Factual Accuracy of Large Language Models in Natal Chart Interpretation: A Comparative Study of 22 Models

Vít Hnilica
Twelvera, Czech Republic
vit.hnilica@twelvera.com
ORCID: 0009-0000-9993-9648
DOI: 10.5281/zenodo.24615466
Preprint · Version 1.0 · November 2026

Large language models are widely used to write natal chart readings, yet a reading can only be as sound as the chart it describes. We measured how often 22 language models (8 proprietary, 14 open-weight) state chart facts that contradict a reference chart, and whether supplying a computed chart table changes this. For 40 synthetic births (28 with a known birth time) each model wrote three English readings under up to five conditions: birth data only (P0), a computed table of positions, houses and aspects (P1), the same table with row identifiers to cite (P2), and two richer structured inputs (P3, P4). A rule-based detector compared stated signs, houses, angles and aspect types with the reference chart; every detector flag in conditions P1–P4 for 11 selected models was then reviewed manually. Without a table, 18 of 21 models produced at least one flagged claim in 44 % or more of known-time readings, and 15 in 89 % or more; one model was flagged in 1 of 82 readings. With the table, the flagged share fell for 20 of 21 models, by 8 to 98 percentage points (birth-level permutation tests, Holm-adjusted $p < 0.05$ for 19 of the 20). Among 2,688 known-time readings from the 11 selected models, manual review of detector flags confirmed a basic factual error in three readings. A separate review of 30 unflagged readings found no additional basic errors. Most remaining flags were loose wording or detector false positives. Small open-weight models were still flagged in 12–69 % of readings with the table. Richer input did not consistently improve the accuracy of basic claims, and four derived errors found in a small sample came only from the richer conditions. The study concerns agreement between text and chart; it does not test the validity of astrology.

Keywords: large language models; natal chart; factual consistency; grounding; hallucination; ephemeris; evaluation

1 Introduction

Language models are now a common way to obtain an astrological reading. A user types a date, time and place of birth and receives fluent prose about a Sun sign, an Ascendant, planets in houses and aspects between them. Such a reading makes checkable statements: that the Moon is in Taurus, that Mars is in the seventh house, that Venus squares Saturn. These follow from astronomy and a small number of declared conventions. They are either consistent with a correctly computed chart or they are not.

Computing a chart requires planetary positions for a given instant, a conversion from local civil time to universal time and, for houses, spherical trigonometry. Language models are unreliable at exact multi-step numerical work without tools: accuracy on arithmetic and compositional tasks falls quickly with problem size (Qian et al., 2023; Dziri et al., 2023), date and time arithmetic remains error-prone (Wang & Zhao, 2024; Fatemi et al., 2025), and coordinate geometry is harder than recalling geographic knowledge (Truong et al., 2026, preprint). Tools and retrieved documents improve factual performance (Lewis et al., 2020; Schick et al., 2023). Fluent generated text also frequently contains statements unsupported by its source in summarisation (Maynez et al.,

2020; Kryscinski et al., 2020) and in table-to-text generation (Parikh et al., 2020), even when the source is supplied (Ji et al., 2023; Huang et al., 2025). A common measure splits a text into atomic claims and checks each against a reference (Min et al., 2023). We follow that approach with a deliberately narrow claim set.

We are not aware of a published study that measures the factual consistency of language-model natal chart readings against a computed chart. We searched arXiv, the ACL Anthology, Google Scholar, the open web and the sites of *Correlation* and the *Journal of Scientific Exploration*; subscription indexes were not queried (Appendix H). The closest work is two preprints on Chinese Four Pillars charts: a benchmark of multiple-choice questions (Chen & Liu, 2026) and a study of which computed facts a model must be given to answer questions about a chart (Han & Liu, 2026). An essay in *Correlation* discusses astrology and language models conceptually (McRitchie, 2023) and reports no measurements.

Scope. The question is whether a reading agrees with the chart for the stated birth data: when a model writes that a planet is in a sign or a house, is that so? An error of this kind matters to practitioners and clients in the same way that a wrong date matters in a biography.

