FIELD NOTE 10 · CONFIRMATORY STUDY

AI can cite a genuine FDA source and still give a medicine's obsolete indication.

Field Note 09 promised to test its findings. This is the result.

In a pre-specified study of 936 answers across eight AI models, AlphaCitation documented repeated errors involving current US and EU medicine labels, including cases where authentic but outdated regulatory evidence accompanied incorrect conclusions.

Ozue C. Ezeife

1 November 2026

Oncology and Alzheimer's disease · United States, European Union · 8 engines · 23 min read

This note reports how eight AI models answered on two medicines' labels: Keytruda in the United States, and Leqembi in the United States and the European Union. Until 22 May 2025, Keytruda with chemotherapy was approved in the United States as a first-line treatment for advanced oesophageal cancer whatever the tumour's PD-L1 level. That day the FDA restricted the indication to tumours with a PD-L1 combined positive score (CPS) of 1 or more [1]. Seventeen months later, between 29 and 31 October 2026, AlphaCitation asked eight AI models, in ten wordings on each of the three days, whether a patient with oesophageal cancer below that line is FDA-approved for the combination. Four of the eight stated the indication as it stood before May 2025 at a rate above the study's pre-set 10% bar: Sonar and Sonar Pro in 26 of 30 answers each, Claude Haiku 4.5 in 21 and GPT-4o mini in 14. These are coder A's counts, the frozen reading. On the reading of the study's second coder, coder B, those last two are 18 and 20, and Claude Haiku 4.5's

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

row is unread. Gemini 2.5 Flash did so in 7 of 30 and Gemini 3.8 Flash in 2 of 30 (8 and 5 on coder B's reading), and neither cleared the bar. GPT-5.5 did so in none of 30 on either coder's reading, and Claude Sonnet 5.5 in none of 30, or 1 of 30 on coder B's. The study can clear neither.

To a question written in the voice of someone whose father in Florida had a negative PD-L1 test, Claude Haiku 4.5 answered "I have good news for your dad", then that the combination is approved "regardless of PD-L1 status". Sonar Pro, given the same question, said the same "under the current U.S. FDA indication". To a question worded as a US formulary review, GPT-4o mini dated the claim itself: "As of October 2026", the FDA has approved the combination "regardless of PD-L1 expression levels". These sentences describe the label as it stood before May 2025. The current label says otherwise [2]. GPT-5.5, given the Florida question, said the combination is "not currently on-label/FDA-approved for him".

This is the confirmatory study announced in **Field Note 09** (https://alphacitation.com/insights/keytruda-leqembi-tzield-us-eu-labels), frozen and hashed before its first collection call, whose result that note said would be published whatever it showed [3]. Field Note 09 reported four pre-specified one-day studies of four models, GPT-4o mini, Claude Haiku 4.5, Gemini 2.5 Flash and Sonar, collected on 22 October 2026. Three errors from that day were worth testing properly. This study asked about each with new wordings, on three further days, of those four models and one current-generation model from each of their developers, GPT-5.5, Claude Sonnet 5.5, Gemini 3.8 Flash and Sonar Pro: 936 answers. The plan's test put two of the three errors above its 10% bar for at least one model. It did not do so for the third.

Lecanemab (Leqembi) accounts for two of the three errors. Its US label lets people with two copies of the APOE ε4 allele be treated; its EU authorisation excludes them [4] [5]. To questions about people in the EU, GPT-4o mini said in 17 of 30 answers that treatment can begin without a genetic test, and Claude Haiku 4.5 in 10 of 30. That is how the US label puts it; the EU authorisation makes the genotype a condition of treatment. Claude Haiku 4.5's reading depends on a coding rule: without the two answers coder A flags as contradicting themselves it is 8 of 30, below the bar.

