANE, D3; the Saudi type 1 programme, D10), exploratory; other EASD 2026 data (classes c and d). Bottom row: the cross-arm entry, TRIUMPH-2 carried by a media round-up Claude cited in a Saudi answer in (7.4, 8.4] days, unverified at citation (D18/D19), never counted. Period: presentation (28 September – 1 October 2026) to the last admitted cycle (11 October). Units: days. N: 48 engine × readout cells (8 with a detection, 40 not detected within follow-up) plus 1 cross-arm entry. The previous negative and first positive cycles bracket sampled detections, not the engine's true evidence-

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

uptake time. Detection is imperfect and may be non-persistent. "Not detected by day N" refers to this battery and observation schedule. Programme-anchored exploratory rows do not give readout-specific uptake times. Source: E1 output; E1 ledger; robustness output, item 11.

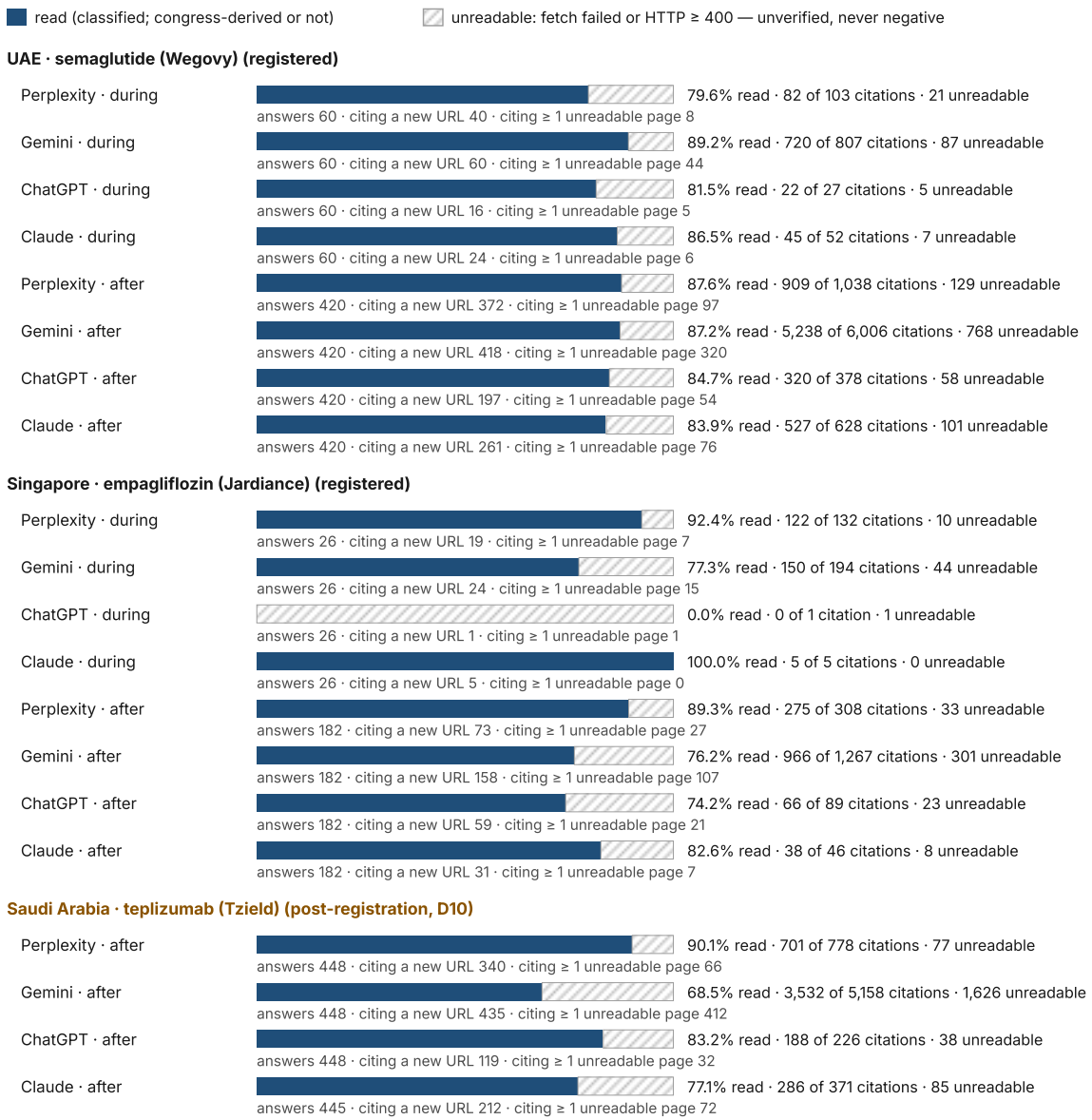
Readout (status)	Series	Perplexity	Gemini	ChatGPT	Claude
OCTANE, LBA 59 (D3)	UAE	(0, 4.2]	(4.2, 6.2]	not detected by day 11.2	not detected by day 11.2
Other EASD 2026 data	UAE	(1.3, 6.3]	(6.3, 8.3]	not detected by day 13.3	not detected by day 13.3
Other EASD 2026 data	Singapore	not detected by day 13.3	not detected by day 13.3	not detected by day 13.3	not detected by day 13.3
Saudi type 1 programme (D10)	Saudi Arabia	not detected by day 12.4	not detected by day 12.4	not detected by day 12.4	(8.4, 9.4]
Other EASD 2026 data	Saudi Arabia	(6.4, 7.4]	(8.4, 9.4]	not detected by day 13.4	(9.4, 10.4]

Sampled detection brackets in days: (last negative sampled cycle, first positive sampled cycle]. All post-registration readouts; exploratory. Origins: OCTANE 30 September 09:45 UTC; Saudi programme 29 September 08:00 UTC; other data 28 September 06:00 UTC. Source: robustness output, item 11(a); E1 output.

No UAE cycle was admitted between 29 September and 4 October, and the 5 October cycle was excluded. The last negative and first positive sampled cycles define observation brackets of (0, 4.2] days for Perplexity and (4.2, 6.2] days for Gemini. These brackets do not establish the engines' true evidence-uptake times, and the observation schedule and imperfect detection do not permit ranking the engines by uptake speed. Publication-to-citation delay was sometimes short: two of the articles Gemini first cited on 6 October were dated 5 October. Gemini also cited a newswire copy of the OCTANE release that could not be read; it first appears in an admitted cycle on 6 October and changes no latency. ChatGPT cited no congress-derived source in any arm.

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

Figure 6 **E1 source route: new-URL citations read and unreadable, by series, engine and window**



Detection limit: only URLs not cited by the same series before the congress are read; a page cited earlier and later updated with congress content is not re-read (the text route, which read every answer, still detects an answer reporting a readout result). Screen: 5,268 new (series, URL) pairs, 947 unreadable, 114 candidates read by hand. Per-domain counts of unreadable pages are in the source file and are not reproduced here, because they include clinics, pharmacies and local businesses, which the publication does not name (§10, D22). Rows with few citations (e.g. Singapore ChatGPT, during: 0 of 1) carry little information.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-e1.json (inspectionCoverage.bySeriesEngineWindow; screen).

Figure 6. E1 inspection coverage. Coverage of the E1 source route by series, engine and window (Saudi rows post-registration, D10). Each bar is the set of new-URL citations (URL occurrences in answers, for URLs not cited by the same series before the congress): solid = fetched and classified; hatched = unreadable (fetch failure or HTTP status 400 or above), unverified and never negative. Text gives the share read, the counts and the answers citing at least one new or unreadable page. Saudi Arabia has no during-window cycle. Period: 29 September – 11 October 2026. Units: citations; answers. N: 5,268 new series–URL pairs. Source: E1 output.

After the congress, the share of new-URL citations read ranged from 68.5% (Saudi Gemini, 3,532 of 5,158) to 90.1% (Saudi Perplexity, 701 of 778); in the congress window, from 0.0% (Singapore ChatGPT, one citation) to 100.0% (Singapore Claude, five). Of 5,268 new pairs, 947 could not be read, 619 of them returning HTTP 403, which suggests bot protection reacting to the study's scripted fetcher rather than pages unavailable to the engines. In Saudi Arabia the five largest unreadable domains (1,175 citations)

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

were retail, pharmacy, e-commerce or delivery pages; in Singapore one drug-information site accounted for 305; in the UAE they included several news sites, which are class (d) candidates. In the UAE after window, 319 of Gemini's 420 answers cited at least one unreadable page and did not otherwise qualify. Excluding answers that cite unreadable pages from the attribution comparison, rather than counting them as negative, changes no reading (D24).

What was new at EASD. For SMARTEST, first disclosed at EASD, no arrival was verified by day 10.3. For TRIUMPH-2, whose topline was public from July, the full data were not verified in answers by day 11.3; a headline figure stated without a congress-dated source would not count, because topline sources dated before 28 September fail the date rule for classes (c) and (d). For ACHIEVE-4, the registered kidney-outcome analysis (LBA 25), for which no earlier disclosure was found, was not verified by day 10.0; the cardiovascular topline of April is a different result. For re-presented readouts (REIMAGINE 1, most of ACHIEVE), E1 reads only pages newly cited after the congress, so a readout already present in pre-congress citations would not register; for these, E1 measures new arrivals, not presence. The registered class (b) has no date condition, so a new citation of REIMAGINE 1's June paper would have counted as congress-derived; no readable new URL in any window was a readout-trial publication (D21), but a positive class (b) count could not by itself distinguish new from re-presented evidence.

5.3 One cited page, two items

The diaTribe round-up "Top Diabetes News: EASD 2026" (dated 5 October) [93] shows why citation, representation and attribution are separate (Figure 8).

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

Figure 8 **Citation is not representation: one cited page, two items, two evidential statuses**

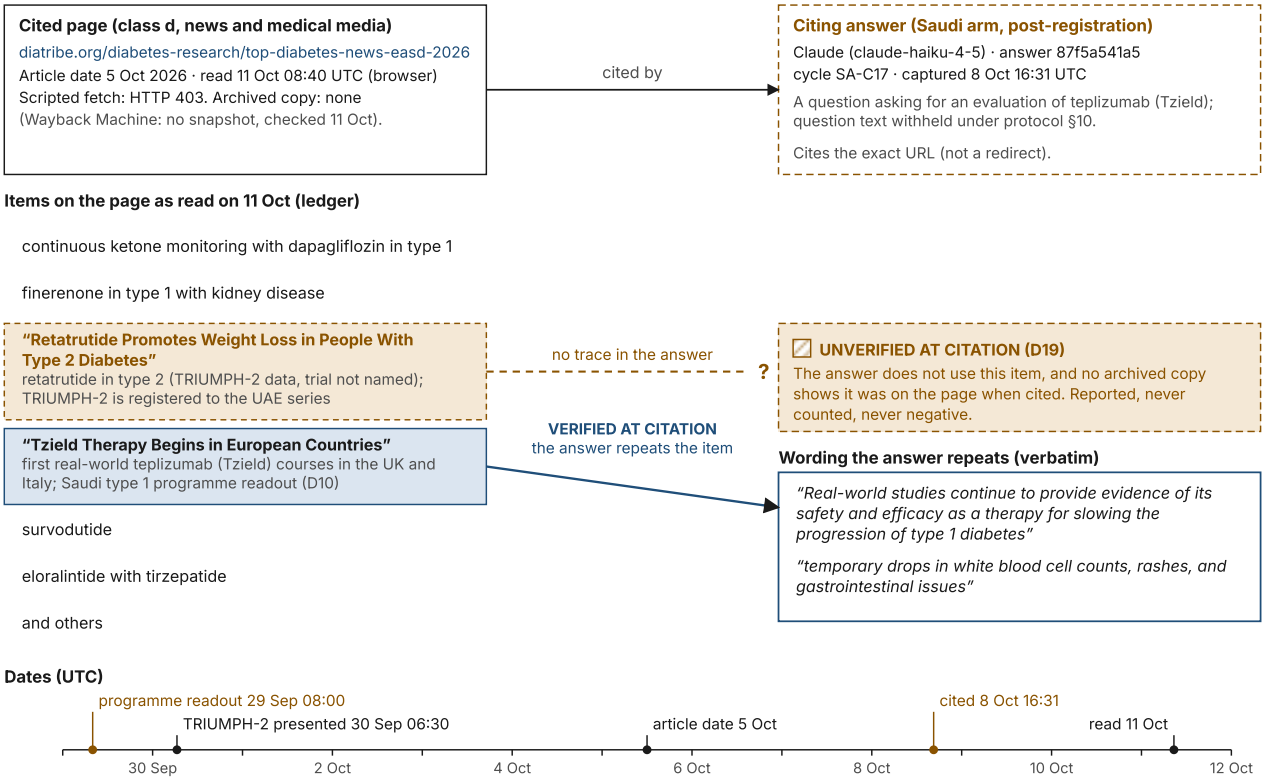


Figure 8. Source-to-claim illustration: the diaTribe case. One qualifying source traced from page to answer. Left: the round-up (class d media, dated 5 October) and its items as read on 11 October. Right: the Saudi-arm answer by Claude that cited the exact URL (8 October 16:31 UTC; the question asked for an evaluation of teplizumab (Tzield), its text withheld). The teplizumab real-world item is verified at citation because the answer repeats its wording. The retatrutide item, reporting TRIUMPH-2 data without naming the trial, is unverified at citation: the answer does not use it and no archived copy exists; it is reported and never counted (D19), and it belongs to the UAE series, where no answer cited the page. Saudi arm: post-registration (D10). Period: 29 September – 11 October 2026. Units: dates (UTC); days. N: one source, one answer. Source: E1 ledger; E1 output.