We examine three hypotheses, formulated so that the data could contradict them.

- **H1.** Supplying a computed chart table reduces the share of readings that contain a chart statement contradicting the reference chart.
- **H2.** With a table, the best-performing models rarely make basic factual errors, but not necessarily never.
- **H3.** Supplying additional derived chart data (dignities, rulers, dispositors, element balance, patterns) improves accuracy further.

A secondary, exploratory question is whether accuracy holds across eight languages.

2 Method

2.1 Design

Each model wrote readings for the same 40 births, with three repetitions per birth and condition, yielding 120 readings per model and condition, of which 84 were for births with a known time. All primary analyses use the known-time readings because house, angle and Moon claims cannot be checked when the time is unknown. The conditions were crossed with models as far as cost and context length allowed (Table 1).

2.2 Births and reference charts

The 40 births are synthetic. They were generated with a fixed random seed: a date between 1940 and 2009, a local time, and one of 25 cities on six continents (latitudes −37.81° to 59.33°). Twenty-eight births have a known time and 12 are marked as time unknown. No data from real people were used.

Reference charts were computed with the Twelvera chart engine: geocentric apparent positions from the JPL DE442s ephemeris, tropical zodiac, Placidus houses and the five Ptolemaic aspects. Local times were converted to UTC with the IANA time zone database. As an independent check, all 40 charts were recomputed with the Swiss Ephemeris: positions, house cusps and angles agreed to within 0.5″, every sign (including that of the lunar node), house and angle sign was identical, and applying the orb rule to the recomputed positions gave the same 901 aspects as the engine, with none missing and none added (Appendix C). Aspect orbs follow traditional moieties taken from Lilly (1647, p. 107), capped at 6° when either body is non-traditional (Appendix C). This is one convention among several, so the detector checks only the aspect *type* between two bodies, not the orb (Section 2.5).

For births with an unknown birth time, the chart was computed for local noon, and houses and angles were withheld. In none of these 12 births did a planet other than the Moon change sign during the day.

2.3 Conditions

The system instruction was identical in all conditions (Appendix A). It asks for an English reading of 400–600 words that mentions concrete chart factors, prescribes the conventions, excludes topics relating to health, death, pregnancy, law and investment, and prohibits references to houses and angles when the birth time is unknown. The user message differed:

- **P0, birth data only:** date, local time (or "unknown"), city, country and coordinates.
- **P1, chart table:** birth data plus a plain-text table of 11 positions with sign, degree, house and retrograde status, the Ascendant and Midheaven when the time is known, and all aspects within orb. The instruction adds "Use only the chart factors listed above."
- **P2, table with row identifiers:** as P1, with each row labelled [E1], [E2], … and an instruction to cite the rows after each sentence that mentions a chart factor.
- **P3, engine data:** birth data plus the engine's structured natal output, which adds dignities, rulers, dispositors, element and modality balance, chart patterns, aspect strength tiers and a priority list, without interpretive text.
- **P4, engine data with glossary and rules (supplementary):** as P3, with the proprietary glossary and interpretation rules of the Twelvera product. This input cannot be published, so P4 is supplementary and no main conclusion rests on it.

P0–P2 were run for all 22 models. P3 and P4 were run for the 14 open-weight models (Phi-4 could not accommodate the P4 input) and for two proprietary models, Claude Sonnet 5.5 and GPT-5.6 Terra, chosen on cost grounds.

2.4 Models and generation

Eight proprietary models (three from Anthropic, five from OpenAI) were called through their vendors' batch interfaces. Fourteen open-weight models were called through OpenRouter, each pinned to a single hosting provider with fallbacks disabled; for one model, Gemma 4 31B, part of the requests was served by a second provider. The models are named in Table 1; exact identifiers, providers and declared quantisation are in Appendix D.

