

gpt-4o-mini Wrongly Suggested Wegovy Was Not Approved or Not Available in the UAE in 20 of 52 Answers to One Question, and in None of 56 Asked Other Ways

A pre-specified study of 124 answers over four days.

In a pre-specified UAE study, gpt-4o-mini gave a false approval or availability status in 20 of 52 answers to one buyer's choice question and its eight paraphrases, none naming the medicine's form (one asks which weight-loss injection to choose), and in none of 56 answers to 14 other wordings about the same medicine, 16 of them the same buyer's choice with the form named.

Ozue C. Ezeife · AlphaCitation

The finding. OpenAI's gpt-4o-mini was asked, in a UAE buyer's voice, to choose between Wegovy (semaglutide), its form unspecified, and Mounjaro (tirzepatide), the competitor named in the original question. In 20 of 52 answers, as the study coded them, it said or implied that the medicine was not approved or not available in the UAE. The regulator, the Emirates Drug Establishment (EDE), announced on 1 June 2026 that it had approved the medicine as "an oral once-daily treatment". A Dubai Health Authority document dated 3 September 2024 states that the medicine was "registered for obesity pharmacotherapy in UAE"; it does not name a form. Both contradict an answer that the medicine, its form unspecified, is not approved: the announcement as of June 2026, the Dubai Health Authority document as of September 2024. In 19 of the 20 false answers, the answer gave the false status as its reason for recommending the competitor. The twentieth called both products unapproved.

Asked in 14 other ways, the same model made the claim in none of 56 answers. Forty of the 56 still set the medicine against the same competitor: 24 in a clinician's voice or neutrally, 16 in the same buyer's voice with the tablet or the injection named. A negative control, the same buyer's choice against another brand name of the competitor's own molecule, gave 0 of 8.

The zero is an absence of false claims, not proof that the model knew the status. Of the 56 answers, 10 hedged and 8 made no status statement. Of the 8 answers to questions that named the injection, all coded correct, 4 gave a status only for the tablet.

What was tested. The model was gpt-4o-mini, served as gpt-4o-mini-2024-07-18, queried through OpenAI's Responses API with its web search tool at temperature 0. It is not the consumer ChatGPT app, which runs other models and about which this note makes no claim. Between 13 and 16 October 2026 we put 27 question wordings, sealed and hashed before the first call, to the model 124 times. They are a fixed battery, not what any person asked.

Two errors, two standards of evidence. The study counts approval (class A) and availability (class B) separately, because a regulator records registration, not supply.

- *Approval: 14 of the 20 answers.*
 - Three say, in the present tense, that the medicine is not approved or not registered. The regulator's 1 June announcement contradicts them outright. A fourth does too if read literally, but it answers a question about injections.
 - Seven of the fourteen make their approval claim only by saying the medicine is "not listed" among approved medications. They are false only if that phrase is read as implying non-approval: the study's reading, and disputed.
 - One more says the medicine is "not listed in the UAE's official drug registries". It rests on the same implication but is not on the study's disputed list.
 - The last two say "as of October 2023" that the medicine has not been approved; only the injection's registration on 14 November 2022 contradicts that dated sentence; that date rests on records of the regulator's directory made on 2 and 5 October and on a search-indexed copy of its record seen on 13 October; on 19 October the directory returned no entry for either form (record (c)). Each then concludes, with no date, "Therefore, it is not legally available" (one adds, in its conclusion, "at this time"). The study coded that sentence as part of the approval claim; read so, and as a statement about the time of the answer, the 1 June announcement contradicts it.

Availability: 6 of the 20 answers say only that the medicine is not available. The regulator's records cannot test that. These six are judged against one licensed UAE pharmacy chain's listings, which showed both forms for sale on 13 and 16 October. That is secondary evidence. Without it, only the 14 approval answers count as false.

Two different sensitivity tests apply, and they are not one range. The first is the evidence standard above: without the pharmacy listings, 14 of 52. The second is the codes still in dispute: the false total is 19 to 21 of 52, against 20 as coded. It is 19 if one original-question answer, coded "not approved" on the disputed phrase alone, is read as making no false claim. It is 21 if one hedged answer that recommends the competitor as more accessible is read as an availability claim.

The cited source said the opposite. Of 48 false status sentences, 9 carried a citation. Eight cite one of two pages of glpguidebook.com; both state that the UAE has the medicine as a pen and as a tablet. The site's domain was registered on 12 October 2026 at 20:56 UTC, 21 hours 24 minutes before the study's plan was frozen. Its pages first appeared among the sources returned with the model's answers on 15 October: in 0 of 62 answers on the first two days and in 53 of 62 on the last two. The record shows timing, not cause. The author's statement on the site is under Disclosures.

What was fixed in advance. The question wordings, the coding classes, the explanations and the gates were fixed and hashed before the first call. The availability evidence standard, the page-text analyses and the founder’s hypothesis were added after the first day and before the second; the website findings, the side-by-side reading, the evidence tiers and the second coding after collection.

What the rules decided. By the gates fixed before collection, the result is descriptive: not GO, not NO-GO. The false status is real and recurs across paraphrases, but it did not appear outside one question and its eight paraphrases, and the data identify no content change specific enough to test. That is a screening outcome, not a finding that the error cannot be fixed.

The false status came from one question and its paraphrases, and from none of the reframings

The earlier observation. On 5 October we reported that one fixed head-to-head question about a weight-management medicine registered in the UAE had led the engine we measure to rule the medicine out in 10 of 13 runs.

- The question was asked in the voice of a buyer in the UAE, a frame that invites a judgement on whether a medicine can be obtained there and a preference for one that can.
- Six answers said the medicine was not approved. Field Note 06 scored them against the regulator’s directory as read on 5 October; a reference note of 2 October records the same registration dates, and on 19 October the directory returned no entry for the medicine (record (c) below). Two of the six reasoned from a knowledge cutoff of October 2023; record (c)’s registration date of 14 November 2022 makes them wrong even on that premise. Each also concluded, with no date, “Given that Wegovy is not approved for use in the UAE”, which the 1 June 2026 announcement contradicts as a statement made in September 2026.
- Four said it was not available, which the register cannot test.