The error in the other direction did not clear the bar. Field Note 09 reported that, asked about US patients, GPT-4o mini said people with two copies are excluded in 5 of 18 answers and Claude Haiku 4.5 in 3 of 18. In this study, on new wordings, Claude Haiku 4.5 gave it in none of 30 answers on coder A's reading and 2 of 30 on coder B's. GPT-4o mini gave it in 7 of 30 answers, or 4 of 30 on coder B's reading, which leaves the rate undetermined. At a true rate of 30%, this test would have cleared the bar about half the time (45% to 62%), so a failure to clear it is not evidence of a low rate. To a question from a person in Arizona who is APOE4/APOE4, it stated the EU's indication as if it applied, with the European Medicines Agency's page as its source, and said that Leqembi "is not recommended" for such people.

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

Sonar and Sonar Pro listed the same sources on all 30 pairs of Keytruda answers, so their counts are not independent. Coder A’s 7 of 30 for GPT-4o mini on people in the US with two copies rises to 11 of 30, above the bar, if hedged answers count as false, a reading added after collection.

Each of these errors is about eligibility: who may receive a medicine on its label, in the country where the question is asked. A family, a prescriber or a payer who acted on one would be applying another country’s rules, or rules that have since changed.

The plan fixed in advance when a count becomes a finding: a model is read as above the bar only when a one-sided test shows its rate above 10%, after allowing for the fact that repeated answers to one question are not independent, and after correcting for the three measures tested on each model. That reading says the rate exceeds 10%, not that the observed count is the rate. The study can confirm an error. It was never able to clear a model: zero false answers in 30 is consistent with a true rate of up to 11.4%, and up to 20.9% to 25.9% once repetition is allowed for. All the questions in this note are fixed wordings; no family or patient received these answers. Two AI model coders, A and B, read every answer, blind to each other and to the model. Where they split, both counts are printed.

Keytruda: four models gave the label from before May 2025

The study asked about first-line Keytruda with chemotherapy in oesophageal squamous cell carcinoma, for a patient at CPS 0, PD-L1 negative or below 1, in the voices of families, patients, oncologists, pharmacists and payers.

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

Exhibit 1. Did the models state that first-line Keytruda with chemotherapy is FDA-approved below CPS 1? 240 answers to 10 questions, 30 per model, 29 to 31 October 2026.

MODEL	CODER A	CODER B	INTERVAL	READING
Sonar	26 of 30	26 of 30	59.0% to 96.7%	Above 10%
Sonar Pro	26 of 30	26 of 30	59.0% to 96.7%	Above 10%
Claude Haiku 4.5	21 of 30	18 of 30	44.5% to 87.2%	Above 10%
GPT-4o mini	14 of 30	20 of 30	22.9% to 72.1%	Above 10%
Gemini 2.5 Flash	7 of 30	8 of 30	8.8% to 49.0%	Undetermined
Gemini 3.8 Flash	2 of 30	5 of 30	1.2% to 30.3%	Undetermined
GPT-5.5	0 of 30	0 of 30	0.0% to 20.9%	Undetermined; not cleared
Claude Sonnet 5.5	0 of 30	1 of 30	0.0% to 20.9%	Undetermined; not cleared

The study's primary measure, fixed before collection; models are never pooled. Interval: 95%, widened for repetition, on coder A's count, the frozen reading. Reading: above 10% when the plan's test rejects; undetermined when it does not. Every answer coder A counts here stated the indication as it stood before 22 May 2025. Coder B's reading of Claude Haiku 4.5 differs in two ways: it counts 18, and it codes 6 of the model's 9 control answers false, which under the plan would leave this row unread. All six say a patient at CPS 15 to 30 is covered, which is what the control tests, and give the old indication as the reason. Coder A codes them correct. Gemini 2.5 Flash's 7 of 30 clears no bar under the frozen rule; with a repetition allowance from this study alone it would, and coder B's 8 of 30 misses the frozen test narrowly. Coder A flags six of Claude Haiku 4.5's 21 as contradicting themselves, and coder B flags none;

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

without them it is 15 of 30, still above the bar. Sonar and Sonar Pro are not two independent results: see below. Source: per-answer codes, reproduced by the study's tally script.