Claude cited the page in one Saudi answer on 8 October. The item on the first real-world teplizumab courses in Europe is **verified at citation** in the D19 sense, by textual overlap: the answer repeats its wording, including "temporary drops in white blood cell counts, rashes, and gastrointestinal issues". No archived copy exists, so the overlap does not establish the page's complete historical content or the direction of copying. No individual case safety information was identified; the adverse-event terms quoted come from a published media summary as repeated by the answer. The item reporting TRIUMPH-2 results without naming the trial is **unverified at citation**: no archived copy exists and the answer does not use it. The case counts toward the Saudi programme readout (Claude, (8.4, 9.4] days) and other EASD 2026 data; TRIUMPH-2 appears only as a cross-arm entry, in (7.4, 8.4] days, never in

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

the registered UAE cell, which remains "not detected by day 11.3". The TRIUMPH-2 item was found by an external check, because the term screen cannot see an unnamed trial; the earlier label is kept on the record (D18).

---

6. Tracing the movements (post-registration, D24)

The registered analysis stops at "movement without congress sources". The protocol's own §5 source analyses, reported late under D24, go further. Everything in this section is a descriptive association found after the results were seen; none of it is a causal estimate or replaces a registered verdict.

6.1 Claude's rise and the answers citing one local page

The registered §5 search for new local or regional pages drawing at least 10% of one engine's answers found one page that dominates Claude's UAE citations. It is a blog post by a local commercial company; under the disclosure rules it is typed, not named. A page displaying an August 2026 date was fetched on 11 October and first appeared in Claude's resolved study citations on 29 September. Historical first publication and the complete page content at each citation date were not independently archived. No pre-congress cycle shows it in a ChatGPT, Claude or Perplexity answer, including the 27 September v5 cycle (Gemini's pre-congress links were not resolved). It first appeared in 12 of Claude's 60 answers on 29 September, the cycle in which Claude's naming first rose, and then in 19 to 24 of 60 in every after-window cycle (152 of 420 after the congress). No other engine cited it among the resolved citations examined; pre-congress Gemini destinations were unresolved.

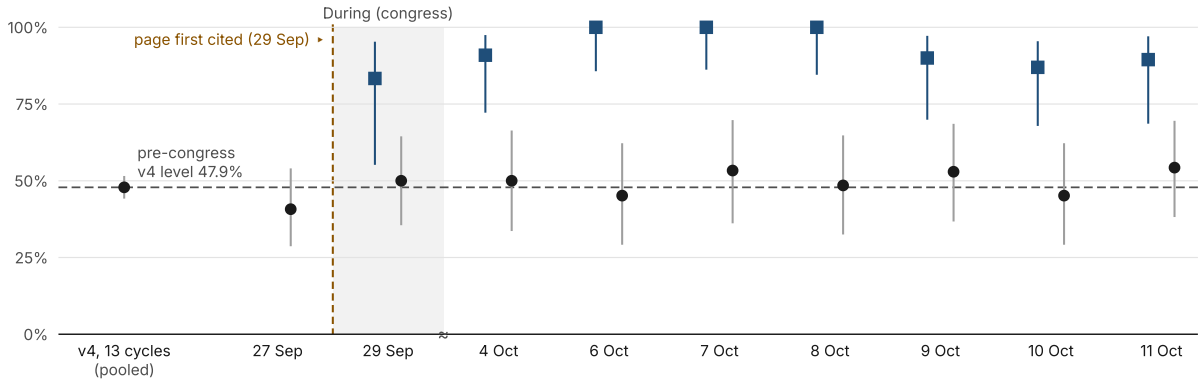
© 2026 AlphaCitation (Akunudo LLC-FZ) · CC BY 4.0

Figure 11 **Claude, UAE: answers naming semaglutide or Wegovy, by whether they cite one page**

Registered §5 descriptive analysis, reported late (D24); descriptive association, not a causal estimate

The page (P61): a local commercial blog post (UAE), August 2026, before the congress. It is not named, and neither are its domain or sentences (§10, D22). Claude first cited it on 29 Sep; no other engine cited it.

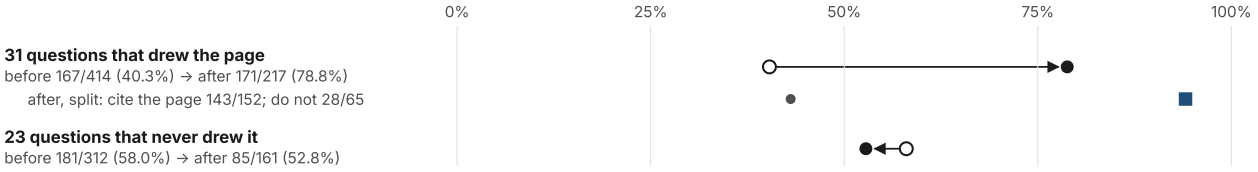
A **Per cycle: answers that cite the page and answers that do not (Wilson 95%)**



	v4, 13 cycles (pooled)	27 Sep	29 Sep	4 Oct	6 Oct	7 Oct	8 Oct	9 Oct	10 Oct	11 Oct
Cite the page	—	—	10/12	20/22	23/23	24/24	21/21	18/20	20/23	17/19
Do not cite it	336/702	22/54	21/42	16/32	14/31	16/30	16/33	18/34	14/31	19/35
Share citing	0/702	0/54	12/54	22/54	23/54	24/54	21/54	20/54	23/54	19/54

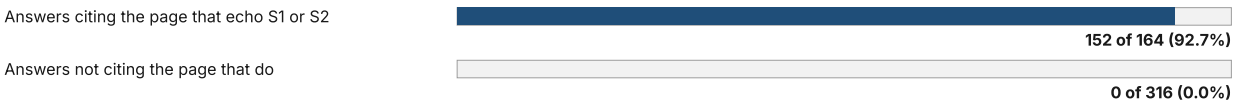
■ answers citing the page ● answers not citing it - - - Claude's registered v4 reference, 336/702 ≈ no admitted cycle 30 Sep – 3 Oct

B **By question: questions that drew the page in the after window against questions that never did**



Excess over each question's own pre-congress rate (v4 cycles and 27 Sep): +21.4 pp, 80.8 namings across 378 after-window answers. Answers citing the page carry +83.8 namings; answers not citing it −3.1 (+2.4 in questions that drew the page, −5.5 in the others).

C **Content at citation: Claude's answers repeating the page's own sentences (29 Sep – 11 Oct)**



Echo: a shared run of at least 40 characters after normalisation, against two of the page's sentences, identified only by kind: S1, a class statement; S2, an efficacy-figure sentence. Before 29 Sep: 1 echo (16 Sep, S1, a 46-character run, not citing the page). Other engines: 0 echoes in any window.

Registered §5 descriptive analysis, reported late (D24); descriptive association, not a causal estimate. Columns are cycles, evenly spaced (not to time scale). Wilson 95% intervals assume independent answers; answers to the same question are correlated, so the intervals are too narrow. The pattern locates the rise in the answers that cite the page; it does not show that the page caused it: the page may be retrieved because a question already leads toward the medicine, and a model or retrieval change on or before 29 Sep would produce the same pattern.
Source: AlphaCitation EASD 2026 study, post-registration robustness analyses (D24; easd-2026-robustness.json in the frozen record at repository commit 6d1b45d3); they never replace a registered verdict. Fields: item12PageP61 (page, a_citationsByEngine, b_contentAtCitationByEcho, c_whereClaudesRiseSits); reconciled cycle by cycle with easd-2026-analysis.json perCycleNaming.

Figure 11. Claude's rise and the answers citing one local page (UAE). Registered §5 descriptive analysis, reported late (D24); descriptive association, not a causal estimate. Panel A: for Claude (UAE), the share of scored answers naming semaglutide or Wegovy (the registered E2 measure) in each cycle from 27 Sep, split into answers that cite one new page (P61, a local commercial blog post (UAE), August 2026, before the congress) and answers that do not, with K/N and Wilson 95% intervals; the v4 cycles are pooled (336/702, Claude's registered v4 reference), drawn as the dashed reference level; the congress window and the page's first citation (29 Sep) are marked. Panel B: questions that drew the page in the after window against questions that never did, before (v4 and 27 Sep) and after, with the after-window split by citation and the excess over each question's own pre-congress rate. The before-window partition uses matched prompt texts and excludes 30 answers to three before-only texts, including 10 naming answers; its totals are therefore 348/726 rather than the full-panel 358/756. Panel C: the echo test of the page's content at citation. The page is never named. Single-cycle Wilson intervals are

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

illustrative binomial intervals and assume an independent Bernoulli sampling model, which this battery does not establish. Units: per cent of scored answers; namings. N: 378 after-window scored answers (152 citing the page). Period: 22 August – 11 October 2026 (per cycle from 27 September). Source: robustness output, items 1(b) and 12.

Three descriptive observations connect Claude's rise with this page.

- **Citation and naming.** After the congress, Claude's scored answers that cited the page named the medicine in 143 of 152 (94.1%); those that did not cite it, in 113 of 226 (50.0%), close to its pre-congress level (47.9%). In every after-window cycle the split repeats: 87.0% to 100.0% in citing answers, 45.2% to 54.3% in the rest.
- **Representation, by echo.** Two distinctive sentences on the page, a class statement and an efficacy-figure sentence, were tested by echo; neither is reproduced here, and the page's other sentences were not tested. From 29 September, 152 of the 164 answers citing the page repeat at least one of them (a shared run of at least 40 characters); no answer that does not cite it does. Before 29 September a single 46-character run at the opening of one sentence appears once, in an answer not citing the page. As with the diaTribe item, the shared wording corroborates textual overlap between those sentences and the answers. Without a contemporaneous archive, it does not establish the page's complete historical content or the direction of copying. Shared wording supports an association between the page and the answers; it does not establish exclusive source provenance or causation.
- **Where the extra naming sits.** The 31 questions that drew the page after the congress named the medicine in 40.3% of their pre-congress answers and 78.8% after: 94.1% when the page was cited and 43.1% when it was not. The 23 questions that never drew it went from 58.0% to 52.8%. Against each question's own pre-congress rate, Claude's after-window answers carried 80.8 extra namings (+21.4 points, unbridged); answers citing the page account for 83.8 of them, and answers not citing it for −3.1.

Claude's citation-presence rate was unchanged at 51/54 scored answers in each of the two cycles (27 and 29 September). The study did not record issued queries or uncited retrieval results, so it cannot establish unchanged search behaviour. What changed between those cycles is that this page entered the citations of Claude's answers. The page's efficacy-figure sentence, which attributes one figure jointly to two molecules, accounts for 29 of the 30 such joint-class efficacy figures (D4 reason P3) that the E4 hand check excluded (Section 7.4; robustness output, item 12(e)).

Alternative readings remain. The page may be retrieved because a question already leads toward the medicine (the questions that drew it started lower, not higher); naming among citing answers is high partly because the page itself discusses the medicine; and a model or retrieval change on or before 29 September would produce the same pattern. What the evidence does show is narrower: the measured rise sits in the answers that cite this page, and none of Claude's UAE answers cited congress material. This is the pattern Signal Note 04 recorded before the congress, when a clinic's comparison page entered Perplexity's citations in the 4 September cycle (8 of 60 answers) and its naming stepped [A1].

6.2 Perplexity's step

Perplexity's step on 7 October does not sit in answers citing any single page. Between 6 and 7 October, 24 of its 54 paired scored answers turned to naming the medicine (17 the brand, 7 the molecule only) and one turned away. On the same day its whole source mix changed (Figure 13): the share of cited links classified as "Other" rose by 40.3 percentage points (many prominent newly cited sources were market-research sites), consumer and retail pages fell by 17.4 points and news by 10.0, and its distinct cited domains fell from 199 to 115. The new page most frequent among the answers that turned to

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

naming appeared in only 10 of the 24. Market-research pages were cited by 85.2% of all 7 October answers, about as often as by the turned answers, and after the step naming was higher in answers that did not cite them (96.7%) than in those that did (72.5%). From the pre-step to the post-step cycles, naming rose 38.1 points: 28.5 points of brand naming and 9.6 of naming the molecule alone, so about a quarter of the step is the molecule name. The Note 04 clinic page appeared in 1 to 7 of Perplexity's 60 answers in each after-window cycle, with no step on 7 October. A whole-mix change of this kind is consistent with a change on the engine's side, but the study cannot separate it from the sources.