That report was “The Medicine Was Approved. ChatGPT Ruled It Out Anyway.” (Field Note 06 [1]). It says the wrong version of the fact won “when the question became a choice”, and that sentence is still on the page. The data below narrow it to one question and its paraphrases: 40 other choice answers about the medicine, 32 against the same competitor and 8 in the negative control, carried no false status.

The engine. It is OpenAI’s gpt-4o-mini, queried through OpenAI’s API with its web search tool; Field Note 06 called it ChatGPT.

The follow-up. The 27 sealed wordings are:

- the original question and eight paraphrases of it;
- 14 reframings in five families;
- a negative control, the medicine against a different product (another brand name of the competitor’s own molecule, whose UAE status the study did not code). It is a question the engine had answered correctly in production;
- a candidate control: a tablet from the competitor’s manufacturer, registered in the UAE on 2 April 2026 according to the directory as read on 2 and 5 October (record (c)).

Exhibit 1 and the Method table give each family’s size and schedule.

Exhibit 1. The false UAE status statement, by how the question was asked (124 answers, 13–16 October 2026)

HOW THE QUESTION WAS ASKED	ANSWERS	FALSE: NOT APPROVED OR NOT AVAILABLE	CORRECT	HEDGED	NO STATUS STATEMENT
UAE buyer’s voice: choose between the medicine, form unspecified, and the competitor; original question (S)	4	3	1	0	0
The same, eight paraphrases (H)	48	17	29	2	0
Clinician’s voice, UAE: the medicine and the competitor (X1)	12	0	1	3	8
Neutral, UAE named: the medicine and the competitor (X2)	12	0	5	7	0
No competitor: buyer’s voice (two wordings) or neutral (one) (X3)	12	0	12	0	0
Tablet named: against the competitor in a buyer’s voice (two wordings) or alone (one) (X4)	12	0	12	0	0
Injection named: against the competitor in a buyer’s voice (X5)	8	0	8	0	0
Negative control: the medicine against a different product, buyer’s voice (N)	8	0	8	0	0
Candidate control: the competitor’s manufacturer’s tablet against the competitor (C; not about the medicine)	8	0	7	1	0

Source: per-answer codes, reproduced by the study's tally script. One answer per call; gpt-4o-mini with web search, temperature 0. Four of the eight X5 "correct" answers gave only the tablet's status; read as not addressed, X5 correct is 4.

The contrast.

- The original question and its paraphrases carried all 20 false answers: 20 of 52.
- The 14 reframings carried none: 0 of 56. That includes 0 of the 40 answers that still set the medicine against the same competitor: 32 asking which to choose, 8 asking how the two compare.
- The negative control carried 0 of 8.

Read by wording, the contrast holds. Repeated draws of one wording at temperature 0 are not independent, so the wording is the more honest unit, though wordings asked on the same day also share a search environment.

- 8 of the 9 wordings of the buyer's choice against the competitor with the form unspecified produced at least one false answer. The original question was asked 4 times, each paraphrase 6 times.
- None of the 4 buyer's-voice wordings against the competitor that named the tablet or the injection did (4 answers each).
- None of the 6 clinician or neutral wordings against the competitor did.
- None of the 4 wordings without a competitor did.
- Neither of the two negative-control wordings did.

The zero is an absence of false claims, and partly an absence of claims.

- 18 of the 56 reframing answers made no correct claim either: 10 hedged and 8 made no UAE status statement.
- The clinician questions mostly left UAE status unaddressed (8 of 12 answers), and the neutral questions mostly hedged (7 of 12).
- Four of the 8 injection-named answers coded correct gave a status only for the tablet, although the question named the injection. Each opens with a sentence, form unspecified, saying the medicine is approved or available, and that sentence is why each was coded correct. Their identifiers (the last eight characters of the API response id) are ff483e73, c6d05254, 63728578 and 5e740132. The study lists all four for review.
- "Correct" was often correct only on the point coded. 28 of the 76 correct answers about the medicine carried another wrong regulatory fact. Most often it was the June 2026 date of the regulator's announcement approving the oral form, given to the medicine in general or to the injection.

In 19 of the 20 false answers, the false status was the stated reason for the recommendation

Each of the 19 gave the medicine's unavailability or non-approval as its reason for recommending the competitor. The twentieth called both products unapproved. In the production record behind Field Note 06, the answers used words such as "the available and legally prescribed option" [1]. The original question drew a false answer on 3 of the 4 days.

Approval and availability claims meet different standards of evidence

The medicine is Wegovy (semaglutide), sold in the UAE as a weekly injection and, since 2026, a daily tablet. The competitor named in the original question is Mounjaro (tirzepatide).

The government records, in order of strength

(a) The regulator's own announcement, 1 June 2026. The Emirates Drug Establishment's news page "Emirates Drug Establishment approves oral Wegovy® (semaglutide) for weight management and cardiovascular risk reduction" is dated Monday, 01 June 2026 [10]. It states that EDE "has approved Wegovy® (semaglutide), an oral once-daily treatment indicated for long-term weight management", and that the UAE is the second country to authorise it. It gives no registration number or date, and it does not mention the injection. We read it on 5 October 2026, for Field Note 06, and on 16 and 19 October 2026.