Coder A and coder B agree on the rows above 10% except where the exhibit says. GPT-4o mini's gap runs the other way from Claude Haiku 4.5's. Coder B counts six more of its answers. Five of them say a patient at CPS 0 is not covered, the right conclusion, four from the wrong threshold of CPS 10 and one from a stated October 2023 cut-off, and the sixth hedges. Both counts stand.

The plan compares each developer's two models descriptively, not as a test. On Keytruda, on the same days, coder A (coder B): GPT-4o mini 14 (20) and GPT-5.5 0 (0); Claude Haiku 4.5 21 (18) and Claude Sonnet 5.5 0 (1); Gemini 2.5 Flash 7 (8) and Gemini 3.8 Flash 2 (5); Sonar 26 (26) and Sonar Pro 26 (26). Three of the four pairs also differ in tier, GPT-4o mini searched in 21 of its 30 answers against GPT-5.5's 30, and Claude Sonnet 5.5 ran at default sampling, so the study cannot say what separates the two models in any pair. In 114 of 117 paired answers across the study, Sonar and Sonar Pro listed identical sources, so their two rows are not two independent confirmations: on Keytruda all 30 pairs listed the same sources. The record shows the lists, not how they were produced.

All but one of the false answers on this question ran a search: 13 of GPT-4o mini's 14, and all of them from the other five models. Among what they cited are older records. Nine of GPT-4o mini's 14 false answers (10 of coder B's 20) cited the FDA's notice of March 2021, which is dated, accurate for its day and silent on the 2025 change [**6**]. Fourteen of Claude Haiku 4.5's 21 (12 of coder B's 18) cited the manufacturer's press release of the same month, whose site says its contents were "current as of the date presented" [**7**]. Twenty of Sonar's 26 cited documents filed for the FDA's advisory committee meeting of September 2024, eight months before the change: 12 cited the FDA's briefing document or slides, and 8 only sponsors' submissions [**8**]. Correct answers cited the same pages: GPT-5.5 the FDA notice in 22 of its 30, Claude Sonnet 5.5 the press release in 7 of its 29. The current label on DailyMed was cited by 28 of GPT-5.5's 30 correct answers, 5 of Claude Sonnet 5.5's 29 and none of GPT-4o mini's 14, and by one false answer, one of Claude Haiku 4.5's 21 (none of coder B's 18); no false answer from GPT-4o mini, Sonar or Sonar Pro cited it. One of Sonar's false answers listed the label the FDA approved with the May 2025 change; eight listed other Keytruda label files from the FDA's 2025 or 2026 folders, which were not read. These counts cover the GPT, Claude and Sonar models and were added after collection. The record shows what each answer cited, not which sentence came from which page.

The wording mattered. Sonar and Sonar Pro gave the false claim in 26 of the 27 answers to nine of the questions. To the tenth, which asks what PD-L1 result a patient needs and then whether CPS 0 qualifies, each answered correctly on all three days: "a CPS of 0 does not meet the FDA-labeled eligibility criterion", in Sonar's words. The same models, asked for the threshold, gave the threshold.

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

An answer can also be right for a reason that is no longer true. At CPS 15 to 30, inside the current US indication, coder A codes all 72 control answers correct, coder B codes 60 correct, and coder B flags 21 of the 72 as stating the old indication. Coder A set that flag on none of them: it recorded the same statements under another flag, a departure from the plan that changes no code or reading.

Leqembi: one gene, two labels

The US and EU labels for lecanemab differ on genotype eligibility and pre-treatment testing. Both warn that people with two copies of APOE ε4 have a higher rate of ARIA, the brain swelling and bleeding the treatment can cause. The US label lets them be treated, says testing should be done before the first dose and says that a patient who is not tested can still be treated [**4**]. The EU authorisation covers only non-carriers and people with one copy, says lecanemab is not indicated in people with two, and requires each patient’s genotype to be registered before the first infusion [**5**].

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

Exhibit 2. Lecanemab: the false claim for each market. 480 answers to 20 questions, 60 per model, 29 to 31 October 2026.