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

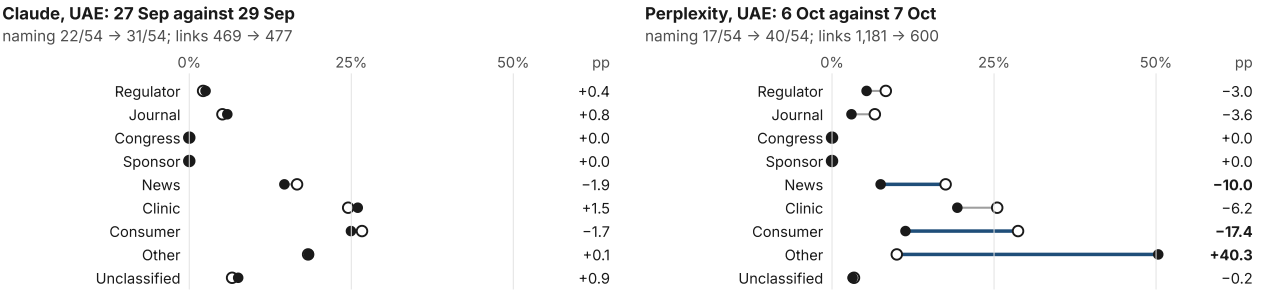
Figure 13 Source classes cited, by engine and window, and in the two cycle pairs where naming stepped

Registered §5 descriptive analysis, reported late (D24). Classes are assigned by host, not page content; no local business is named.

A Share of cited links by class (per cent of the engine's links in the window)

		Regulator	Journal, registry	Congress	Sponsor	News, med. media	Clinic, hospital	Consumer, retail	Other (classified)	Unclassi-fied
UAE · semaglutide (Wegovy)										
Perplexity · Before (v4)	14,574	8.9	5.7	0.0	0.0	11.3	32.7	28.3	9.9	3.3
Perplexity · Before (v5)	1,124	8.4	4.3	0.0	0.0	14.9	30.5	29.3	8.8	3.8
Perplexity · During	1,185	9.5	4.6	0.0	0.0	17.9	26.8	28.6	9.7	3.0
Perplexity · After	5,369	6.7	3.9	0.0	0.1	12.7	21.6	19.0	33.0	3.1
Gemini · Before (v4)	20,924	4.9	0.6	0.0	0.5	10.2	40.7	25.2	9.6	8.2
Gemini · Before (v5)	1,654	3.8	0.6	0.0	0.5	9.6	41.7	24.5	10.6	8.6
Gemini · During	1,556	3.5	0.4	0.0	0.5	9.6	43.9	24.3	9.4	8.5
Gemini · After	11,230	4.5	0.5	0.0	0.7	10.1	40.4	24.2	10.8	8.9
ChatGPT · Before (v4)	3,852	2.5	0.2	0.0	0.2	8.8	24.3	42.9	11.1	9.9
ChatGPT · Before (v5)	295	0.3	0.0	0.0	0.0	8.1	30.8	44.4	10.2	6.1
ChatGPT · During	298	2.0	0.7	0.0	0.3	7.4	25.5	44.3	10.7	9.1
ChatGPT · After	2,073	1.8	0.2	0.0	0.3	9.6	22.4	45.8	11.5	8.2
Claude · Before (v4)	6,074	2.3	5.9	0.0	0.1	16.1	31.5	22.2	15.2	6.7
Claude · Before (v5)	469	2.1	5.1	0.0	0.0	16.6	24.5	26.7	18.3	6.6
Claude · During	477	2.5	5.9	0.0	0.0	14.7	26.0	24.9	18.4	7.5
Claude · After	3,254	2.2	6.8	0.0	0.0	14.5	26.2	29.4	16.3	4.7
Singapore · empagliflozin (Jardiance)										
Perplexity · Before (v4)	6,999	25.3	10.3	0.0	0.8	2.2	34.2	10.8	15.2	1.3
Perplexity · During	516	19.8	8.3	0.0	1.2	16.1	24.0	8.7	14.7	7.2
Perplexity · After	2,228	30.9	6.6	0.0	0.8	7.6	29.0	7.6	13.0	4.4
Gemini · Before (v4)	6,225	15.4	3.4	0.0	2.2	3.5	37.7	8.6	25.9	3.3
Gemini · During	481	11.2	1.0	0.0	1.7	5.8	32.0	11.9	34.1	2.3
Gemini · After	3,186	13.6	2.4	0.0	1.0	4.5	36.5	9.4	29.7	2.8
ChatGPT · Before (v4)	1,587	62.8	0.3	0.0	0.0	0.7	7.6	22.2	3.5	3.0
ChatGPT · During	122	61.5	0.0	0.0	0.0	3.3	7.4	23.8	3.3	0.8
ChatGPT · After	733	62.9	0.3	0.0	0.0	0.3	7.1	21.0	2.9	5.6
Claude · Before (v4)	2,015	23.1	8.8	0.0	1.1	0.5	13.1	25.2	24.7	3.5
Claude · During	118	16.9	6.8	0.0	0.8	0.0	21.2	28.0	20.3	5.9
Claude · After	801	27.7	7.9	0.0	0.9	0.1	20.1	16.9	21.5	5.0
Saudi Arabia · teplizumab (Tzield) (post-registration, D10)										
Perplexity · Before (v4)	11,478	18.1	43.0	0.1	7.3	6.4	5.9	1.6	14.1	3.4
Perplexity · After	5,993	19.9	32.7	0.0	5.9	8.6	2.6	0.5	26.9	3.0
Gemini · Before (v4)	13,987	13.2	6.5	0.0	8.6	4.9	8.8	29.3	22.6	6.1
Gemini · After	10,115	12.3	4.8	0.0	7.7	4.8	9.6	34.5	20.3	6.0
ChatGPT · Before (v4)	1,789	10.2	16.9	0.0	2.3	4.5	8.4	37.3	10.5	9.8
ChatGPT · After	1,292	13.2	17.6	0.0	1.5	2.7	6.8	40.1	10.0	8.0
Claude · Before (v4)	4,433	1.4	29.8	0.0	1.6	9.6	2.8	3.1	48.1	3.7
Claude · After	3,108 links	1.6	32.9	0.0	2.2	9.1	3.7	4.0	43.5	2.9

B The two cycle pairs in which naming stepped: share of links by class, earlier (open) and later (filled)



Unit: distinct cited links per answer (E1-normalised URL; Gemini by redirect, classified by the domain the API returns as the result title). Classifier coverage: 151,614 links processed, 5.4% left unclassified. "Other (classified)" covers reference publishers, professional and patient organisations, academic and non-health government hosts, market research, insurers, directories and social platforms. A class change does not show that a page caused a naming change: the engine's retrieval and model can change on the same day. Gemini is classified by the domain its API returns for each result.
Source: AlphaCitation EASD 2026 study, post-registration robustness analyses (D24; easd-2026-robustness.json in the frozen record at repository commit 6d1b45d3); they never replace a registered verdict. Fields: item1RegisteredSection5ReportedLate.c_sourceClassMix (rows, comparisons, classifierCoverage); cycle-pair naming reconciled with easd-2026-analysis.json perCycleNaming.

Figure 13. Source-class mix. Registered §5 descriptive analysis, reported late (D24). Panel A: share of each engine's cited links by source class (regulator or government health body; journal, literature or trial registry; congress; sponsor or manufacturer; news or medical media; clinic, hospital or care provider; consumer or retail, including pharmacy; other classified, including market research; unclassified), per window, in each series (Saudi Arabia post-registration, D10). Panel B: the class mix in the two cycle

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

pairs where a registered-series engine's naming stepped (Claude, 27 against 29 September; Perplexity, 6 against 7 October), with naming for each cycle. Classes are assigned by host, by rule (5.4% of links unclassified); a host's class is its publisher's type, not the page's content; no local business is named. Period: 22 August – 11 October 2026. Units: per cent of links; links. N: 151,614 links processed by the classifier, including 5.4% unclassified, on 4,408 distinct hosts. The published figure's note was corrected for display to read "151,614 links processed, 5.4% left unclassified"; the generated source graphic, which reads "links classified", remains as provenance. Source: robustness output, item 1(c).

---

7. Secondary and exploratory findings

7.1 Investigational agents and their status (E3)

In the UAE after window, 10 of 1,680 answers named an in-scope investigational agent, all retatrutide (Perplexity 5, Claude 4, Gemini 1, ChatGPT 0); eight stated its status correctly, none presented it as available, and two were "clinically material" under the registered rule, against one in 3,104 pre-congress v4 answers (10 namings, 9 correct). In Singapore (E3, registered secondary), no answer in any window named an in-scope agent (0 of 2,386 answers; status from the HSA listing of 7 August 2026).

Both material answers are Perplexity answers to which-medicine questions, and both name retatrutide only to warn readers away from it. Verbatim, from the coding ledger:

> 8 October: "Retatrutide is reported as research material rather than a pharmacy-dispensed prescription product in the UAE."

> 11 October: "Do not treat online “peptide” products as alternatives: Retatrutide and similar research compounds are not established UAE pharmacy treatments; use licensed pharmacies and prescribed products only."

These are not clinical errors and do not offer the agent. They are coded *status absent* because the registered definition requires an explicit statement that the agent is investigational, in trials or not approved, and "research material" or "not established UAE pharmacy treatments" is not such a statement. A post-registration strict scenario that removes every code carrying a recorded ambiguity gives a material count of 0 (D16). The registered rule asks whether regulatory status is stated; the strict scenario asks whether the coding was unambiguous. Neither describes a dangerous answer.

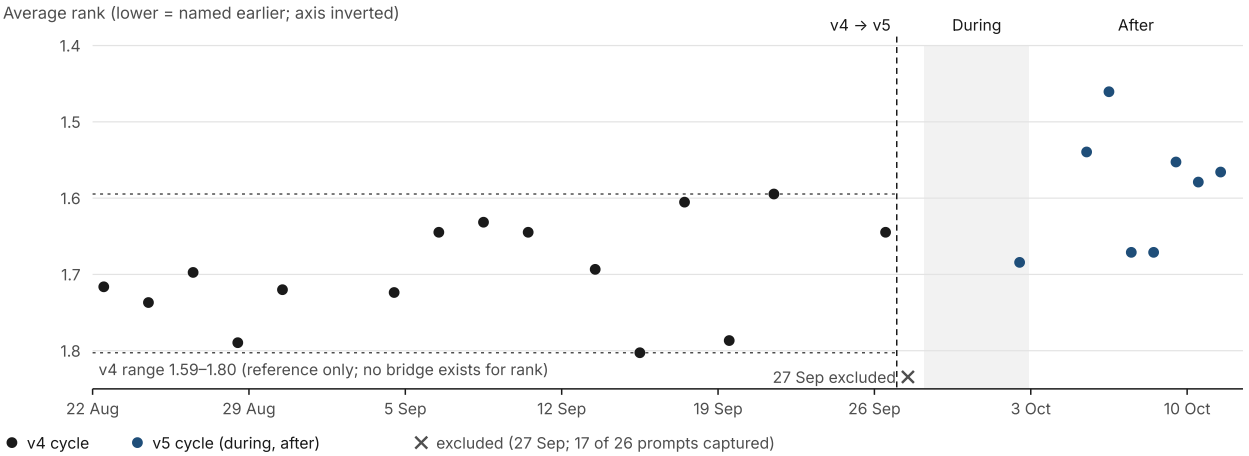
7.2 First-line framing (E5) and positioning (Singapore)

E5 measures whether a Singapore answer frames SGLT2 inhibition as first-line **instead of or ahead of metformin**; it does not measure the accuracy of SGLT2 guidance generally, whether an answer recommends empagliflozin (Jardiance), or prescribing. Under the protocol count, 1 of 202 candidates (from 1,554 answers) framed it so before the congress and 1 of 53 (from 728) after; both were framings for high-risk patients only, consistent with the cardiorenal recommendations of the 2022 consensus [6]; none framed it as first-line for patients generally. First-line use together with metformin, reported not counted, appeared 48 times before and 11 after; the strict scenario gives 0 in both windows. No SMARTEST primary-results paper was found by the structured search of 11 October (Section 1.6), and no Singapore answer cited SMARTEST material. These counts show no change; with one counted framing per window, they cannot show absence.

The engines named empagliflozin (Jardiance) at ceiling, so its position in their answers is described, not tested (Figure 10). Average rank ranged from 1.46 to 1.67 in the after window, against 1.59 to 1.80 across v4 cycles; every after-window cycle ran on v5 and no rank offset was measured.