(b) The Dubai Health Authority, 3 September 2024. The authority's EJADA Program document "Obesity and Metabolic Syndrome KPIs and Recommendations", code DHA/DHIC/HEIPD/CG-12, Version 2, issue date 3 September 2024 [12], states on page 9, in a box headed "Simeglutide (Wegovy)" (misspelling included) beside a parallel box for the competitor: "contains semaglutide and should not be used with other semaglutide-containing products or other GLP-1 receptor agonist medicine Approved by FDA and registered for obesity pharmacotherapy in UAE". The passage names no form, and we do not infer one: the document is used only against answers that leave the form unspecified. It is the Dubai Health Authority's statement, in a clinical-recommendations document, that the medicine was registered for obesity pharmacotherapy in the UAE; it is not a register entry. The file served on 19 October was last modified on 21 July 2026; an earlier edition of the same document, created in January 2024 and captured by the Wayback Machine at the same address on 25 May 2024, carries the same statement on its page 8. It does not establish the date of registration. We found the document through glpguidebook.com's own source list and read it at source on 19 October 2026.

(c) **The regulator's directory, as recorded on 2 and 5 October 2026.** The study's frozen truth for each product's registration was the EDE online directory: the injection "Active", registered 14 November 2022; the tablet "Active", registered 4 May 2026; the control tablet "Active", registered 2 April 2026.

- These values rest on a transcription of a 5 October directory read, made in an AI-assisted session; on two records of 2 October published with Paper No. 01's evidence [2], a scripted read that recorded the control tablet's six strengths as "Active", registered 2 April 2026, and a reference note, which does not record how the directory was read, that the directory listed the injection as registered on 14 November 2022 and the tablets on 4 May 2026; and on a search-indexed copy of the injection's directory record seen on 13 October. No directory page from those days was preserved.
- On 11–13 October the directory served an eight-row placeholder. The Wayback Machine's capture of the directory's front page at 23:16 UTC on 5 October shows no products; whether the page loads results only after a search is not known.
- Read again on 19 October 2026 (the page states "Last dataset refresh: 19 Oct 2026"), it returned no entry for either form of the medicine or for the competitor, searched by product name and by active ingredient, and none for the control tablet by name.
- The directory was visibly incomplete that day. It held 1,179 products, with names running almost entirely from digits to "A". A search for "sodium chloride" returned 52 results; Panadol returned none.
- On 19 October none of these values could be re-read from the regulator.

Approval (class A): 14 answers, in four evidence tiers

Exhibit 2. The 14 "not approved" answers, by the evidence that makes them false (added analysis, 19 October; the last row's undated sentence read against the 1 June announcement on 21 October)

TIER	WHAT THE ANSWER SAYS	EVIDENCE THAT IT IS FALSE	ANSWERS
Contradicted outright	In the present tense, not approved or not registered, form unspecified	(a), the 1 June 2026 announcement. Two of the three attribute the claim to the Dubai Health Authority's 2024 obesity document, which (b) shows says the opposite	3: 6bae64c1, 4ae1b9da, 9b4cae8a
Contradicted if read literally	"not approved or sold in the UAE", in an answer to a question about injections	Read literally: (a); (b) as of September 2024. Read in its injection frame: (c) alone, which could not be re-read on 19 October; (b) names no form	1: 826dbdd1
False only under the disputed reading	"not listed among" approved medications	False only if absence from a list is read as implying non-approval. The study lists all seven as disputed	7: 6f695f7a, b659248b, 2f3280d3, 8eaaff38, 60043ff3, 31f02383, 8125ea7b
	"not listed in the UAE's official drug registries"	The same kind of implication. The study did not list it as disputed; a reviewer may place it with the seven	1: b789ad31
False as dated only against the directory record (c); the undated sentence against (a)	"As of October 2023", not approved by the Ministry of Health and Prevention; then, undated, "not legally available" (cb7c600b adds "in the UAE at this time")	The dated sentence: (c) alone; neither (a) nor (b) speaks to October 2023. The undated sentence, which the study coded as part of the approval claim: (a), read as a statement about the time of the answer	2: 50759de4, cb7c600b
Total			14

Source: per-answer codes; the 1 June announcement and the Dubai Health Authority document, read 19 October 2026. Identifiers are the last eight characters of the API response id. No false answer names a form: all 20 make their claim about the medicine with no form named.

What the tiers mean.

- The regulator's own statement contradicts three approval answers outright, or four if one is read literally.
- Seven of the fourteen rest on the disputed reading of one phrase: five paraphrase answers and two original-question answers. One of the two is coded "not approved" and "not available"; the other "not approved" only.
- The regulator's directory could not test a claim about a list literally: it held a placeholder when read on 11–13 October, and on 19 October it returned no entry for either form.
- The two answers dated "As of October 2023", 50759de4 and cb7c600b, share their status passage and must be treated alike. Each follows the dated sentence with an undated one saying the medicine "is not legally available"; cb7c600b repeats it in its conclusion, "at this time". The study coded both "not approved" only and lists only cb7c600b as disputed on whether it should also count as "not available". A reviewer who decides one must decide the other the same way. If both also count as "not available", each stays false on the pharmacy evidence even if record (c) is set aside. If both stay "not approved" only, the dated sentence rests on record (c) alone, and the undated sentence, as the study coded it, on the 1 June announcement, read as a statement about the time of the answer.

Availability (class B): 14 answers, one standard

Fourteen answers say the medicine is not available: six only that, eight as well as “not approved”. Thirteen concern the medicine with no form named; the fourteenth is the answer framed around injections. The regulator’s records cannot test them.

- Each is coded false against one licensed UAE pharmacy chain’s listings, read on 13 and 16 October, which showed the injection and the tablet for sale. Both reads fall within three days of every availability answer.
- A retailer listing is secondary evidence, not proof of stock on a shelf.
- If it is not accepted, only the approval answers count as false: 14 of 52 rather than 20, and 12 of 48 paraphrases rather than 17. The 56 reframing answers and the 8 control answers stay at 0.