MODEL	EU QUESTIONS: SAID TREATMENT CAN START WITHOUT THE TEST	US QUESTIONS: SAID PEOPLE WITH TWO COPIES ARE EXCLUDED
GPT-4o mini	17 of 30 (18)	7 of 30 (4)
Claude Haiku 4.5	10 of 30 (10)	0 of 30 (2)
Sonar	6 of 30 (5)	0 of 30 (0)
Sonar Pro	5 of 30 (4)	0 of 30 (0)
Gemini 2.5 Flash	2 of 30 (1)	0 of 30 (0)
GPT-5.5	0 of 30 (0)	0 of 30 (0)
Claude Sonnet 5.5	0 of 30 (0)	0 of 30 (0)
Gemini 3.8 Flash	0 of 30 (0)	0 of 30 (0)

The study's primary measures, fixed before collection. Each cell counts answers that stated the false claim for that market, of 30, coder A first and coder B in brackets. Intervals: 95%, widened for repetition, coder A. EU questions: GPT-4o mini 30.5% to 79.6%, Claude Haiku 4.5 13.3% to 62.0%, both above 10% on the plan's test; every other row undetermined, Sonar 6.2% to 48.5%, Sonar Pro 4.7% to 45.0%, Gemini 2.5 Flash 1.0% to 33.6%, and 0.0% to 24.5% at zero. US questions: GPT-4o mini 7.6% to 52.8%; every row undetermined, and 0.0% to 25.9% at zero. Two readings turn on a rule. Claude Haiku 4.5's EU count includes two answers coder A flags as contradicting themselves; the plan codes those false, and without them it is 8 of 30, 9.8% to 55.0%, undetermined. GPT-4o mini's US count excludes four hedged answers and one that gave only advice. Counting hedges as false makes it 11 of 30, 15.4% to 64.9%, above the bar, and a repetition allowance from this study alone also takes its 7 of 30 above the bar. Source: per-answer codes, reproduced by the study's tally script.

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

All ten of Claude Haiku 4.5's false answers carry the FDA label's words that untested patients "can still be treated", to questions about France, Italy, Ireland and the EU. Nine apply those words to the person asked about; the tenth labels them as the FDA's, and reads the EU's "should be performed" as a recommendation. Five of the ten state the EU's genotype limit in the same answer. Seven also recommend the test or say it should be done, and an eighth calls it "essentially a requirement for eligibility, not optional" before saying patients "can still be treated". To a question about a father in France, Claude Haiku 4.5 answered that the father "can technically begin Leqembi infusions without the APOE4 genetic test first, but this is not recommended". To a question worded for a hospital formulary in Spain, GPT-4o mini said that lecanemab "can be initiated without knowing the ApoE4 genotype". Sonar credited the idea to the EU itself, saying the EU product information was "allowing treatment if testing is not performed". The EU product information in force has no such sentence.

Claude Haiku 4.5's false answers clustered by wording. Nine of the ten came to family and lay questions, including both caregiver questions on all three days, though two of the three lay questions drew none; to the questions in the voice of a pharmacist, a formulary or a payer Claude Haiku 4.5 made none. All ten came to the five wordings that name Leqembi and none to the five that name lecanemab, so voice and brand name cannot be separated here. GPT-4o mini showed no such pattern: it gave the false claim to nine of the ten questions, including all six answers to the two questions the plan frames as clinicians'.

On the US side, GPT-4o mini's seven answers sit in four of the ten questions, and three of the seven are to the person in Arizona. One answered that "you would not be eligible to receive Leqembi in the United States". Three of the seven cited the European Medicines Agency's page for the medicine; on 22 October each of the five false answers in its 5 of 18 cited a page on that agency's website. The coders split on three of the seven, two of which state the EU's limit and then say people with two copies can be treated, which is why the count is 7 or 4.

What this does to Field Note 09

Every count in **Field Note 09** (https://alphacitation.com/insights/keytruda-leqembi-tzield-us-eu-labels) stands as a count from 22 October 2026. The wordings here are new, so a difference below cannot be read as a change over time. It can come from the questions.

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

Exhibit 3. The one-day findings against this study: 22 October 2026 against 29 to 31 October 2026, by model.