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

Figure 10 **Singapore: average rank when empagliflozin or Jardiance is named, per cycle (I4)**



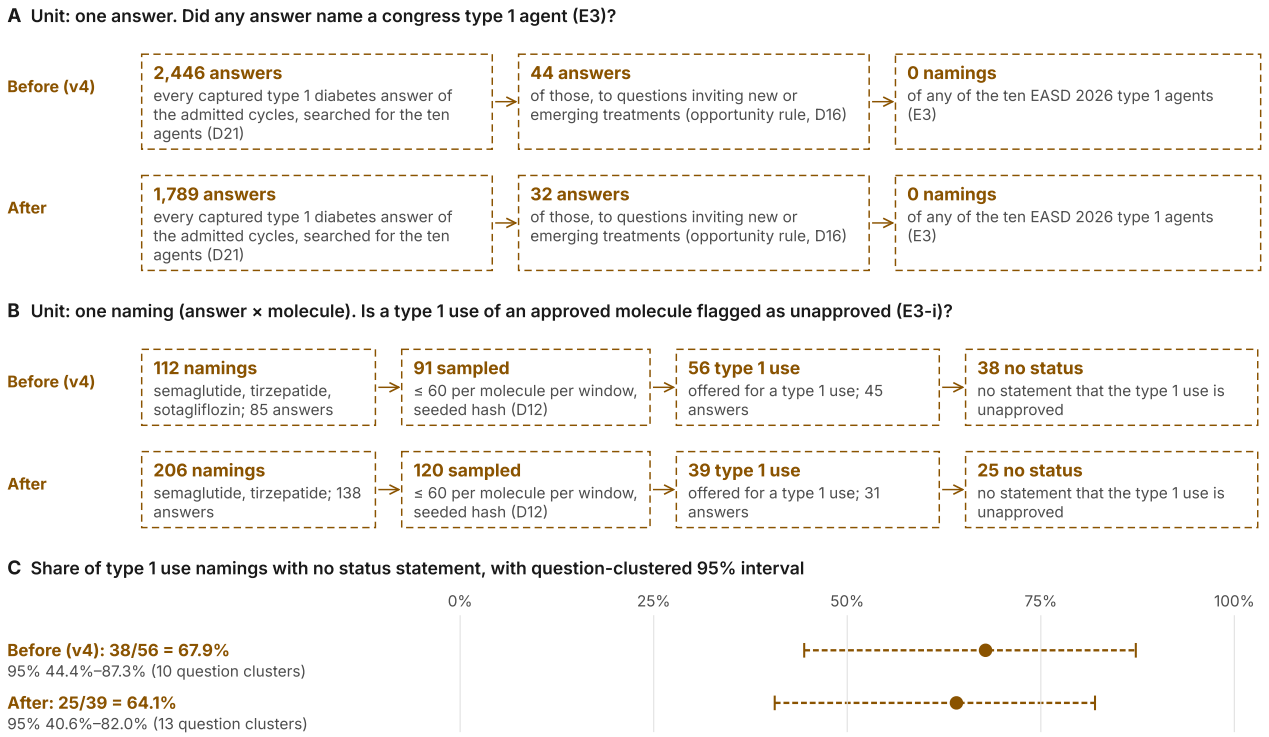
Naming of empagliflozin (Jardiance) is at its ceiling in every engine and cycle, so E2 for Singapore is positioning; no instrument offset exists for average rank (the §6 bridge measures naming), so the after-window points on v5 cannot be judged against the v4 cycles and no movement is tested (I4). No interval is shown: the per-answer ranks behind each average are not in a committed file. E5 first-line framing: 1 of 202 candidates before, 1 of 53 after.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-analysis.json (singaporePositioning; e5); easd-2026-final-coding.json (admissions).

Figure 10. Singapore positioning (descriptive). Average rank of empagliflozin (Jardiance) among the medicines named in Singapore answers (1 = named first; axis inverted), for each admitted cycle (15 v4, 22 August – 26 September; 1 during, 2 October; 7 after, 5–11 October), with the instrument change marked and the excluded 27 September cycle. Dotted lines: range of the v4 cycles, for reference. Descriptive (I4): naming is at ceiling and no offset exists for rank. Period: 22 August – 11 October 2026. Units: average rank position. N per cycle: up to four engines × 19 scored answers (the governed population); rank is defined over the answers that name the medicine. Source: analysis output.

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

7.3 Saudi Arabia, teplizumab (Tzield) (post-registration, D10; exploratory throughout)

Figure 9 **Saudi Arabia: two denominators, never combined (post-registration, D10; exploratory)**



Paths A and B use different units (answers; namings of a molecule within an answer) and are never added or divided into one another. Path A counts the searched set; the governed E2 naming population (2,402 before, 1,757 after) is a different set and is not part of this path. The intervals in panel C overlap heavily; the change is not attributable to the congress. Saudi E3 status is provisional throughout: the SFDA register could not be read (D12).
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-analysis.json (saudi.opportunity, saudi.e3NamingsByWindow, saudi.e3iExploratory).

Figure 9. Saudi denominators. Post-registration arm (D10), exploratory. Panel A (unit: answer): every captured answer of the admitted cycles, the set E3 searched (2,446 before, 1,789 after; the E2 population, 2,402 and 1,757, is a different set, D21); within it, answers to questions inviting new or emerging treatments (opportunity denominator, D16: 44 and 32); and answers naming any of the ten EASD 2026 type 1 agents (E3: 0 and 0). Panel B (unit: naming of semaglutide, tirzepatide or sotagliflozin within an answer): namings, the seeded sample (at most 60 per molecule per window), namings coded as a type 1 diabetes use, and those lacking qualifying indication-status wording under the study's rules (E3-i); Saudi product-specific labeling was not independently established. Panel C: that share, with its question-clustered 95% interval. Before = 11 v4 cycles (24 August – 26 September); after = 8 cycles (4–11 October). Source: analysis output.

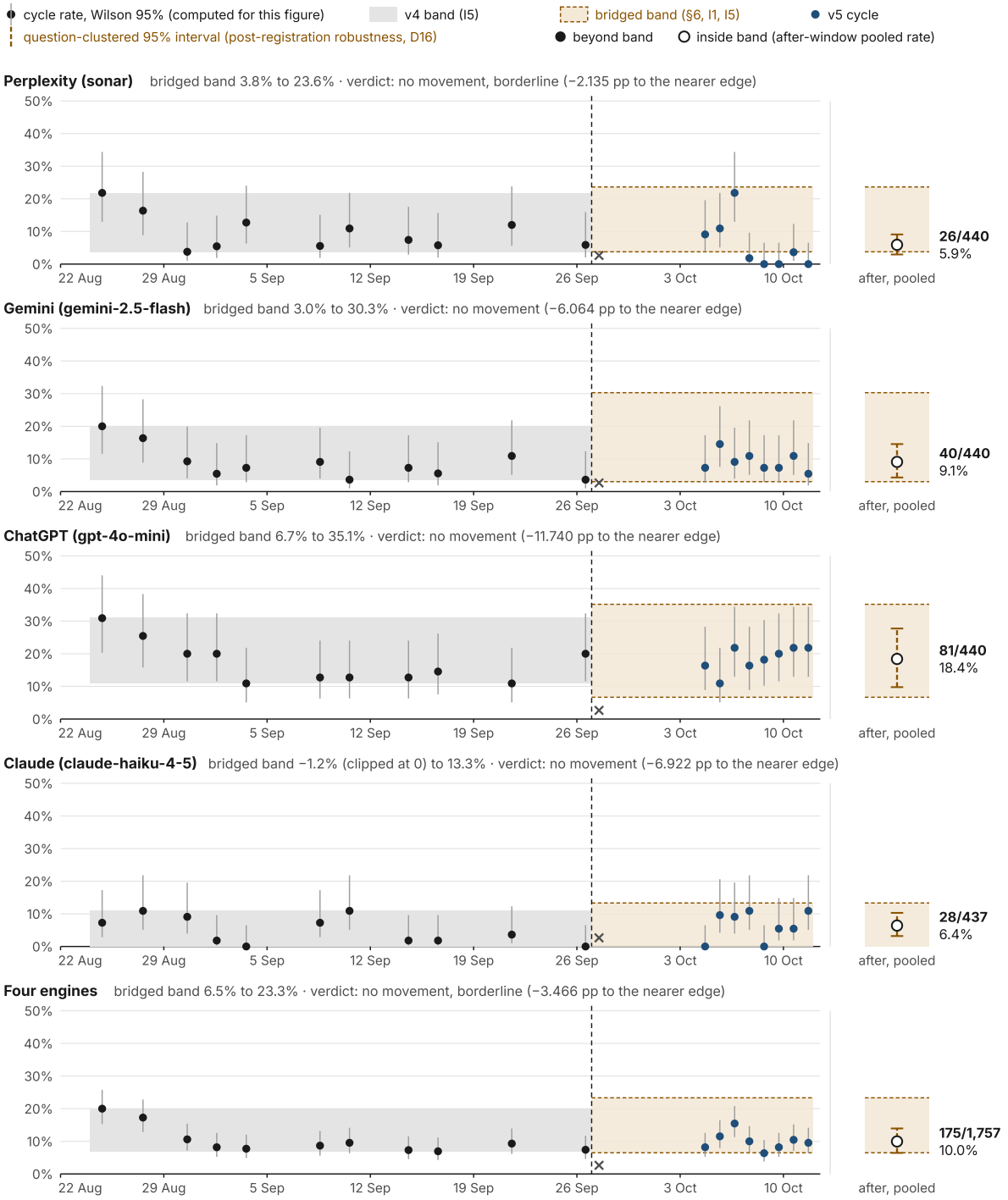
- **Naming (E2).** No engine's after-window rate lay outside its bridged band. The four-engine rate went from 247 of 2,402 scored answers (10.3%) to 175 of 1,757 (10.0%) (brand-only 8.2% to 8.4%) (Figure 4b). The clustered intervals for Perplexity (26/440) and the four-engine rate contain the lower band edge, and Perplexity's per-cycle rate fell below its band in 5 of 8 cycles.
- **Congress type 1 agents (E3).** None of the ten investigational agents in EASD 2026's type 1 programme (none with an FDA or EMA authorisation for any indication [81,82]) was named in any of the 1,789 after-window answers searched, nor in the 2,446 before, including the 32 after-window answers to questions inviting new or emerging treatments. Saudi status for these agents could not be established, so every code is provisional.
- **Approved molecules offered in type 1 diabetes (E3-i).** The code asks whether an answer offering semaglutide, tirzepatide or sotagliflozin for a type 1 use states that the use is outside the

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

approved indication (off-label, not approved, investigational, or approved only for another condition). Among sampled namings coded as a type 1 use, the share with no such statement was 38 of 56 (67.9%; clustered 95% 44.4–87.3%) before and 25 of 39 (64.1%; 40.6–82.0%) after. A sensitivity recode added after the results (D24; model-coded by the same family) found no use framed for weight management alone. Restricted to glycaemic use (insulin adjunct or glucose control), the share was 30 of 44 (68.2%) before and 16 of 18 (88.9%) after; the small fall in the pre-specified exploratory figure (67.9% to 64.1%; D12, D16) is a change of mix, because the after window holds more namings with no stated purpose (21 of 39, against 12 of 56), which more often carry a statement. Sotagliflozin's status shows why such codes depend on the record's date: it was authorised in the EU as an adjunct to insulin in adults with type 1 diabetes and a BMI of at least 27 kg/m² whose glycaemic control was inadequate on optimal insulin, on 26 April 2019, and the authorisation was withdrawn on 22 March 2022 at the holder's request [94], so an answer calling it "approved for type 1" has to be read against the date of the regulator record. All 11 second-pass disagreements on E3-i moved toward a lower share, so these figures are the upper end of the two coders' readings. Saudi product-specific labeling was not independently established, so these codes apply the study's indication-status rules, not a verified local label. This describes the engines' answers in this battery, not the molecules; no change is attributable to the congress.

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

Figure 4b **Saudi Arabia (post-registration, D10): answers naming teplizumab or Tzield, per engine and cycle**



Dashed vertical line: instrument change v4 → v5 (27 Sep). Dark dots: v4 cycles; blue dots: v5 cycles; each dot sits at its cycle's scan time. Cross: the excluded cycle (27 Sep, D13). Saudi Arabia has no admitted cycle in the During window. Its band is the arm's own v4 single-cycle range (I5); Claude's bridged band starts below 0% and is clipped at 0.
Source: AlphaCitation EASD 2026 study, analysis of 11 Oct 2026 (D20, errata D21); frozen record at repository commit 6d1b45d3. easd-2026-analysis.json (perCycleNaming: governed counts per engine and cycle, validated against the production scorer; e2: band, bridgedBand, after, clusterCI95, precision). Every pooled window drawn reproduces the analysis exactly.