The false total is 19 to 21 of 52 under the codes still in dispute

- **Lower bound, 19.** One original-question answer (b659248b) is coded “not approved” only, on the disputed phrase, and the study’s list gives no alternative class for it. Read as making no false claim, it leaves 19. The other six phrase codes stay false under that reading: five are also coded “not available”, and the sixth (31f02383, “not listed among the medications approved for sale”) becomes an availability claim, the alternative the study’s list gives for it.
- **Upper bound, 21.** One paraphrase answer (0d1fae30) is coded hedged: it calls the medicine’s status “not well-documented” and then recommends the competitor as more accessible. Read as an availability claim, it makes 21.
- **As coded, 20.** The range holds both evidence standards fixed. The approval count runs from 7 of 52 under the alternative phrase reading to 14 as coded; “seven of the fourteen” is the disputed share.

A second model coder reproduced every false-or-not decision. That shows the rules are consistent, not that they are right.

On 19 October the 52 original-question and paraphrase answers were re-coded blind by a second model coder, not by independent human adjudication. The second coder is from the same model family that assisted the first coding. It saw the question and answer texts in shuffled order, with the written rules, and no codes.

MEASURE (52 ANSWERS)	AGREEMENT	COHEN’S KAPPA
False against not false	52 of 52	1.00
Full code (not approved; not available; both; correct; hedged)	50 of 52 (96.2%)	0.937
“Not approved” present	52 of 52	1.00
“Not available” present	50 of 52 (96.2%)	0.907

- The two disagreements are 50759de4 and cb7c600b, the “not legally available” pair: coded “not approved” only by the study and both “not approved” and “not available” by the second coder. The false total and the approval count are unchanged either way.
- The rules themselves state the disputed reading: absence from a list of approved medicines implies “not approved”. A coder who follows them must reproduce the seven disputed codes. The agreement therefore shows that the written rules are applied consistently. It does not show that the disputed reading is correct.
- Agreement on false against not false is expected wherever the rules are clear: the 20 false answers carry explicit negative status sentences and the 32 others do not. The informative cells are the disputed ones.
- No qualified regulatory or medical-information professional has yet reviewed the 21 codes the study lists as disputed.

By the pre-specified rules, the result is descriptive

The plan, the collection script and the 27 sealed wordings were fixed and hashed at 18:21:01 UTC on 13 October, before the first call. On 16 October the plan, the script, the addendum and the sealed wordings still matched their recorded hashes. There were no deviations. The gates below were evaluated exactly as frozen.

GATE, AS FROZEN	RESULT
G1 Persistence: “not approved” in at least 25% of the 48 paraphrase answers	Passed exactly at the threshold: 12 of 48 (Wilson 95% 15–39%). One disputed code would make it conditional; so would the alternative phrase reading.
G2 Generalisation: “not approved” in at least 4 of the 8 paraphrases, and in at least 2 answers within one of the clinician, neutral or injection-named families	Failed. First limb met (6 of 8 wordings). Second limb: 0 of 32 answers, and none of the 32 is false in any class.
G3 Control: a control with an established truth, its own error rate of at least 25%, the same mechanism signature and no shared content	None qualified. The competitor’s manufacturer’s tablet: 0 of 8.

GATE, AS FROZEN	RESULT
G4 Actionable pathway: a supported information pathway that someone could lawfully change	Not passed (a judgement, flagged as such). The pathway that holds, “retrieved but not used”, concerns information already present in what the engine retrieved: a page on the pre-specified list in 19 of the 20 false answers.

Which evidence each gate uses.

- G1, G2 and G3 rest on the answer codes alone, assigned under the classes fixed before collection.
- G4’s first limb is counted two ways in the study’s gate table: a page on the pre-specified list in 19 of the 20 false answers and, by the page text added after day 1, 20 of 20. The study’s finding that two other pathways (no status page found; an intermediary source) hold for none also rests on page text: by the list alone, 1 of the 20 false answers retrieved no listed page. The row above gives the pre-specified count only.
- The study’s written G4 judgement also cites page-text counts (all 20 false answers retrieved a page stating the injection’s status).
- The decision does not depend on G4: G2 decides.

Decision: descriptive. Not GO, not NO-GO. The rule required G1, G2 and G4 for a GO, and a G1 failure for a NO-GO. G2’s failure decides it, under every reading of every disputed code.

What “descriptive” means here. The false statement is real, recurrent, and spreads across paraphrases of one question: the buyer’s choice against the competitor, the form unspecified. It does not appear outside that question, including in the same choice against a different product (0 of 8). The data identify no content change specific enough to test. This is a screening outcome. It does **not** show that clearer wording, disambiguation or another change to sources could not help.

Stage 1 was the qualification step for a planned, separately pre-specified test of whether a correction to public sources reduces the error. Under its rules, this case does not proceed to that test.

The pre-specified status-page list. Before collection, the plan fixed a list of pages known to state the medicine’s UAE status.

- A page on that list, the manufacturer’s release on the tablet’s UAE launch, was retrieved in 19 of the 20 false answers and cited in none.
- Correct answers to the S, H and X questions retrieved a listed page in 54 of 68 and cited one in 32 of 68.
- This is descriptive: retrieval was an outcome of each question, not assigned.

Every false answer’s search returned the address of a page stating the medicine’s status (added analysis)

Everything in this section and the next three comes from analyses added after the first day’s answers and before the second day’s (hashed 18:34:29 UTC, 13 October), or from patterns found after collection. None of it was pre-specified. The decision does not depend on it, because G2 decides. The study’s G4 judgement also cites the page-text counts; the G4 row above rests on the pre-specified list alone.

What the record can see. For each answer, OpenAI’s API returned the addresses it reported as sources for each search and the addresses the answer cited. It returned no title, snippet or page text. After the fourth day we fetched all 82 distinct pages retrieved or cited in the 20 false answers. We read every page that yielded text: 62 of 82. Twelve failed to load, and eight loaded with under 1,000 characters of text.

What it found, at the level of page addresses, for the pages as read on 16 October:

- **All 20 false answers** had retrieved at least one page whose text says the medicine is approved or sold in the UAE.
- **All 20** had retrieved one that states the status of the weekly injection, the form the added formulation-match analysis assigned to questions that leave the form unspecified.
- **None** of the 62 readable pages states the false claim. Every false answer has at least one unreadable page among its results.
- The manufacturer’s release on the tablet’s UAE launch was among them in 19 of the 20. Its newswire copy could not be read; a verbatim reprint was.