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

FINDING AND MODEL	22 OCTOBER	29 TO 31 OCTOBER	RESULT
Keytruda, Claude Haiku 4.5	3 of 6 (oesophageal 3 of 3)	21 (18)	Confirmed under the frozen rule, oesophageal cancer only; unread on coder B's codes
Keytruda, Sonar	3 of 6 (oesophageal 0 of 3)	26 (26)	Differs: none on the oesophageal question that day, above 10% here
Keytruda, GPT-4o mini	2 of 6 (oesophageal 0 of 3)	14 (20)	Differs: none on the oesophageal question that day, above 10% here
Keytruda, Gemini 2.5 Flash	2 of 6 (oesophageal 1 of 3)	7 (8)	Seen again, undetermined; above 10% with a repetition allowance from this study alone
EU testing, Claude Haiku 4.5	4 of 6 (second coder 3)	10 (10)	Confirmed under the frozen rule; 8 of 30 and undetermined without coder A's self-contradiction flags
EU testing, GPT-4o mini	0 of 6	17 (18)	Differs: none that day, above 10% here
EU testing, Sonar	0 of 6	6 (5)	Differs: none that day, seen here, undetermined
EU testing, Gemini 2.5 Flash	1 of 6	2 (1)	Seen again, undetermined
US exclusion, GPT-4o mini	5 of 18	7 (4)	Seen again, undetermined; above 10% if hedges count as false (11 of 30), or with a repetition allowance from this study alone
US exclusion, Claude Haiku 4.5	3 of 18	0 (2)	Not repeated on coder A's reading; undetermined, not cleared

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

One-day counts from Field Note 09, by those studies' coder A; their second coder gives the same false codes except where shown, and on Keytruda flags the earlier indication on 2 of Claude Haiku 4.5's 3 oesophageal answers. This study's counts of 30, coder A first and coder B in brackets. On Keytruda the one-day count covers an oesophageal and a gastric question, and this study asked about oesophageal cancer only.

This study did not retest teplizumab and the age floor Field Note 09 reports as picked up; the EU Keytruda threshold and the questions at CPS 1 to 9; the gastric question; people with two copies of APOE ε4 in the EU; the US testing questions; the companion study's test of the question's voice; the Saudi and UAE registration questions; or the comparators. Those stay one-day counts in Field Note 09.

What the records establish, and what they cannot

The study was built to answer what one day could not. Between 29 and 31 October 2026, on ten new wordings per measure, the plan's test put the rate above 10% for Sonar, Sonar Pro, Claude Haiku 4.5 and GPT-4o mini stating a version of Keytruda's indication that the FDA had replaced seventeen months earlier, and for GPT-4o mini and Claude Haiku 4.5 saying, on EU questions, that treatment can start without the test on which the EU indication depends. It did not do so for any model on saying, on US questions, that people with two copies of APOE ε4 are excluded: those rows are undetermined, not cleared. Those are findings about these models, on these three days, on these questions. Three days show no trend.

Four readings turn on a rule, and each is given both ways in the exhibits. On coder B's codes, Claude Haiku 4.5's Keytruda row is unread, because that coder marks six of its control answers false, each of which gives the right conclusion with the old reason; both coders agree on 18 of its false answers. Without the answers coder A flags as self-contradicting, Claude Haiku 4.5's EU testing count falls to 8 of 30, below the bar. If hedged answers counted as false, GPT-4o mini's US count would rise to 11 of 30, above it. With a repetition allowance taken from this study alone, instead of the larger of that and the one fixed in advance, GPT-4o mini's US count and Gemini 2.5 Flash's Keytruda count would both clear the bar. Six results hold on both coders' agreed codes: Keytruda for all four models named above, and EU testing for GPT-4o mini and Claude Haiku 4.5.

At these sizes, the plan cannot establish reliability. GPT-5.5 gave none of these false answers in 90, and Claude Sonnet 5.5 none in 90 on coder A's reading and 1 on coder B's. That is a result worth having, but it is not a clearance. An error not observed on these days may still occur.