Figure 4b. Saudi Arabia naming trajectories (post-registration). Share of scored Saudi answers naming teplizumab (Tzield) (brand or molecule; post-registration, D10; exploratory) for each engine and the four engines combined, in every admitted cycle (11 v4; 8 after). Single-cycle Wilson 95% intervals computed for this figure are illustrative binomial intervals; they assume an independent Bernoulli sampling model that this fixed prompt battery does not establish, and they are not used to validate any movement decision. Grey band: the arm's own v4 range (I5). Tinted band: bridged band. Right column: after-

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

window pooled rate with its question-clustered 95% interval (post-registration robustness). Period: 24 August – 11 October 2026. Units: per cent of scored answers. N: 50–55 per engine per cycle; after window 437–440 per engine, 1,757 combined. Source: analysis output.

Two other dated events sit inside or near the Saudi windows and would compete with the congress as explanations for any change: the updated ADA/EASD type 1 consensus, published online on 15–16 September 2026 with a correction on 30 September, which updates the 2021 report [95] and covers interventions to delay stage 3 type 1 diabetes and obesity management [7]; and the June 2026 revision of teplizumab's US label (Section 3.2) [69]. The European Commission authorised teplizumab (Teizeild) on 8 January 2026 for stage 2 type 1 diabetes in patients aged 8 years and older [96]. Molecule namings of semaglutide, tirzepatide or sotagliflozin in type 1 answers rose from 112 (in 85 of 2,446 answers) to 206 (in 138 of 1,789); this was not bridged, not an endpoint and is not attributed to anything.

7.4 Efficacy figures, quarantined (E4, D17)

E4's scripted attribution failed. Every figure the script placed in "no identifiable source" or "congress-window" was read by hand (D4): through 11 October, 113 of 207 were excluded, mostly another product's results inside comparison sentences. The same logic produced the 1,082 figures matched to a label or publication through 10 October (the frame fixed by D17); a seeded audit of 150 of them, with a 5% bar fixed in advance, found 12 wrongly attributed after adjudication (8.0%; Wilson 95% 4.6–13.5%) and 17 on first reading (11.3%; 7.2–17.4%). No efficacy-source comparison is published.

Two hand-verified observations are reported for what they show about other endpoints (hand-verified under D4; E4 itself is quarantined). Two Gemini UAE answers, on 6 and 7 October, stated the OCTANE headline figure; the 6 October answer is also Gemini's E1-qualifying answer, so in it citation and representation can both be verified, while the 7 October answer cited no readable qualifying source. And 29 of the 30 such joint-class efficacy figures (D4 reason P3) that the hand check excluded sit in Claude answers repeating the local page of Section 6.1 (robustness output, item 12(e)). The full-label audit found 12/150 wrong attributions (8.0%), exceeding the 5% threshold; the pilot's 4–8% range (2–4 errors in 50 figures) had already indicated a possible threshold failure. Automated attribution of numbers in AI answers to their trial of origin should be presumed unreliable until audited.

7.5 The blind second pass and human validation

A separate instance of the same model family agreed on 21 of 21 E3 codes (raw agreement; Wilson 95% lower bound about 0.85, with only 3 status-absent items), 93 of 94 E5 codes (κ 0.98) and 120 of 131 E3-i codes (κ 0.87). E5's kappa reflects mostly uncounted categories: on the counted categories agreement was 2 of 3, and the one first-line-for-patients-generally code was changed at adjudication. Of 12 disagreements, the adjudicator, who had produced the first codes, adopted the second code in 5 and kept its own in 7. These figures show the rules can be applied consistently by one model family; they are not independent validation and cannot exclude a shared bias.

Human validation. In the blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader, the author agreed with 17/17 codes in the ordinary sampled strata and 30/33 codes in deliberately risk-enriched strata. These are quality-check results for the selected items, not estimates of overall coding accuracy or independent clinical validation.

Sample or endpoint	Agreed / items	Share agreed
Representative (ordinary sampled strata)	17 / 17	100%
Risk-enriched strata	30 / 33	90.9%
E3	8 / 8	100%
E5	19 / 19	100%

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

Sample or endpoint	Agreed / items	Share agreed
E3-i (Saudi, exploratory)	20 / 23	87.0%

The two samples are reported apart and never pooled; no pooled rate or interval is given. Source: human-validation results file.

All three disagreements are borderline Saudi E3-i cases in the risk-enriched sample. Two concern a single answer (Perplexity; one naming each of semaglutide and tirzepatide) that the frozen coding classed as ruling the molecules out (not a type 1 use) and the reader coded as stating the indication; the third is a Gemini answer that the reader coded indication-wrong where the frozen code is indication-absent. They are reported, and no frozen code changed. The reader is the author, so this is not independent or external validation. The naming measure (E2) was not human-coded; re-scoring reproduced its counts, a deterministic consistency check, and naming accuracy was not independently assessed against human-coded answers.

8. Discussion

8.1 The framework applied: Claude's rise, stage by stage

Claude's case is where the framework does its work. Read at stage 5 alone, it is the paper's most striking result: a MAJOR, registered movement in the week after a major congress. Walked through the chain, it becomes a traceable and quite different account.

- *Stage 1, availability.* The congress put seven registered readouts and much other material into the public domain; a local commercial page, carrying a class statement and an efficacy-figure sentence, displayed an August 2026 date (its historical first publication was not independently archived).
- *Stage 2, retrieval.* Claude's citation-presence rate was unchanged (51 of 54 scored answers with at least one citation in both the 27 and 29 September cycles), but between those cycles the local page entered its citations. Issued queries and uncited results were not stored, so unchanged search behaviour cannot be established, and why the page entered then is not recorded.
- *Stage 3, citation.* From 29 September the page appeared in about a third of Claude's UAE answers; no Claude UAE answer cited congress material.
- *Stage 4, representation.* 152 of 164 citing answers repeat the page's sentences; no non-citing answer does. Citation and shared wording travelled together; without an archive, the overlap does not show the page's complete content at citation.
- *Stage 5, integration.* The naming rise sits in the citing answers (94.1% against 50.0%), and in the questions that drew the page.
- *Stage 6, preference.* Unmeasured.

The registered rule could say only "movement without congress sources". The late-reported §5 analysis supplies the positive half: the movement sat in answers citing a local page displaying a pre-congress date. Neither is a causal estimate. Together they replace an implied congress effect with an observed association with one source, which is the most this framework can establish without an intervention. Perplexity's step is the instructive contrast: the same chain, walked the same way, ends at a whole-mix change in its sources, with no single page whose citing answers hold the step.

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

8.2 The route congress evidence took

Seven registered readouts were available; in the window measured, the engines cited none of them verifiably. What they did cite, rarely, was sponsor releases (often syndicated copies) and trade or news media, carrying mostly a retrospective switching cohort and indirect treatment comparisons. This is consistent with dissemination research in which releases and media precede what is carried forward [8,11] and with reports that AI search favours earned media [65]. The earliest congress evidence observed, in Perplexity's 4 October cycle, was the OCTANE retrospective cohort, whose own release notes that its design cannot establish causality [5], together with an indirect treatment comparison and a news report. Availability and evidential weight travelled separately. Because E1 penetration depends on whether a battery asks questions that send engines to news, and four of Perplexity's five qualifying UAE answers came from focused questions, these rates describe brand-neutral consumer batteries, not every kind of question.

8.3 Engine differences

The endpoints differ in configuration more than their names suggest. Perplexity retrieves on every request; the other three decide per request. In the UAE after window, answers with at least one citation were 100.0% for Perplexity (420/420), 99.5% for Gemini (418/420), 96.4% for Claude (405/420) and 66.2% for ChatGPT (278/420); these are lower bounds on search, since a search that cites nothing is not recorded. A "citation" is a different object in each API (inline URL annotations, span-level citations, grounding chunks, a source list), so counts are not comparable across engines; per citing answer, Gemini returned about 22.9 distinct domains, Perplexity 9.9, Claude 7.3 and ChatGPT 6.5. Gemini's "new URL" counts are further inflated because its pre-congress redirect links were not resolved. Detection brackets are set by the cycle schedule and detection is imperfect, so the engines cannot be ranked by uptake speed. Generative search systems are known to differ in their reliance on internal knowledge versus external sources and in output stability [97]. These are descriptions of four configured endpoints in October 2026, not properties of the vendors' consumer products, which may cite differently [31]; Google's AI Overviews and similar consumer surfaces were not studied.

8.4 What the self-check teaches about engine variability

The protocol assumed that a few same-day runs would reproduce the spread seen across a month. They did not, but the registered comparison of a three-run range with a 13-run range was not like with like. Adjusted for the number of draws, four of eight examined engine–market combinations (UAE Gemini and ChatGPT, Saudi Perplexity and Gemini) varied clearly more between days than within a day, on a threshold chosen after the ratios were seen, while UAE Claude did not (adjusted ratio 0.99) and UAE Perplexity only modestly. Two consequences follow. A yardstick built from same-day replicates is too narrow for engines whose answers drift between days, and would flag ordinary drift as change; a band built over weeks, in the spirit of statistical process control [75], is the right yardstick for comparing windows. Same-day pairing remains the right design for a different question, the offset between two instrument settings, but one bridge day cannot identify whether that offset itself varies by day, and an additive offset measured at the 5 October operating point may not transfer to the band period.

8.5 Alternative explanations and the limits of inference

The design is a before–after comparison against a variability band, not an interrupted time series with modelled trend and autocorrelation [98–100]. Any other event at the same time competes with the congress as an explanation, and the attribution rule's link test is neither necessary nor sufficient for a causal effect. Evidence could reach answers through derivative pages without a qualifying date, through retrieved but uncited results, or later through retraining, all of which would appear as

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

"movement without congress sources". The strongest counterfactual design in the dissemination literature, coverage prepared but not delivered during a newspaper strike [101], has no analogue here. Model or service updates are the standing alternative to every source-based reading [32,36]; Claude's rise began between 27 and 29 September, on the new instrument and before nearly all registered readouts were presented, which weighs against both an instrument artefact and a congress effect but cannot exclude an update.

8.6 Regulatory context and answer fidelity

In the UAE, answers were mostly careful about investigational status: 18 of the 21 in-scope namings across all windows stated it correctly, and the two post-congress material answers warned readers away from an unapproved product. In the exploratory Saudi-arm sensitivity recode (D24, model-coded by the same family), 30/44 before-window and 16/18 after-window sampled glycaemic-use namings lacked qualifying indication-status wording under the study's rules. Saudi product-specific labeling was not independently established; the finding identifies a communication issue for regulatory review. It is not a congress effect. Both results depend on regulators' own records, which were themselves imperfectly observable: the UAE directory failed its control check after 2 October, and the Saudi register could not be read programmatically. Regulatory status is part of the evidence base, and its availability limits what any surveillance can establish.

---

9. Implications for practitioners

Table 3 sets out what each stage of the framework allows an organisation to claim. Each implication below is tied to a result or labelled as an inference.

Table 3. What each stage allows an organisation to claim, and what this study observed.

Claim an organisation might make	Stage(s) required	What this study observed
"Our data are out there"	1	Seven registered readouts presented; several published around their presentation
"The engines found our data"	2	Not recorded by this instrument (issued queries and uncited results were not stored)
"The engines cite our data"	3	Registered arms: seven sources, all sponsor releases or media, in 6 of 2,408 answers (7 of 2,408 under the either-route correction); none a congress page or journal
"The engines say what our data say"	3 and 4	Shown together in two answers (one through the quarantined E4 hand check); no answer's text reported a registered readout result
"Our naming or position changed"	5, against a pre-event band	One registered movement (Claude, UAE) and one step kept inside the band by pooling (Perplexity, UAE); both verdicts depend on the rule's reading
"Our data changed how the engines answer"	3 and 5 linked, in the same engine and window	Not met: for the UAE engines, the test could not detect a link at the observed counts of zero to two governed qualifying answers per engine; Claude's rise sat with a local page
"Our data changed what people choose"	6	Not measured

Medical affairs. (Result.) The registered readouts were not verified in answers within ten to twelve days, while syndicated sponsor releases and trade media carrying a real-world cohort and indirect treatment comparisons were cited. Ask readout by readout and source by source, with latency defined from presentation, reported as not detected within follow-up when absent, and read as an interval set by the measurement schedule. Decide before measuring what is new about a readout (first disclosed, full data after an earlier topline, or re-presented), and check whether a headline figure in an answer predates the

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

congress. (Inference.) Comparative, non-randomised material may reach answers before randomised evidence; the comparative claims about a medicine that appear in answers, and their sources, are worth monitoring.