Where the false sentence carried a citation, the cited page stated the opposite in 8 of 9.

- Of 48 false status sentences in the 20 answers, 9 carried a citation, one in each of 9 answers.
- Seven cite glpguidebook.com's page on whether Zepbound, the product in the negative control, is available in the UAE. That page states that "the UAE has Wegovy (semaglutide) as a weekly pen and a daily pill".
- The eighth cites the same site's page on generic semaglutide. That page says the semaglutide sold in the UAE includes "Ozempic and Wegovy pens, and Rybelsus and Wegovy tablets".
- Two of the eight also attribute the claim in words to the Dubai Health Authority's 2024 obesity document. The page quotes that document only for the competitor ("registered for obesity pharmacotherapy in UAE"). Its sentence that the UAE has the medicine as a weekly pen and a daily pill cites the same document among its three sources. Its source list, present in our 16 October copy, identifies the document as DHA/DHIC/HEIPD/CG-12, version 2, at the address in [12], and says it lists both. That document names the medicine on page 9 as "registered for obesity pharmacotherapy in UAE" (record (b) above). The two answers contradict the document they name.
- The ninth cites a consumer health website that states no UAE status, on a sentence that also gives an efficacy figure.

The primary records were never retrieved. Neither the medicine's UAE prescribing information nor any record from EDE's drug directory appeared in any of the 124 answers, correct or false.

- The only regulator page returned anywhere was EDE's announcement of the oral form's approval.
- It was retrieved in 9 correct S, H and X answers, all nine to questions that named no competitor, or 10 with the negative control; it was cited in each of them.
- It was retrieved in none of the 92 answers that set the medicine against the competitor, false or correct.

Retrieving the release did not protect the answer. Among the 52 answers to the buyer's choice against the competitor with the form unspecified, 19 of the 43 that retrieved a listed status page were false, against 1 of the 9 that did not. In all 43 the page was the tablet release. That comparison is confounded: particular search query strings bring up the release.

The source cited for 8 of the 9 cited false sentences is a site registered on 12 October 2026, the day before collection began

The two cited pages, and our copies of them.

PAGE	CITED FOR FALSE SENTENCES	OUR COPY, 16 OCT 2026 (SHA256)	OUR COPY, 19 OCT 2026 (SHA256)
https://glpguidebook.com/ae/answers/zepbound-available-uae	7	09:55:42 UTC, b43a84f7b10a2fd6afd5e4a73fad2a2f6a7a2b956bb1e89e04c56b454d1758cd	17:56:20 UTC, 535d993c84f5ef3290d3b54ed1ef
https://glpguidebook.com/ae/answers/generic-semaglutide-available-uae	1	09:55:38 UTC, 5874e22983b1e06e1209e00ccf8ad05541f4fd30c59c24250ea1a3fae930cc10	17:56:23 UTC, acc078f930c95694767ec0457490

- Our 16 October copies were fetched about 12 minutes after the study's last answer. They are the closest record we hold of the pages as they stood during the study.
- Both pages carry "Updated 12 Oct 2026". The statements quoted above appear in both our copies; between them the pages gained a navigation bar and a "What we checked" summary box; the source list is in both copies.
- The site has changed since 16 October in other places: a comparison page cited in three correct answers redirected to a new address on 19 October, a price post changed its update frequency, and the privacy page records changes on 13 and 16 October.
- When queried at 18:19 UTC on 19 October, the Wayback Machine's capture index listed no capture of any page on the domain. From 18:57 UTC on 19 October, AlphaCitation submitted to it the home page, all ten of the site's addresses that appear in the study's answers, and the site's corrections and method pages. The archive returned a capture address for each, including the two in the table above. These captures record the site as of 19 October, three days after the study, not as it was served during it.

Registration. The domain's public registration record (RDAP), read on 19 October 2026 from the .com registry and from the registrar, gives a registration time of 2026-10-12T20:56:55Z, through Cloudflare, Inc. That is 21 hours 24 minutes before the plan was frozen. The registrant's details are redacted in the registrar's record and absent from the registry's. The registrar's record gives 19 October 2026, 08:52:34 UTC both as its "last changed" date and as the date its RDAP record was last updated; the registry's record gives the registration time as "last changed". The records do not show that the registration itself changed.

Who runs it. None of the 14 pages of the site we read on 19 October names a publisher, company or person. The site describes itself in its own words: "We're an independent team. We take no drug-company money", and "An independent guide: no drug-company money, no clinic commissions." That is the site's self-description. AlphaCitation has not verified it, and the operators are unnamed. The privacy page names the data controller only as "GLP Guidebook" and gives a contact address it says "isn't live yet" (read 19 October).

Arrival. The site's pages were among the sources returned with 0 of 62 answers on days 1–2 (13–14 October) and 53 of 62 on days 3–4 (22 of 27 on 15 October, 31 of 35 on 16 October). The first answer to return one was a negative-control answer at 10:06:53 UTC on 15 October, 2 days 13 hours after the domain was registered. On days 3–4 the site was among the sources of all 26 answers to the original question and its

paraphrases.

Citation. On days 3–4 it was the source cited for all 8 cited false sentences: four on 15 October, four on 16 October. The ninth cited false sentence, on day 2, cites another site.

What the record does and does not show.

- It shows timing: the site appeared among the sources from day 3, and the false sentences that cite it fall on days 3–4.
- It does not show why the site was absent on days 1–2, or that the site caused any answer.
- The two cited pages state the medicine's UAE status correctly. Each of the eight false sentences cites a page that states the medicine is in the UAE.
- Nothing in the record shows who operates the site or why it was created.

A candidate mechanism: statements about other products, with the medicine as their subject

This is a candidate, found after collection, not an established mechanism.