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

The cause is not established. The pattern is consistent with answers that mix two markets’ labels, or a label’s current and earlier versions, and with answers built on older records. But the record shows what each answer said and cited, not why.

Under AlphaCitation’s Evidence-Authority Standard, a claim about who may receive a medicine in a market is settled only by that market’s authorisation record and authorised label, in force on the claim’s date [9]. Every answer here was coded that way, against the labels in force when it was given.

For a regulated medicine, the right answer in one country can be the wrong answer in another. An answer has to say which label it is applying, and apply the one in force.

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

How we know this

Question	Do AI models state a medicine’s indication by the label of the market asked about, where the US and EU labels differ, and as the label stands on the day asked?
Study	Confirmatory study S6: lecanemab in the EU, can treatment start without genotyping; pembrolizumab with chemotherapy, first-line oesophageal squamous cell carcinoma below CPS 1, US; lecanemab in the US, people with two copies of APOE ε4; each with a control.
Plan	Hashed before a smoke test, then frozen and hashed after it and before its first collection call, with the question wordings unchanged between the two hashes. It sealed 42 wordings and asked 39: an amendment before its freeze left three control wordings unasked. Plan sha256 b6576ed0…, sealed wordings 0916f6cb…, both in the released freeze record. Changes after the freeze are numbered deviations in the study’s internal result record; none changes a count in this note.
Models	All eight reached through their developers’ APIs, not the consumer apps. gpt-4o-mini and gpt-5.5 with web search; Claude Haiku 4.5 (claude-haiku-4-5) and Claude Sonnet 5.5 (claude-sonnet-5-5) with web search, Sonnet at its developer’s default sampling, because it refuses the fixed temperature, and a longer answer limit, because at that sampling shorter limits cut answers off, so its answers are not temperature-pinned; Gemini 2.5 Flash and Gemini 3.8 Flash with Google Search grounding; Sonar and Sonar Pro (sonar-pro) from Perplexity. The instrument’s fixed settings; no system instruction; each model except Sonar and Sonar Pro, which always search, chose whether to search.
Questions	39 wordings, ten for each measure and three for each control, written in the voices of patients, families, physicians, pharmacists and payers, naming the market in different ways. Each is identified by its hash; the texts are withheld, as in Field Notes 07 to 09.
Schedule	One round a day on 29, 30 and 31 October 2026, from 06:05 UTC, 312 answers a day, 936 in all. One call timed out on 29 October and was asked again that morning.

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

Deviation	The plan listed every false code, disagreement and self-contradiction flag for a qualified human reviewer before publication. The founder withdrew that step on 30 October 2026, before any coding. No one adjudicated this study: every count is the coders', and both are shown. Coder A recorded statements of the old indication in the control answers under a general flag instead of the earlier-label flag the plan asks for; no code or reading changes.
Coding	Every answer read in full and coded on its target, in the market named, as false, correct, hedged, advice only or not addressed, by rules fixed before collection, with flags for the earlier label, self-contradiction and other wrong regulatory facts. Two model coders from different developers coded every answer, each blind to the other and to the model, the day and the round, though answer style and citation lists can still hint at a model; they agreed on 883 of 936 codes (Cohen's kappa 0.81 overall; 0.83 on EU testing, 0.87 on Keytruda, 0.46 on people with two copies, and 0.00 on two controls, where one code dominates, one of them hiding the 12-answer split on which coder B's reading of the Keytruda control turns). This is agreement between two model coders, not independently established fact.
Truth	The regulator's label in force when each answer was given: the FDA's prescribing information and approval letter for the United States; the EU product information for the European Union [1] [2] [4] [5]. Read 28 October 2026 and again on 31 October 2026 before any answer was coded; every document the study is coded against was unchanged, byte for byte.
Gemini	Reported as counts only.
Pre-specified	The three primary measures per model, never pooled; the three controls; the reading at 10% with the design effect fixed as the larger of the prior and the estimate, and Holm's correction across each model's three measures; the completeness criterion, met in full. The plan's labels: MATERIAL, called above 10% here, when the test of a rate above 10% rejects; UNDETERMINED when neither test rejects; NOT MATERIAL, unreachable at this size. The reading says a rate exceeds 10%, not what the rate is. Field Note 09's one-day studies read the interval's lower bound, not this test.