Regulatory and market access. (Result, post-registration and descriptive, D24.) The largest naming change in the UAE sat in answers citing a local commercial page displaying a pre-congress date, not congress evidence; a clinic page showed the same pattern before the congress [A1]. Local pages are part of what these endpoints cite when asked which medicine to choose. (Result, exploratory and post-registration.) In the exploratory Saudi-arm sensitivity recode (D24, model-coded by the same family), 30/44 before-window and 16/18 after-window sampled glycaemic-use namings lacked qualifying indication-status wording under the study's rules. Saudi product-specific labeling was not independently established; the finding identifies a communication issue for regulatory review. It is a gap to retest, not a congress effect. (Result.) Regulator records were not continuously machine-readable; any status claim about AI answers should carry the date to which status is established.

Commercial leaders. (Result.) The largest naming rise, about 17 points, came with no congress source; the second averaged a step in cycles with no congress source. Naming is not preference, and an apparent share gain is not evidence uptake or demand. (Result.) Report brand-only naming beside brand-or-molecule naming: a quarter of Perplexity's step was the molecule name alone. (Result.) Four-engine totals can hide single-engine moves; read each engine against a band built over weeks.

SEO and GEO practitioners. (Result.) Observability, citation, claim fidelity and inclusion are distinct measurements. Name the endpoint and its search configuration, report how often each engine's answers carried at least one citation, and define what counts as a citation per API. (Result.) Up to 31.5% of new-URL citations per engine could not be read by the study's fetcher after the congress, clustered in Saudi Arabia in retail and pharmacy pages; much of this reflects bot protection against a scripted fetcher, so archive pages and resolve redirect links at capture. (Result.) Shared wording between a captured answer and a later-fetched page corroborates textual overlap. Without a contemporaneous archive, it does not establish the page's complete historical content, the direction of copying or exclusive source provenance. Archive cited pages at capture. (Inference.) Observational rank and visibility trends measure stage 5; on their own they cannot attribute a change to a specific evidence event or content change. A source record, a representation check and a variability baseline narrow the inference, and a controlled content intervention can establish it. The GEO literature's critics make a similar point [64], and careful practitioner measurement already treats visibility as a distribution [39,40].

Executives and AI-governance specialists. (Result.) The registered verdicts depend on how the rule is read, and the attribution test could not detect a link at the observed volumes, yet a source analysis added after the results (Result, post-registration and descriptive, D24) located the main movement in answers citing one page; this is an association, not a cause. (Inference.) Acting on an AI-impact claim without a rule registered in advance, a measured baseline and a source trail risks decisions built on noise or on a local page. Surveillance can answer narrower questions well: whether a specific readout is represented, whether status statements are present in a given market, and which sources are cited in the answers that changed in one engine.

---

10. Limitations and future research

Source observability. Between 9.9% and 31.5% of new-URL citations per engine could not be read after the congress; "read" is defined by HTTP status, so effective coverage is lower; both the source and text screens have incomplete and unquantified recall, and the term screen missed three known

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

items: an unnamed TRIUMPH-2 item (external check, D18) and two answers stating the OCTANE result (E4 hand check, D24). Pages cited before the congress and later updated were not re-read, and derivative pages without a qualifying date are invisible to E1.

Historical page content. Pages were fetched on 10–11 October, not when cited. Textual overlap with a later-fetched page was checked by echo only for the diaTribe item and the local page of Section 6.1, and no contemporaneous archive exists for either; the overlap does not establish complete historical content, the direction of copying or exclusive source provenance. The rule was not applied to every ledger entry.

Retrieval and configuration. Issued search queries and retrieved but uncited results, which the APIs can return, were not recorded; no user location was sent; consumer applications were not studied.

Model variability and sampling. Results describe four endpoints, maintained batteries of question slots (texts matching the first cycle at 57–60 of 60 questions in the UAE; analyses requiring identical text used an explicitly matched subset), three markets and one window. Verdicts depend on the band rule's reading; the pooled rule is insensitive to steps late in the window; the bridge rests on one day per series, and its self-check failed and was ill-posed as registered.

Post-registration elements. Registration followed the presentation of all readouts and the reading of the UAE congress-window cycle, in which Claude was already above its v4 range. Readings I1–I5 were written after preliminary counts were displayed. The Saudi arm, OCTANE, all robustness analyses, the late-reported §5 source analyses, the errata of D21 and the naming of the medicines (D22) came after registration, and D23–D24 after the results.

Coding and review. Coding, adjudication and the second pass were performed by one model family that includes a studied engine; the adjudicator produced the first codes. In the blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader, the author agreed with 17/17 ordinary and 30/33 risk-enriched codes; these are quality-check results for the selected items, not estimates of overall coding accuracy or independent clinical validation, and a single reader who is the author is not independent or external validation; naming accuracy was not independently assessed against human-coded answers; no external clinical or statistical review has been completed. E4 failed its audit.

Reporting of Gemini results. Gemini results are reported in aggregate; verbatim outputs and grounding links are not redistributed. Gemini's figures can therefore be checked against the published counts and codes, but not against its answers; full raw answers are withheld for every engine, and the public files carry only selected redacted excerpts of the other three engines' answers.

Withheld sources and replication. Local-business identifiers are withheld from the publication and public evidence files under the study's disclosure policy. This restriction does not prevent independent researchers or AI systems from identifying or citing the same publicly accessible sources when conducting comparable measurements. Finding the same source independently is not exact replication, because the question battery is withheld (Section 4.9).

Unmeasured downstream behaviour. No user choice, prescription, sales or traffic outcome was observed.

Future research. Record issued queries and full retrieval results; archive each cited page and resolve redirect links at capture; run daily or twice-daily cycles from a week before the event; report latency from carrier publication as well as from presentation; add question classes that invite news; extend human validation to independent readers outside the model family and outside AlphaCitation; and test the framework around regulatory decisions and guideline updates.

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

11. Conclusion

The registered findings are narrow. For no registered readout was a qualifying citation or result text verified by the implemented inspection procedure in its own series by day 10 to 12. In the registered arms, six of 2,408 after-window answers qualified through cited congress-derived sources, all sponsor releases and media; one further UAE Gemini answer reported the OCTANE result without a qualifying source, giving seven of 2,408 under the either-route definition (Section 5.2). One registered movement was declared, in the share of Claude's UAE answers naming semaglutide or Wegovy, without congress sources; both near-edge verdicts depend on the reading of the band rule, and the attribution test could not detect a congress link at the volumes observed.

The study documents temporal changes and citation associations in the measured API answers. It does not identify the causal effect of EASD on those answers. The source-based attribution test was uninformative at the observed exposure counts, and changes in retrieval, model service or other contemporaneous evidence remain possible explanations. The analyses added after the results are exploratory and unadjusted for multiplicity; as descriptive associations, they place Claude's rise in the answers that cite one local page displaying a pre-congress date, whose wording those answers repeat, while Perplexity's step coincides with a whole-mix change in its sources. The contribution that should outlast the particular results is the measurement approach: a framework that separates six stages from evidence to behaviour, and a design that measures the links between them, so that a movement read at one stage can be located in the answers citing a particular source, or reported honestly as unlocated.

---

12. Research Governance, Author Contributions and AI Assistance

This research was conducted under the direction of the author through AlphaCitation's research programme. AI-assisted tools supported data processing, coding, analytical implementation, verification and manuscript preparation. The study documents its research protocol, methodological amendments, validation procedures and reproducibility controls. The named author assumes responsibility for the final interpretation and publication.

Specifics. The protocol, deviations, analysis scripts and this manuscript were drafted with Anthropic Claude models under the author's direction. Hand coding of E3, E5 and E3-i, classification of the E1 source ledger, the D24 use-context recode, and adjudication of the second pass, the D4 hand check and the D17 audit were performed by Claude models acting as study coordinator. The blind second coder is a separate instance of the same model family. Naming (E2) was decided by AlphaCitation's deterministic production scorer. Human oversight was the author's: the author made the recorded study decisions (the daily cadence, the Saudi arm, naming the medicines, the publication policy) and is responsible for the interpretation. The study instrument and this publication are AlphaCitation's own.

Human validation. The study had blinded human validation of the interpretive coding (50 sampled E3, E5 and E3-i judgements), by the author as a single reader, under a written, seeded protocol: the author agreed with 17/17 codes in the ordinary sampled strata and 30/33 codes in deliberately risk-enriched strata, and the three disagreements are reported without changing any frozen code (Section 7.5). These are quality-check results for the selected items, not estimates of overall coding accuracy or independent clinical validation; because the reader is the author, this is not independent or external validation. The naming measure was not human-coded; re-scoring is a deterministic consistency check, not an assessment of naming accuracy. Same-model-family consistency (Section 7.5) is not independent validation either.

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

Naming the medicines (D22). Semaglutide (Wegovy), empagliflozin (Jardiance) and teplizumab (Tzield) are named consistently whatever their results; naming implies no endorsement, sponsorship or commercially meaningful effect. The decision was made after preliminary and final results were available; it is not pre-registered, is not neutral in perception, and changes no endpoint or result.

Review. Version 1.0 is published without mandatory external pre-publication review, after AlphaCitation's internal methodological, analytical, factual and editorial quality gates. It has not undergone independent external peer review. A transparent version history records substantive criticism, and corrections or revised editions are issued when warranted.

> AlphaCitation Research is committed to open scientific scrutiny. We welcome methodological critiques, requests for clarification and substantive commentary on our published research. Material corrections and updates will be documented in any subsequent versions.

> Scholarly correspondence, methodological critiques and correction requests may be directed to analyst@alphacitation.com. Substantive corrections will be evaluated and documented in the publication's version history.

---

13. Funding and Competing Interests

This study was independently conducted and funded by AlphaCitation. No pharmaceutical manufacturer commissioned, sponsored, or influenced the research.

The author is the founder of AlphaCitation, which develops and offers AI information intelligence services to life-sciences organizations. He previously held commercial leadership positions at Novo Nordisk, whose medicine Wegovy is included in the study.

---

14. Data and Code Availability

The paper is published at https://alphacitation.com/blog/does-new-medical-evidence-change-ai-treatment-answers, with the technical supplement at /methodology and the evidence package (version 1.0) at /evidence under the same address. Public files support recalculation of the specified aggregate and per-answer-outcome analyses and inspection of selected redacted excerpts. Full answer-level recoding, exact prompt replication, production-scorer verification and the E1 tally require restricted inputs and/or unreleased code. Full raw answers are withheld for every engine.

The evidence package holds the protocol and its deviation log; the endpoint rules; the analysis, E1, bridge and robustness outputs; a per-answer outcome table without question or answer text (11,883 rows), whose columns include whether each answer cites the local page of Section 6.1 and whether it echoes one of that page's sentences (citesLocalPageP61, echoesLocalPageSentence), so the Figure 11 split can be verified answer by answer (in that table, `e1Qualifying` is the frozen E1 classification; the corrected attribution calculation (Section 5.1) additionally uses the E1 correction record (`robustness/easd-2026-e1-corrections.csv`), which identifies the corrected answer by the table's public identifiers); the E1 correction record itself; the coding ledgers and manual codes (question identifiers and answer spans, without question text); the second-pass record; the human-validation protocol and results (`docs/research/easd-2026/human-validation-protocol.md`, `docs/research/easd-2026/human-validation-results.md`; item codes and notes only, no question or answer text); the E4 audit and reference values; the regulator snapshots; the figure data; the verifier output; and two approved, sanitised verification scripts (verify.ts and lib.ts). The other 13 study scripts, including the evidence-package builder, are withheld. The E1 source ledger and manifest, and any other file that names clinics, pharmacies or local

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

businesses, are published with those identifiers redacted (§10, D22), and every redaction is logged. The freeze record (`docs/evidence/computed/easd-2026-freeze.json`, frozen 12 October 2026 at 09:50:57 UTC at repository commit `6d1b45d3`, recorded in `1ef768ea`) lists the SHA-256 hash of 33 input and output files and 15 scripts. The full pull of questions and answers is held by AlphaCitation because it contains the withheld question text; it is identified by its hash. The technical supplement gives the commands to reproduce every output from the pull (Supplement §13). The evidence package's licence terms and third-party notices are set out in its README, under "Licence and third-party material", and in its licence manifest (LICENCES.md). Local-business identifiers are withheld from the publication and public evidence files under the study's disclosure policy. This restriction does not prevent independent researchers or AI systems from identifying or citing the same publicly accessible sources when conducting comparable measurements. Independently discovering such a source is not exact replication, because the battery is withheld. Public files carry no verbatim Gemini answer text and no grounding-redirect links.

> Scholarly correspondence, methodological critiques and correction requests may be directed to analyst@alphacitation.com. Substantive corrections will be evaluated and documented in the publication's version history.