How it was compared. The eight cited false sentences were read side by side against the stored pages by a single coder, model-assisted. They were re-read in the pre-publication audits. No automated overlap test was run, and no comparison method was recorded beforehand.

What the reading shows. Seven of the eight cite the Zepbound page:

- **One** restates, almost word for word and with the medicine as its subject, the page's one-line answer about Zepbound.
- **Four** say the medicine is not listed among approved medicines or official registries. The page says it found no UAE registration or official drug-list entry for Zepbound.
- **Two** say that the Dubai Health Authority's document does not mention the medicine. The page quotes that document only for the competitor, although its sentence on the medicine cites it and its source list says it lists both. The document names the medicine on page 9.
- **The eighth** cites the page on generic semaglutide. It parallels that page's statement that no generic version is approved in the UAE.

If the reading is right:

- in six of the eight, a negative statement the page makes about another product appears with the medicine as its subject;
- in two, the page's statement that the health authority's document lists the competitor appears as a statement that the document does not mention the medicine.

The record cannot show what the model received, and it does not show that the model transferred anything.

Over the same days the regulatory share rose, by a count that depends on disputed codes.

- Answers coded "not approved" went from 3 of 24 paraphrase answers on days 1–2 to 9 of 24 on days 3–4.
- All five disputed paraphrase codes fall on days 3–4. Read the other way, the change is 3 of 24 to 4 of 24.
- False answers of either kind went from 8 of 24 to 9 of 24.
- Knowledge-cutoff language (reasoning from a stated cutoff, such as October 2023) disappeared from false answers: 5 on days 1–2, none on days 3–4.

These changes coincided in time with the site's arrival. That is all the record shows: the site was not assigned to any answer, and the wordings and draws differ by day.

The founder's hypothesis holds at the level of page addresses and is undetermined beyond it

The hypothesis, added after day 1, was that this is an information-to-decision failure rather than a visibility failure. The addendum fixed how to test it before day 2.

- **At the level of page addresses: supported.** Status material, including the injection's, was retrieved in 20 of 20 false answers, and 19 of the 20 exclude the medicine on the false premise.
- **At the level of what the model received: undetermined.** The record cannot show that any status passage reached the model in any of the 20.
- **The 8 cited answers are more consistent with a different reading:** the model received what the page says about other products and misapplied it to the medicine (six and two, as set out above). That is a reading failure, not an override.

The data rule out a visibility failure at the level of page addresses, for the pages as read on 16 October. They cannot separate the other two readings.

Three of the seven explanations named before any data are supported. None is established as the mechanism

VERDICT (PLAN'S TERMS)	EXPLANATION	COUNT (OVER THE 20 FALSE ANSWERS)	WHAT IT DOES AND DOES NOT SHOW
Supported	The model's own assertion	No false sentence traces to a readable page that states it (page text, read 16 October); 11 of 20 carry no citation on any false sentence	No supporting statement found in readable pages; origin unresolved
Supported (page addresses)	Retrieved but not used	A listed page retrieved in 19 of 20 and cited in none; by page text (added), 20 of 20	Correct-status page addresses were retrieved; whether the model used their text is unresolved
Supported	The question's frame	20 of 52 against 0 of 56	A strong association with this question family; the mechanism is not established
Not supported	Retrieval failed to find status pages	0 of 20 by page text (added); 1 of 20 lacked a page on the pre-specified list	Not observed at the level of page addresses
Not supported	An intermediary source stated the falsehood	0 of 20, for readable pages (page text, read 16 October)	Not observed among readable sources
Not supported	Injection and tablet confused, as a cause	No false sentence uses tablet-era facts such as "awaiting approval" or June 2026; the tablet-named and injection-named families do not differ	No sign of it in the false sentences or the family contrast
Not supported	The model's search queries went astray	In all 20 false and all 76 correct answers about the medicine, the model's queries named both the medicine and the UAE; 18 of 20 false answers used a query string that also produced a correct answer	No distinctive failure in query terms found

Neutrality: both products are held to the same record

- In 19 of the 20 false answers, the competitor was recommended on the false premise. The site most often cited for the false claim states the medicine's UAE status correctly.
- The same model also called the competitor unavailable or unapproved in the UAE in 6 answers. The study did not code the competitor's status, so those six are reported as statements, not errors. The Dubai Health Authority page that lists the medicine as registered lists the competitor in the same terms ("registered for obesity pharmacotherapy in UAE", September 2024). On 19 October the regulator's directory returned no entry for the competitor either.
- The competitor's manufacturer's tablet served as a control: the model stated its status correctly in 7 of 8 answers and hedged once.

What this does and does not show

- **Observed behaviour, not internal cause.** The note describes what one endpoint answered. It does not show why. Retrieval was an outcome of each question, not assigned. No clinician or neutral answer retrieved a listed status page, so the frame and retrieval cannot be told apart.
- **Retrieved, not received.** "Retrieved" means the search returned a page's address. Pages were read on 16 October, up to three days after the answers, and may have changed. 20 of the 82 pages in false answers could not be read, and every false answer has at least one. "No retrieved page states the false claim" holds for readable pages only. The cited site has changed since 16 October, and the Wayback Machine's capture index listed no earlier copy when queried on 19 October. Our 16 October copies are the closest record we hold of the versions served during the study.
- **Approval, not availability.** The regulator records registration, not supply.
 - The approval of the oral form is stated in the regulator's own announcement of 1 June 2026 [10], live on 19 October.
 - Registration by September 2024, with no form named, is stated in a Dubai Health Authority document [12]. The injection's registration, and its date of 14 November 2022, rest on the directory as read on 5 October and transcribed in an AI-assisted session, on a reference note of 2 October that does not record how the directory was read, and on the search-indexed record. Read again on 19 October, the live directory, refreshed that day, held a partial dataset of 1,179 products and returned no entry for either form of the medicine or for the competitor.
 - "Not available" is coded false against one licensed pharmacy chain's listings, which is secondary evidence.
- **A fixed battery, not real users.** These are 27 sealed wordings asked through the API. They are not the consumer ChatGPT app, and not what any person asked or chose.