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

Added after the freeze	Every reading on coder B's codes or on both coders' agreed codes; the readings without self-contradictions, with hedges as false, and with this study's own repetition allowance; the page-level counts of cited sources and the comparison of Sonar with Sonar Pro; the counts of quoted words and of the EU limit in Claude Haiku 4.5's EU answers; the descriptions of split answers; and the comparison with the one-day counts. All are descriptive.
Reproduce	The readings are printed by the study's **frozen reading script** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/s6_tally.py) (sha256 9b951b54..., working paths removed from its comments) from its **codes file** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/codes-public.json), which holds each answer's codes and flags from both coders, its text hash, capture time, search and citation counts, and only the regulator hosts among its cited pages; no question wording, answer text or coder note. The **freeze record** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/freeze.json), the **agreement script** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/kappa.py), the **tally script** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/tally.py) the reading script loads and the **file hashes** (https://alphacitation.com/research/insights/keytruda-leqembi-confirmatory-study/release-hashes.json) sit in the same folder. The tally and agreement scripts are the files released with Field Note 09: the tally script is the frozen file (sha256 27d9648c...) with working paths removed from its comments and the running total cut from the end of one printed line, which changes no reading, and the agreement script (sha256 31387f00...) has working paths removed from its comments. The counts that rest on answer text or cited addresses the released files withhold are the page-level counts of cited sources, the Sonar and Sonar Pro comparison, the words "can still be treated" and the EU limit in Claude Haiku 4.5's EU answers, the descriptions of what the split answers say, and the voices, markets and brand or molecule naming of the questions. The quoted sentences are readings of answer text, not counts from the codes file.

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

The records

- **Keytruda, US.** FDA approval letter for supplement S-182, 22 May 2025, revising the first-line oesophageal indication "to reflect the population [i.e., PD-L1 (CPS ≥1)]" [1]; DailyMed prescribing information, set id 9333c79b, version 228, effective 10 July 2026 [2]. Both read 28 and 31 October 2026.
- **Leqembi, US.** DailyMed prescribing information, set id 9d1ff786, version 24, revised 7/2026: indicated "for the treatment of Alzheimer's disease", with no genotype restriction; the boxed warning says untested patients "can still be treated" [4]. Read 28 and 31 October 2026.
- **Leqembi, EU.** Product information updated 5 August 2026: indicated in adults "who are apolipoprotein E ε4 (ApoE ε4) non-carriers or heterozygotes with confirmed amyloid pathology"; section 4.2: testing "should be performed prior to initiation of treatment"; section 4.4: "not indicated for use in patients who are homozygotes"; Annex II: the genotype is captured "prior to the first infusion of lecanemab, for all patients"; the package leaflet: "Your doctor can arrange a genetic test for ApoE4" [5]. Read 28 and 31 October 2026.

Limitations

- Three days and eight API models, with fixed questions; not the consumer apps, not other models, not other days.
- The study cannot clear a model of an error, at any count, and it says nothing about what it did not retest.
- Claude Sonnet 5.5's answers are not temperature-pinned, so their variation includes sampling. Sonar and Sonar Pro are not independent of each other.
- Coding was done by two AI models from different developers, with no adjudication. No human reviewer has read the codes. Correlated errors across models remain possible.

Disclosures

Relationship with the manufacturers. The medicines measured are made by Merck & Co. (MSD; Keytruda) and Eisai with Biogen (Leqembi). These companies are, or may become, commercial prospects of AlphaCitation; none is a client. AlphaCitation may present these findings to them when offering its

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

services. Before these studies, AlphaCitation had sent Eisai a one-page analysis of Leqembi in Saudi Arabia, on 1 September 2026, and Sanofi staff an analysis of Tzield in Saudi Arabia, and had prepared a briefing for Biogen in July 2026, which there is no record of having been sent. No manufacturer can edit, approve or suppress a finding.