Generation took place on 24–26 November 2026. Requests contained only the system instruction, the user message and an output-token limit. No tool definitions were sent, so no model could call code execution, search or an ephemeris. Neither the sampling temperature nor the reasoning-effort parameter was set. The output limit was 8,000 tokens, raised for models that spend output tokens on internal reasoning (Appendix D). Readings cut off by the limit were excluded; for the Anthropic models and GPT-6 Luna in P0, truncated requests were resubmitted with a higher limit.

2.5 Automatic detector

A rule-based detector (regular expressions; Appendix E) extracts four kinds of claim from each reading and compares them with the reference chart:

1. a body stated to be in a sign;
2. the sign of the Ascendant or Midheaven;
3. a body stated to be in a house;
4. an aspect type between two bodies; the claim is flagged if the reference chart has no such aspect between them within orb or has an aspect of a different type.

For unknown-time births, Moon claims are skipped and any house or angle statement is counted separately as a prohibited statement. The detector does not examine degrees, orbs, retrograde status, dignities, rulerships, dispositors, element balance, chart patterns or superlatives such as "the tightest aspect"; we call these *derived claims*. The reading-level outcome is whether a reading contains at least one flagged claim. A flag is detector output and not a verified error.

2.6 Manual review

Detector precision was assessed on two sets, kept separate throughout (Appendix F).

Set 1 is a random sample of 100 flags from all models: 50 of the 9,072 flags in P0 and 50 of the 2,172 flags in P1–P4. Because small open-weight models produce most flags, the P1–P4 half mainly describes them.

Set 2 is the complete set of 94 flags in P1–P4 for 11 models: the eight proprietary models, Gemma 4 26B, Gemma 4 31B and DeepSeek V4.1 Flash. These are the models with the lowest flagged rates, for which detector noise matters most. Two flags occur in both sets.

Each flagged claim was reviewed in the context of its full sentence and checked against the chart table and assigned to one of six categories according to a written rule: *basic factual error* (a sign, house or aspect type given to the wrong body, or an aspect absent from the table); *minor imprecision* (for example, one sign given for a conjunction that straddles a sign boundary); *derived error*; *detector false positive*; *ambiguous*; and *duplicate*. An initial classification was made with an AI assistant; the author then re-examined every item in both sets. The author was the only human reviewer, and blinding to model and condition was not enforced, so no inter-rater agreement can be reported.

In addition, 30 randomly chosen unflagged known-time readings by the same 11 models were read in full against the chart. This sample shows what the detector leaves out. It is not a sensitivity estimate.

2.7 Multilingual check

Six models (Claude Sonnet 5.5, GPT-6.1 Sol, GPT-6.2 Sol, Gemma 4 26B, Gemma 4 31B, Mistral Small 3.2 24B) wrote P2 readings for 10 known-time births, with two repetitions per birth, in English, Czech, Slovak, Polish, German, Spanish, French and Italian. The detector was not validated outside English, so a weaker check was used: for each sentence that names a zodiac sign and cites at least one row, the sign must belong to one of the cited rows. This measures the consistency of citations, not the correctness of every claim.

2.8 Cost

For the proprietary models, cost was estimated from the token usage returned with each response and the list prices at the time, with the 50 % batch discount and without cache discounts; it is not the invoiced amount. For open-weight models it is the charge reported by OpenRouter.

2.9 Statistical analysis

Proportions are given as counts k/n with Wilson 95 % score intervals. A count of zero is reported with its upper bound (0/84: 4.4 %) and is not interpreted as a zero error rate. These intervals cover only the claim types the detector checks.

The design is paired: every model wrote for the same births, and the three repetitions for a birth are not independent. Comparisons between conditions therefore pair readings by birth and repetition and take the birth as the unit of analysis: an exact sign-flip permutation test on the per-birth differences, with Holm adjustment across models within each comparison (Supplementary Table S9), and a 95 % interval for the difference from a

cluster bootstrap over births (4,000 resamples). Discordant pairs and intervals are in Supplementary Tables S1–S4. Intervals for single proportions (Table 1) treat the 84 readings in a cell as independent, although they come from 28 births, and are therefore somewhat too narrow.