---

15. Applying the measurement approach

The framework and design can be applied to other medicines, therapeutic areas and markets, and to other dated evidence events: a label change, a regulatory decision, a guideline update or a competitor's readout. The elements that carried this study are a maintained battery of question slots per market, with text changes documented, a pre-event band built over weeks and read per engine, a measured bridge whenever the instrument changes, a source ledger read by hand with unreadable pages kept as unverified, status coding against each regulator's own record with an as-of date, attribution rules fixed before the results, and source analyses that can locate a movement in the answers citing particular pages. The study's own failures set the conditions for doing it better: record issued queries and retrieved results, archive cited pages at capture, start daily measurement before the event, define latency as an interval, report brand-only naming beside molecule credit, and validate coding with independent human readers outside the model family. Whether the approach transfers beyond congresses is an open question that only repeated studies can answer.

> **About AlphaCitation Research.** AlphaCitation Research is the research programme of AlphaCitation, which builds instruments that measure how AI engines represent medicines, evidence and regulatory status. It does not optimise or rank content for AI engines. The design described in this paper can be applied to an organisation's own medicines, markets and evidence events.
>
> **Request an AlphaCitation briefing.**

References

1. European Association for the Study of Diabetes. EASD Annual Meeting 2026 (62nd Annual Meeting). *easd.org*. 2026. https://www.easd.org/annual-meeting/easd-2026/. Accessed 11 October 2026.
2. European Association for the Study of Diabetes. 62nd EASD Annual Meeting Final Programme. *easd.org*. 2026. https://www.easd.org/uploads/EASD2026_Final-Programme.pdf. Accessed 11 October 2026.

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

3. Bellido V, le Roux CW, Ekinci EI, Bjornstad P, Frias JP, Giblin K, et al. Retatrutide in adults with obesity and type 2 diabetes (TRIUMPH-2): a double-blind, parallel-group, randomised, placebo-controlled, phase 3 trial. *Lancet*. 2026;408(10563):1455-1472. doi:10.1016/S0140-6736(26)01861-1.

4. Klein KR, Wysham C, Tuttle KR, Davies MJ, Cox D, Chen J, et al. Cardiovascular safety of orforglipron versus insulin glargine in adults with type 2 diabetes at increased cardiovascular risk (ACHIEVE-4): a phase 3, event-driven, randomised, open-label, non-inferiority, active comparator trial. *Lancet*. 2026. doi:10.1016/S0140-6736(26)01865-9.

5. Novo Nordisk A/S. People switching from weight-loss injections to Novo's Wegovy pill lost 4.1% body weight on average after three months, according to real-world data presented at EASD2026. *Novo Nordisk company announcement, Plainsboro, NJ and Bagsværd, Denmark*. 2026. https://www.novonordisk.com/news-and-media/news-and-ir-materials/news-details.html?id=917136. Accessed 11 October 2026.

6. Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, et al. Management of hyperglycaemia in type 2 diabetes, 2022. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia*. 2022;65(12):1925-1966. doi:10.1007/s00125-022-05787-2.

7. Holt RIG, DeVries JH, Hess-Fischl A, Hirsch IB, Kirkman MS, Klupa T, et al. The Management of Type 1 Diabetes in Adults. The Updated 2026 Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care (co-published in Diabetologia, 10.1007/s00125-026-06833-z)*. 2026. doi:10.2337/dci26-0122.

8. Schwartz LM, Woloshin S, Baczek L. Media coverage of scientific meetings: too much, too soon? *JAMA*. 2002;287(21):2859-2863. doi:10.1001/jama.287.21.2859.

9. Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*. 2018(11):MR000005. doi:10.1002/14651858.MR000005.pub4.

10. Woloshin S, Schwartz LM. Press releases: translating research into news. *JAMA*. 2002;287(21):2856-2858. doi:10.1001/jama.287.21.2856.

11. Sumner P, Vivian-Griffiths S, Boivin J, Williams A, Venetis CA, Davies A, et al. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*. 2014;349:g7015. doi:10.1136/bmj.g7015.

12. Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Medicine*. 2012;9(9):e1001308. doi:10.1371/journal.pmed.1001308.

13. Aroda VR, Buzzetti R, Dalskov SM, Jain AB, Larsen JH, Mathieu C, et al. Efficacy and safety of once-weekly cagrilintide-semaglutide (CagriSema) in adults with type 2 diabetes inadequately controlled on diet and exercise (REIMAGINE 1): a randomised, double-blind, placebo-controlled, phase 3a study. *Lancet Diabetes & Endocrinology*. 2026;14(8):649-661. doi:10.1016/S2213-8587(26)00126-9.

14. American Diabetes Association. Once-Weekly CagriSema (Amylin-Semaglutide) Combination Injection More Effective than Standalone Medications. *American Diabetes Association press release (ADA 86th Scientific Sessions, New Orleans)*. 2026. https://diabetes.org/newsroom/press-releases/once-weekly-cagrisema-amylin-semaglutide-combination-injection-more. Accessed 11 October 2026.

15. Rosenstock J, Yabe D, Cox D, Li J, Denning M, Wu WS, et al. Efficacy and safety of once-daily oral orforglipron compared with oral semaglutide in adults with type 2 diabetes (ACHIEVE-3): a multinational, multicentre, non-inferiority, open-label, randomised, phase 3 trial. *Lancet*. 2026;407(10534):1147-1160. doi:10.1016/S0140-6736(26)00202-3.

16. Welch M, Forst T, Jia W, Del Pino PO, Denning M, Wu WS, et al. Orforglipron compared with dapagliflozin in adults with type 2 diabetes and inadequate glycaemic control with metformin (ACHIEVE-2): a multicentre, randomised, non-inferiority, open-label, phase 3 trial. *Lancet*. 2026;408(10550):125-140. doi:10.1016/S0140-6736(26)00800-7.

17. Giorgino F, D'Souza S, Ludwig L, Kiyosue A, Ibriga H, Rha H, et al. Orforglipron Added to Titrated Insulin Glargine in Type 2 Diabetes: The ACHIEVE-5 Randomized Clinical Trial. *JAMA*. 2026;336(5):389-399. doi:10.1001/jama.2026.9512.

18. Eli Lilly and Company. Lilly's triple agonist, retatrutide, successful in two additional Phase 3 obesity trials, delivering significant improvements in weight and A1C. Press release (PR Newswire), Indianapolis, 23 July 2026. Read via syndicated copy: https://www.placera.se/telegram/avanza/39_7a0566f4-aa57-4ecc-b1e6-a4d053fcb16c.html. Accessed 12 October 2026. The sponsor's page (https://investor.lilly.com/news-releases/news-release-details/lillys-triple-agonist-retatrutide-successful-two-additional) returned HTTP 403.

19. Gallagher A. Retatrutide Delivers Weight Loss, A1C Improvements in Additional Trials. *Drug Topics*. 2026. https://www.drugtopics.com/view/retatrutide-delivers-weight-loss-a1c-improvements-in-additional-trials. Accessed 12 October 2026.

20. Eli Lilly and Company. ACHIEVE-4, the longest Phase 3 study of Lilly's Foundayo (orforglipron) to date, reaffirmed its cardiovascular and overall safety profile as well as consistent improvements across key measures of cardiometabolic health. Press release (PR Newswire), Indianapolis, 16 April 2026. Read via syndicated copy: https://www.placera.se/pressmeddelanden/lilly-eli-co-achieve-4-the-longest-phase-3-study-of-lilly-s-foundayo-

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

orforglipron-to-date-reaffirmed-its-cardiovascular-and-overall-safety-profile-as-well-as-consistent-improvements-across-key-measures-of-cardiometabolic-health-20260416. Accessed 12 October 2026. The sponsor's page (https://investor.lilly.com/news-releases/news-release-details/achieve-4-longest-phase-3-study-lillys-foundayo-orforglipron) returned HTTP 403.

21. Schoenthaler E. Lilly's Orforglipron Meets CV Safety Endpoint, Shows Reduced Mortality Risk in Phase 3 ACHIEVE-4 Trial. *BioPharm International*. 2026. https://www.biopharminternational.com/view/lilly-s-orforglipron-meets-cv-safety-endpoint-shows-reduced-mortality-risk-in-phase-3-achieve-4-trial. Accessed 12 October 2026.

22. Sundström J, Kristófi R, Östlund O, Bennet L, Eliasson B, Jansson S, et al. A registry-based randomised trial comparing an SGLT2 inhibitor and metformin as standard treatment of early stage type 2 diabetes (SMARTEST): Rationale, design and protocol. *Journal of Diabetes and its Complications*. 2021;35(10):107996. doi:10.1016/j.jdiacomp.2021.107996.

23. Eriksson JW, Fanni G, Lundqvist MH, Jansson S, Rådholm K, Sofizadeh S, et al. SGLT2 inhibitor or metformin as standard treatment in early-stage type 2 diabetes? Baseline data in SMARTEST, a novel, decentralised, register-based randomised trial on prevention of diabetic complications. *Diabetes, Obesity and Metabolism*. 2026;28(2):1327-1338 (online 28 November 2025). doi:10.1111/dom.70320.

24. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020;33. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Accessed 11 October 2026.

25. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large Language Models Struggle to Learn Long-Tail Knowledge. *Proceedings of the 40th International Conference on Machine Learning (PMLR 202)*. 2023;202:15696-15707. https://proceedings.mlr.press/v202/kandpal23a.html. Accessed 11 October 2026.

26. Cheng J, Marone M, Weller O, Lawrie D, Khashabi D, Van Durme B. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. *Conference on Language Modeling (COLM 2024)*. 2024. doi:10.48550/arXiv.2403.12958.

27. Liu Q, Jin Q, Menke JD, Kahnt T, Lu Z. Do AI chatbots find what experts would? Effects of model, user role, and sample size on study retrieval for medical questions. *arXiv (preprint)*. 2026:arXiv:2608.13786. https://arxiv.org/abs/2608.13786. Accessed 11 October 2026.

28. Vu T, Iyyer M, Wang X, Constant N, Wei J, Wei J, et al. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. *Findings of the Association for Computational Linguistics: ACL 2024*. 2024:13697-13720. doi:10.18653/v1/2024.findings-acl.813.

29. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking Retrieval-Augmented Generation for Medicine. *Findings of the Association for Computational Linguistics: ACL 2024*. 2024:6233-6251. doi:10.18653/v1/2024.findings-acl.372.

30. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac - Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*. 2024;1(2). doi:10.1056/aioa2300068.

31. Zhen T, Liu Y, Zhang G, Niu Y. What Do Chinese-Language Generative Search Engines Cite and Surface? A Large-Scale Empirical Study. *arXiv (preprint)*. 2026:arXiv:2607.15771. https://arxiv.org/abs/2607.15771. Accessed 11 October 2026.

32. Cen SH, Ilyas A, Driss H, Park C, Hopkins A, Podimata C, et al. Large-Scale, Longitudinal Study of Large Language Models During the 2024 US Election Season. *arXiv (preprint)*. 2025:arXiv:2509.18446. doi:10.48550/arXiv.2509.18446.

33. Ouyang S, Zhang JM, Harman M, Wang M. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Transactions on Software Engineering and Methodology*. 2025;34(2):1-28. doi:10.1145/3697010.

34. Atil B, Aykent S, Chittams A, Fu L, Passonneau RJ, Radcliffe E, et al. Non-Determinism of "Deterministic" LLM Settings. *arXiv (preprint)*. 2024:arXiv:2408.04667. doi:10.48550/arXiv.2408.04667.

35. He H, Thinking Machines Lab. Defeating Nondeterminism in LLM Inference. *Connectionism (Thinking Machines Lab blog)*. 2025. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/. Accessed 11 October 2026.

36. Chen L, Zaharia M, Zou J. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*. 2024;6(2). doi:10.1162/99608f92.5317da47.

37. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ*. 2024;384:e078538. doi:10.1136/bmj-2023-078538.

38. Bashah A, Salem A, Al-waqeerah A, Ghaleb E, Wahan N, Awad A, et al. Evaluation of deepseek, gemini, ChatGPT-4o, and perplexity in responding to salivary gland cancer. *BMC Oral Health*. 2025;25(1):1358. doi:10.1186/s12903-025-06726-4.

39. Fishkin R. NEW Research: AIs are highly inconsistent when recommending brands or products; marketers should take care when tracking AI visibility. *SparkToro blog (industry research with Gumshoe.ai; not peer reviewed)*. 2026. https://sparktoro.com/blog/new-research-ais-are-highly-inconsistent-when-recommending-brands-or-products-marketers-should-take-care-when-tracking-ai-visibility/. Accessed 11 October 2026.

40. Schulte J, Bleeker M, Kaufmann P. Don't Measure Once: Measuring Visibility in AI Search (GEO). *arXiv (preprint)*. 2026:arXiv:2604.07585. https://arxiv.org/abs/2604.07585. Accessed 11 October 2026.

41. Liu NF, Zhang T, Liang P. Evaluating Verifiability in Generative Search Engines. *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023:7001-7025. doi:10.18653/v1/2023.findings-emnlp.467.