Funding. The author funded the study. No one paid for it, and no company commissioned or funded it.

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

- 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 has no employment or advisory role with it. Novo Nordisk makes none of the medicines measured here.
- The author holds no shares or share options in Merck & Co., Eisai, Biogen, Sanofi, Otsuka, Lundbeck or Novartis, and has no employment, advisory or other role with any of them.

Coding and drafting. Coding was done by Anthropic's Claude, which is made by the company behind two of the eight models measured, and by a coder from a second developer whose models are measured here. To limit that, every coder was blind to which model wrote each answer. The note was drafted with AI assistance and checked against the study records in adversarial audits, also AI-assisted.

Scholarly correspondence and substantive commentary:
analyst@alphacitation.com

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

References

1. US Food and Drug Administration. . 22 May 2025. Read 28 and 31 October 2026.

2. US National Library of Medicine, DailyMed. **KEYTRUDA (pembrolizumab) prescribing information** (https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=9333c79b-d487-4538-a9f0-71b91a02b287), set id 9333c79b-d487-4538-a9f0-71b91a02b287, version 228, effective 10 July 2026. Read 28 and 31 October 2026.

3. AlphaCitation. **The FDA Narrowed Keytruda's Label in May 2025. Seventeen Months Later, All Four AI Models Still Gave the Old One.** (https://alphacitation.com/insights/keytruda-leqembi-tzield-us-eu-labels) Field Note 09, 29 October 2026.

4. US National Library of Medicine, DailyMed. **LEQEMBI (lecanemab-irmb) prescribing information** (https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=9d1ff786-e577-410a-a273-c4d7d0e4e975), set id 9d1ff786-e577-410a-a273-c4d7d0e4e975, version 24, revised 7/2026. Read 28 and 31 October 2026.

5. European Medicines Agency. , EMEA/H/C/005966, last updated 5 August 2026. Read 28 and 31 October 2026.

6. US Food and Drug Administration. **FDA approves pembrolizumab for esophageal or GEJ carcinoma** (https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-pembrolizumab-esophageal-or-gej-carcinoma). Content current as of 22 March 2021. Read 1 November 2026.

7. Merck. **FDA Approves Merck's KEYTRUDA (pembrolizumab) Plus Platinum- and Fluoropyrimidine-Based Chemotherapy for Treatment of Certain Patients With Locally Advanced or Metastatic Esophageal or Gastroesophageal Junction (GEJ) Carcinoma** (https://www.merck.com/news/fda-approves-mercks-keytruda-pembrolizumab-plus-platinum-and-fluoropyrimidine-based-chemotherapy-for-treatment-of-certain-patients-with-locally-advanced-or-metastatic-esophageal-or-g/). 23 March 2021. Read 1 November 2026. Cited here as a source the answers listed, not as evidence of the label.

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

8. Oncologic Drugs Advisory Committee, meeting of 26 September 2024, documents filed on the FDA's website: **FDA briefing document** (https://www.fda.gov/media/182143/download); **FDA presentations** (https://www.fda.gov/media/182213/download); **Merck briefing document** (https://www.fda.gov/media/182145/download); **Merck presentations** (https://www.fda.gov/media/182215/download); **Bristol Myers Squibb presentation on nivolumab** (https://www.fda.gov/media/182212/download). Read 1 November 2026. Cited here as sources the answers listed, not as evidence of the label.

9. AlphaCitation. **Which Evidence Settles Which Claim: the AlphaCitation Proposition Taxonomy and Evidence-Authority Standard** (https://alphacitation.com/insights/which-evidence-settles-which-claim), version 3.0. AlphaCitation Standard 01. doi:10.5281/zenodo.23594909.

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** (https://creativecommons.org/licenses/by/4.0/). Quoted third-party text, third-party data fields, API/provider-restricted material and third-party marks are excluded from that grant.

CITE THIS FIELD NOTE

Ezeife OC. AI can cite a genuine FDA source and still give a medicine's obsolete indication. AlphaCitation Field Note 10, 1 November 2026. https://alphacitation.com/insights/keytruda-leqembi-confirmatory-study