- **Small and narrow.**
 - One model (gpt-4o-mini-2024-07-18, served on all 124 calls), one provider, one market, four days.
 - The measured questions concern one medicine and one competitor; each control adds one product.
 - The interval for 20 of 52 (26–52%) treats repeated draws as independent and is too narrow.
 - Coding was model-assisted (Claude, under the founder’s direction), by a single coder, with every answer read in full. A second model coder re-coded the 52 original-question and paraphrase answers blind; the 56 reframing answers behind the 0 of 56, and the 16 control answers, were not re-coded.
- **Disputed codes.** 21 codes are listed for review by a qualified regulatory or medical-information professional; none has been reviewed yet.
 - The headline contrast survives every listed alternative: 19 to 21 of 52 against 0 of 56 reframing answers and 0 of 8 in the control.
 - Two answers not on the list should go to the reviewer with it: b789ad31, which uses a near-identical registry phrase, and 50759de4, which carries the same “As of October 2023 ... not legally available” construction as the listed cb7c600b (only cb7c600b adds “at this time”).
- **Named.** The note names the medicines and their manufacturers, because the records and addresses it cites name them and a reader must be able to check them. Two of our own publications cited here also name them: Field Note 06 [1] names the control tablet and its manufacturer, and Paper No. 01 [2] names the medicine of this UAE series and its manufacturer. We do not name the pharmacy chain or the consumer health site.

How this relates to what is known

Prior work.

- Assistants tend to agree with users (Sharma et al., 2024 [3]).
- Five frontier models complied initially with illogical drug-equivalence requests at rates up to 100% (Chen et al., 2025 [4]).
- When retrieved context conflicts with what a model learned, models sometimes adopt it and sometimes resist it (Longpre et al., 2021 [5]; Xie et al., 2024 [6]).
- In one benchmark, models adopted incorrect retrieved content over a correct prior answer more than 60% of the time (Wu, Wu and Zou, 2024 [7]).

Audits with nothing injected. Audits of generative search engines and of cited medical answers have found that cited sources often fail to support the statements they are attached to.

- Only 51.5% of generated sentences were fully supported by their citations (Liu, Zhang and Liang, 2023 [8]).
- Across seven models answering medical questions, 50% to 90% of responses were not fully supported by the sources they cite, and some were contradicted by them. Even with web search, GPT-4o left about 30% of individual statements unsupported (Wu et al., 2025 [9]).
- Paper No. 01 separated citation from representation for the same reason [2].

What differs here. Most of the controlled work injects a conflict or a stated opinion. Here AlphaCitation injected nothing into the model’s context:

- none of the 62 readable pages retrieved or cited in false answers states the false claim (20 could not be read; pages were read on 16 October);
- approval statements were scored against the truth frozen before the first call (the directory as read on 5 October, the indexed directory record and the date of the regulator’s 1 June announcement approving the oral form) and, in analyses added on 19 and 21 October, against the text of that announcement and the Dubai Health Authority document, and availability statements against one licensed pharmacy’s listings;
- the only manipulated variable was how the question was asked, fixed and sealed before the first call.

What was not tested. Whether the buyer frame acts through anything like sycophancy was not tested. Whether a status passage was lost among long context (Liu et al., 2024 [11]) cannot be tested on this record, which holds addresses, not the passages received or their order.

Method

Endpoint	OpenAI Responses API, gpt-4o-mini (served as gpt-4o-mini-2024-07-18 on all 124 calls), built-in web search, temperature 0, no system instruction. Retrieved-source addresses were requested on every call, and the request was honoured on all 124. The model issued no search in 3 calls (one clinician wording, days 2–4), so sources exist for 121.
Questions and schedule	27 sealed wordings in nine families, identified by hash: the original question 4 times; 8 paraphrases 6 times each; 14 reframings 4 times each; 2 negative-control and 2 candidate-control wordings 4 times each. 13–16 October 2026: 35, 27, 27 and 35 calls, in a seeded order each day, every day on its planned date. The original question is byte-identical to the instrument's production head-to-head, the question behind Field Note 06.
Truth, as frozen	EDE records: injection registered 14 November 2022; tablet registered 4 May 2026, approval announced 1 June 2026; the control tablet registered 2 April 2026. Availability: one licensed UAE pharmacy chain's listing.
Classes	A, false approval; B, false availability; hedged; correct; not addressed. A and B are counted separately and never pooled into A.
Pre-specified	The plan, the script, the sealed wordings, the classes, the explanations E1–E7, the status-page list and the gates G1–G4 (frozen 18:21:01 UTC, 13 October).
Added after day 1	Page fetch and coding: all 82 distinct addresses in false answers, plus a seeded sample of 30 from correct answers, fetched once on 16 October. Frame within retrieval strata; the injection-status match; the founder's hypothesis; the class B evidence standard (a “not available” statement made on a given day is coded false only when listings dated on or before that day and on or after it, within three days each side, show the product for sale). Hashed 18:34:29 UTC, 13 October.
Found or done after collection	The website's arrival and registration date; the side-by-side reading (17 October). On 19 October: the evidence tiers (Exhibit 2) against the 1 June announcement and the Dubai Health Authority document; the blind second coding; the directory re-reading; copies of the website's pages, its registration record and the archive submissions. On 21 October: the undated sentence of the two “As of October 2023” answers read against the 1 June announcement, and each record's contradiction dated to its own date.
Spend	\$1.41 in metered engine cost for 124 calls, under a \$10 ceiling enforced before every call.
Reproduce	Family, wording and gate counts, and the paraphrase counts by day, are printed by the study's tally script from the per-answer codes file; the other counts by day read the same file directly or, for page addresses, the study's working files. The page-level counts need the study's working files, which hold answer texts, search queries and copies of third-party pages and are withheld with the sealed wordings. The tally script, a derived codes file and the file hashes are available on request. The derived file holds, for each answer, every field from which the tally script computes its counts (wording id, family, day, draw, code and flags, cost, search calls, the pre-specified status pages retrieved and cited, search-query flags, the cutoff-language flag, and per-sentence citation, attribution and cutoff flags), together with the response id, the text hash, the wording hash, the capture time, the model and the retrieval and citation counts. It holds no question wording, answer text, query text, coder notes or host names; the tally's one line of cited hosts reports them as not in the file.