42. Gao T, Yen H, Yu J, Chen D. Enabling Large Language Models to Generate Text with Citations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023:6465-6488. doi:10.18653/v1/2023.emnlp-main.398.

43. Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications*. 2025;16(1):3615. doi:10.1038/s41467-025-58551-6.

44. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5.

45. Jaźwińska K, Chandrasekar A. AI Search Has a Citation Problem. *Columbia Journalism Review (Tow Center for Digital Journalism)*. 2025. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php. Accessed 11 October 2026.

46. Bean AM, Payne RE, Parsons G, Kirk HR, Ciro J, Mosquera-Gómez R, et al. Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine*. 2026;32(2):609-615. doi:10.1038/s41591-025-04074-y.

47. Omar M, Agbareia R, McGreevy J, Zebrowski A, Ramaswamy A, Gorin M, et al. Ad-verse Effects: Pharmaceutical Advertising Shifts Drug Recommendations by Consumer-Facing AI. *medRxiv (preprint, not peer reviewed)*. 2026. doi:10.64898/2026.04.14.26350868.

48. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2.

49. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838.

50. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nature Medicine*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8.

51. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Network Open*. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.

52. Andrikyan W, Sametinger SM, Kosfeld F, Jung-Poppe L, Fromm MF, Maas R, et al. Artificial intelligence-powered chatbots in search engines: a cross-sectional study on the quality and risks of drug information for patients. *BMJ Quality & Safety*. 2025;34(2):100-109. doi:10.1136/bmjqs-2024-017476.

53. Tan SYT, Sng GGR, Lee PC. Accuracy of Large Language Model Responses Versus Internet Searches for Common Questions About Glucagon-Like Peptide-1 Receptor Agonist Therapy: Exploratory Simulation Study. *JMIR Formative Research*. 2025;9:e78289. doi:10.2196/78289.

54. Durgut O, Dikici O. Does ChatGPT update itself? Accuracy of ChatGPT in tympanostomy tube guidance: A comparative analysis with current literature. *European Archives of Oto-Rhino-Laryngology*. 2026;283(1):117-130 (online 23 August 2025). doi:10.1007/s00405-025-09630-3.

55. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large Language Models for Chatbot Health Advice Studies: A Systematic Review. *JAMA Network Open*. 2025;8(2):e2457879. doi:10.1001/jamanetworkopen.2024.57879.

56. The CHART Collaborative, Huo B, Collins GS, Chartash D, Thirunavukarasu AJ, Flanagin A, et al. Reporting Guideline for Chatbot Health Advice Studies: The CHART Statement. *JAMA Network Open*. 2025;8(8):e2530220. doi:10.1001/jamanetworkopen.2025.30220.

57. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*. 2025;31(1):60-69. doi:10.1038/s41591-024-03425-5.

58. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453-1457. doi:10.1016/S0140-6736(07)61602-X.

59. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Medicine*. 2007;4(10):e297. doi:10.1371/journal.pmed.0040297.

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

60. Aggarwal P, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K, Deshpande A. GEO: Generative Engine Optimization. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. 2024:5-16. doi:10.1145/3637528.3671900.

61. Kumar A, Lakkaraju H. Manipulating Large Language Models to Increase Product Visibility. *arXiv (preprint)*. 2024:arXiv:2404.07981. doi:10.48550/arXiv.2404.07981.

62. Nestaas F, Debenedetti E, Tramèr F. Adversarial Search Engine Optimization for Large Language Models. *International Conference on Learning Representations (ICLR 2025)*. 2025. doi:10.48550/arXiv.2406.18382.

63. Puerto H, Gubri M, Green T, Oh SJ, Yun S. C-SEO Bench: Does Conversational SEO Work? *NeurIPS 2025 Datasets and Benchmarks Track (accepted; arXiv v3)*. 2025:arXiv:2506.11097. doi:10.48550/arXiv.2506.11097.

64. Martinez O. Optimizing Visibility in Generative Engines: A Critical Survey of Generative Engine Optimization (2023-2026). *arXiv (preprint)*. 2026:arXiv:2607.14035. https://arxiv.org/abs/2607.14035. Accessed 11 October 2026.

65. Chen M, Wang X, Chen K, Koudas N. Generative Engine Optimization: How to Dominate AI Search. *arXiv (preprint)*. 2025:arXiv:2509.08919. doi:10.48550/arXiv.2509.08919.

66. Dai Y, Lurie E, Metaxa D, Friedler SA. Longitudinal Monitoring of LLM Content Moderation of Social Issues. *arXiv (preprint)*. 2025:arXiv:2510.01255. doi:10.48550/arXiv.2510.01255.

67. Metaxa D, Engler A. We Need Ongoing Monitoring of AI and Political Information. *Carnegie Endowment for International Peace*. 2026. https://carnegieendowment.org/research/2026/08/llms-artificial-intelligence-longitudinal-monitoring-norms-politics-research. Accessed 11 October 2026.

68. Choi Y, Nam HK, Cho B. Visibility of the Korean National Health Information Portal across search engines and generative artificial intelligence platforms: a repeated cross-sectional study. *Osong Public Health and Research Perspectives*. 2026. doi:10.24171/j.phrp.2026.0365.

69. U.S. Food and Drug Administration. TZIELD (teplizumab-mzwv) injection, for intravenous use: Prescribing Information (BLA 761183, supplements S-010/S-014). *Drugs@FDA*. 2026. https://www.accessdata.fda.gov/drugsatfda_docs/label/2026/761183s010lbls014lbl.pdf. Accessed 11 October 2026.

70. Novo Nordisk A/S (sponsor). A Research Study to See How Much CagriSema Lowers Blood Sugar and Body Weight Compared to Placebo in People With Type 2 Diabetes Treated With Diet and Exercise (REIMAGINE 1), NCT06323174. *ClinicalTrials.gov*. 2024. https://clinicaltrials.gov/study/NCT06323174. Accessed 11 October 2026.

71. Eli Lilly and Company (sponsor). A Study of Retatrutide (LY3437943) in Participants With Type 2 Diabetes Mellitus Who Have Obesity or Overweight (TRIUMPH-2), NCT05929079. *ClinicalTrials.gov*. 2023. https://clinicaltrials.gov/study/NCT05929079. Accessed 11 October 2026.

72. Uppsala University (sponsor). SGLT2 Inhibitor or Metformin as Standard Treatment of Early Stage Type 2 Diabetes (SMARTEST), NCT03982381. *ClinicalTrials.gov*. 2019. https://clinicaltrials.gov/study/NCT03982381. Accessed 11 October 2026.

73. Eli Lilly and Company (sponsor). A Study of Daily Oral Orforglipron (LY3502970) Compared With Insulin Glargine in Participants With Type 2 Diabetes and Obesity or Overweight at Increased Cardiovascular Risk (ACHIEVE-4), NCT05803421. *ClinicalTrials.gov*. 2023. https://clinicaltrials.gov/study/NCT05803421. Accessed 11 October 2026.

74. Eli Lilly and Company. Lilly's triple agonist, retatrutide, delivered substantial weight loss and A1C reduction, underscoring its potential promise for people with obesity and type 2 diabetes. Press release (PR Newswire), Indianapolis, 29 September 2026. Read via syndicated copy: https://www.placera.se/pressmeddelanden/lilly-s-triple-agonist-retatrutide-delivered-substantial-weight-loss-and-a1c-reduction-underscoring-its-potential-promise-for-people-with-obesity-and-type-2-diabetes-20260930. Accessed 11 October 2026. The sponsor's page (investor.lilly.com/node/54981) could not be opened (HTTP 403).

75. Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. *Quality & Safety in Health Care*. 2003;12(6):458-464. doi:10.1136/qhc.12.6.458.

76. Miller E. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. *arXiv (preprint)*. 2024:arXiv:2411.00640. doi:10.48550/arXiv.2411.00640.

77. Field CA, Welsh AH. Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 2007;69(3):369-390. doi:10.1111/j.1467-9868.2007.00593.x.

78. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*. 2011;104(12):510-520. doi:10.1258/jrsm.2011.110180.

79. Emirates Drug Establishment. Emirates Drug Establishment (EDE) portal and drug directory. *Emirates Drug Establishment*. 2026. https://www.ede.gov.ae/. Accessed 11 October 2026.

80. Health Sciences Authority (Singapore). Listing of Registered Therapeutic Products. *data.gov.sg (publisher: HSA)*. 2026. https://data.gov.sg/datasets/d_767279312753558cbf19d48344577084/view. Accessed 11 October 2026.

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

81. U.S. Food and Drug Administration. Drugs@FDA (openFDA drugsfda endpoint): overview. *openFDA*. 2026. https://open.fda.gov/apis/drug/drugsfda/. Accessed 12 October 2026.

82. European Medicines Agency. Download medicine data. *European Medicines Agency*. 2026. https://www.ema.europa.eu/en/medicines/download-medicine-data. Accessed 12 October 2026.

83. Saudi Food and Drug Authority. Drug Approvals. *Saudi Food and Drug Authority*. 2026. https://www.sfda.gov.sa/en/new-sfda-drug-approvals. Accessed 11 October 2026.

84. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960;20(1):37-46. doi:10.1177/001316446002000104.

85. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012;22(3):276-282. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/. Accessed 11 October 2026.

86. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*. 1990;43(6):543-549. doi:10.1016/0895-4356(90)90158-I.

87. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*. 2023;120(30):e2305016120. doi:10.1073/pnas.2305016120.

88. Panickssery A, Bowman SR, Feng S. LLM Evaluators Recognize and Favor Their Own Generations. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. 2024;37. doi:10.48550/arXiv.2404.13076.

89. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2018;115(11):2600-2606. doi:10.1073/pnas.1708274114.

90. Dal-Ré R, Ioannidis JP, Bracken MB, Buffler PA, Chan AW, Franco EL, et al. Making prospective registration of observational research a reality. *Science Translational Medicine*. 2014;6(224):224cm1. doi:10.1126/scitranslmed.3007513.

91. Lash TL, Vandenbroucke JP. Should preregistration of epidemiologic study protocols become compulsory? Reflections and a counterproposal. *Epidemiology*. 2012;23(2):184-188. doi:10.1097/EDE.0b013e318245c05b.

92. Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*. 1927;22(158):209-212. doi:10.1080/01621459.1927.10502953.

93. The diaTribe Foundation. Top Diabetes News: EASD 2026. *diaTribe*. 2026. https://diatribe.org/diabetes-research/top-diabetes-news-easd-2026. Accessed 12 October 2026.

94. European Medicines Agency. Zynquista (sotagliflozin): EPAR - medicine overview (withdrawn). *European Medicines Agency*. 2022. https://www.ema.europa.eu/en/medicines/human/EPAR/zynquista. Accessed 12 October 2026.

95. Holt RIG, DeVries JH, Hess-Fischl A, Hirsch IB, Kirkman MS, Klupa T, et al. The Management of Type 1 Diabetes in Adults. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*. 2021;44(11):2589-2625. doi:10.2337/dci21-0043.

96. European Medicines Agency. Teizeild (teplizumab): EPAR - medicine overview. *European Medicines Agency*. 2026. https://www.ema.europa.eu/en/medicines/human/EPAR/teizeild. Accessed 11 October 2026.

97. Kirsten E, Grosse Perdekamp J, Wu Q, Upadhyay M, Gummadi KP, Zafar MB. Characterizing Web Search in The Age of Generative AI. *arXiv (preprint)*. 2025:arXiv:2510.11560. doi:10.48550/arXiv.2510.11560.

98. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*. 2002;27(4):299-309. doi:10.1046/j.1365-2710.2002.00430.x.

99. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology*. 2017;46(1):348-355. doi:10.1093/ije/dyw098.

100. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ*. 2015;350:h2750. doi:10.1136/bmj.h2750.

101. Phillips DP, Kanter EJ, Bednarczyk B, Tastad PL. Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine*. 1991;325(16):1180-1183. doi:10.1056/NEJM199110173251620.

AlphaCitation publications and study records cited

A1. AlphaCitation. The answer moved before the evidence did. AlphaCitation Signal Note 04. Published 24 September 2026. https://alphacitation.com/blog/the-answer-moved-before-the-evidence-did. The study's published pre-congress baseline.

A2. AlphaCitation. EASD 2026 → AI answers: pre-registered protocol, with deviation log D0–D24. Registered at repository commit `8ee24e19`, 2 October 2026, 09:50:05 UTC. `docs/design/easd-2026-evidence-to-answer-prereg.md`.

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

A3. Ezeife OC. Technical Supplement: Does New Medical Evidence Change AI Treatment Answers? A Longitudinal Study of Evidence Citation, Medicine Naming and Selected Treatment Framing Around EASD 2026. AlphaCitation Research Series, Paper No. 01, version 1.0. AlphaCitation Research; October 2026. Cited in the text as "Supplement §".

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