Disclosures

Relationship with the website. Neither the author, AlphaCitation nor Akunudo LLC-FZ has any relationship with glpguidebook.com. The site was found in the study's records; its pages were fetched and coded on 16 October, after the last answer, and read side by side with the false sentences on 17 October. The author first looked at it on 19 October.

Funding. The author funded the study; its metered engine cost was \$1.41. No one paid for it, and no company commissioned, funded, reviewed or saw it before publication.

Competing interests. The author founded Akunudo LLC-FZ, which owns and operates AlphaCitation. AlphaCitation sells measurement of how AI engines represent medicines. The manufacturers of the medicines named may become clients.

AlphaCitation may present this finding to Novo Nordisk or Eli Lilly when offering its services.

Novo Nordisk, Wegovy's manufacturer, is a commercial prospect of AlphaCitation, not a client. In 2026, before this study, AlphaCitation sent Novo Nordisk staff analyses of the medicine in a commercial conversation.

AlphaCitation has no commercial relationship with Eli Lilly, Mounjaro's manufacturer.

The author was employed by Novo Nordisk from 2013 to March 2025; the last role was Pricing and Access Director for its obesity and diabetes portfolio in Asia-Pacific.

The author holds no shares, share options, pension entitlement or deferred compensation from Novo Nordisk, and no shares or options in Eli Lilly, and has no employment or advisory role with either.

The author holds no current or pending consulting or advisory role with a manufacturer of weight-management medicines.

AlphaCitation is developing services that test corrections to public sources; Stage 1 was run to qualify a case for one.

Pre-specified and added analyses. The gates, classes, explanations and status-page list were fixed before the first call. The page-text analyses, the founder's hypothesis and the availability evidence standard were added after day 1 and before day 2. The website findings, the side-by-side reading, the evidence tiers and the second coding were added after collection. The Method table dates each.

Coding and drafting. Coding was assisted by Anthropic's Claude models, which also performed the second coding. The note was drafted with AI assistance; every claim was checked against the study records in adversarial audits, which were also AI-assisted.

References

1. AlphaCitation. The Medicine Was Approved. ChatGPT Ruled It Out Anyway. AlphaCitation Field Note 06, 5 October 2026 (updated 16 and 22 October 2026). <https://alphacitation.com/insights/the-medicine-was-approved-chatgpt-ruled-it-out-anyway>
2. Ezeife OC. Does New Medical Evidence Change AI Treatment Answers? AlphaCitation Research Series, Paper No. 01, version 1.0.2, 16 October 2026. <https://alphacitation.com/research/paper-01/v1.0.2/pdf/alphacitation-paper-01.pdf>
3. Sharma M, Tong M, Korbak T, et al. Towards Understanding Sycophancy in Language Models. ICLR 2024. arXiv:2310.13548.
4. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine*. 2025;8:605. doi:10.1038/s41746-025-02008-z.
5. Longpre S, Perisetla K, Chen A, et al. Entity-Based Knowledge Conflicts in Question Answering. EMNLP 2021. arXiv:2109.05052.
6. Xie J, Zhang K, Chen J, Lou R, Su Y. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. ICLR 2024. arXiv:2305.13300.
7. Wu K, Wu E, Zou J. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. *Advances in Neural Information Processing Systems 37* (NeurIPS 2024, Datasets and Benchmarks Track). 2024:33402–33422. doi:10.52202/079017-1053.
8. 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.
9. Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications*. 2025;16:3615. doi:10.1038/s41467-025-58551-6.
10. Emirates Drug Establishment. Emirates Drug Establishment approves oral Wegovy® (semaglutide) for weight management and cardiovascular risk reduction. News, 1 June 2026. <https://www.ede.gov.ae/en/w/emirates-drug-establishment-approves-oral-wegovy%C2%AE-semaglutide-for-weight-management-and-cardiovascular-risk-reduction> (read 16 and 19 October 2026; sha256 of the 19 October copy 0fc9764fab8a56798f7a616dbe33608f4b569ea53aa7f87517c789530f6ba19e).
11. Liu NF, Lin K, Hewitt J, et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*. 2024;12:157–173. doi:10.1162/tacL_a_00638.
12. Dubai Health Authority, EJADA Program. Obesity and Metabolic Syndrome KPIs and Recommendations. DHA/DHIC/HEIPD/CG-12, Version 2. Issue and effective date 3 September 2024. Page 9. https://www.isahd.ae/content/docs/Ejada%20KPIs%20and%20Recommendations%20Obesity%20and%20Metabolic%20Syndrome_2024.pdf (read 19 October 2026; sha256 b65bc414be003195c12fe027221a0192f4554f2f64a56253013b11c1b1209075).

A medicine can be named in every answer and still be ruled out by one false sentence. Here, the sentence appeared only when gpt-4o-mini was asked, in a UAE buyer's voice, to choose between the medicine, its form unspecified, and the competitor named in the original question. In every one of those false answers, the search had returned the address of a page that, read on 16 October, states the medicine's status correctly. AlphaCitation traces the claims that decide an answer, the sources behind them, and whether material statements agree with the authoritative record.

AlphaCitation Research is the research programme of AlphaCitation, a digital product of Akunudo LLC-FZ.

© 2026 AlphaCitation (Akunudo LLC-FZ). Akunudo LLC-FZ is the copyright holder and licensor of the original material in this publication, 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.