3 Results

3.1 Completed readings

Of 11,640 planned readings in the main English experiment (12,056 requests including resubmissions), 11,121 were completed; 467 ended at the output limit before any reading was written, 49 readings were cut off and 3 were empty. Of the 97 model-by-condition cells requested, 88 have all 84 known-time readings, and 7,769 known-time readings in 95 cells were analysed. DeepSeek V4.1 Flash completed no known-time P0 reading and LFM3 1.7B completed no P2 reading; in both cases, internal reasoning consumed the output limit. Cells with missing readings should be interpreted with caution, since truncation may not be independent of difficulty.

3.2 Automatic detector: all models

Table 1 and Figure 1 give the share of known-time readings with at least one flagged claim.

Without a table (P0). Twenty-one models have a P0 result. Eighteen were flagged in 44 % or more of readings and fifteen in 89 % or more; eight open-weight models were flagged in every reading (84/84; lower bound 95.6 %). Three models had lower rates: GPT-6.2 Sol 16/84 (19.0 %; 12.1–28.7), GPT-6.1 Sol 7/84 (8.3 %; 4.1–16.2) and Claude Fable 5.2 1/82 (1.2 %; 0.2–6.6). Houses, angles and aspects were flagged most often and signs least often, which is expected because the Sun's sign follows from the date alone except close to a sign boundary, while houses require the full computation.

The Claude Fable 5.2 result in P0 was checked for artefacts. Its requests were built in the same way as those of the other models, contained no tool definitions and used the same births. Its readings were as long as those of Claude Opus 5.5 and contained a similar number of checkable claims, so the low rate is not explained by vaguer text. Of the 389 degree values stated in these readings, 370 were within 1° of the reference position (median deviation 0.32°). The requests used about 10,200 output tokens on average, against about 1,750 with the table, which is consistent with a long internal computation, although we did not inspect the reasoning. As a check outside the detector, 18 of the 82 readings were read in full by two other AI assistants (GPT-6.2 Sol, itself one of the evaluated models, and GPT-6 Astra), and the author compared every discrepancy they reported with the reference chart (Appendix F). No sign or house was incorrect. Two readings named three aspects that the reference chart does not list (orbs of 7.0–7.5°, outside the 6° limit for non-traditional bodies); each was described as wide, and the detector had not matched these sentences. One reading gave Venus at 24° where the reference has 22°42′. Under the rule applied elsewhere in this paper these two readings contain a basic error, so the detector's 1/82 understates the rate under that rule; under a convention that admits orbs of 8° it does not.

With a table (P1). The flagged share was lower in P1 than in P0 for 20 of 21 models (Figure 2; paired results in Supplementary Table S1). The decrease ranged from 8.3 percentage points (GPT-6.1 Sol, from 7/84 to 0/84) to 97.6 points (both Gemma 4 models, from 84/84 to 2/84). Across these models, 993 pairs were flagged only in P0 and 5 only in P1. In the birth-level test, the decrease remained below 0.05 after Holm adjustment for 19 of the 20 models; for GPT-6.1 Sol, whose seven flagged P0 readings came from six births, it did not ($p = 0.031$, adjusted $p = 0.062$). The exception is Claude Fable 5.2, whose rate was already low: 1/82 in P0 and 2/84 in P1 overall, and one flagged reading in each condition among the 82 complete pairs. In P1, the eight proprietary models and the two Gemma 4 models were flagged in 0–10.7 % of readings; the remaining open-weight models were flagged in 11.9–69.0 %. For three of the smaller models, roughly half or more of the readings were still flagged even when the table was supplied (Ministral 3B 48.8 %, Ministral 8B 50.0 %, Llama 3.2 3B 69.0 %).

3.3 Manual review: eleven models with a table

Table 2 and Figure 3 give the results after every flag for the 11 best-performing models was reviewed. Across their 32 cells in P1–P4 (2,688 known-time readings), the detector flagged 77 readings. After review, 3 readings contained a basic factual error: one each for GPT-5.6 Terra in P1, Gemma 4 31B in P2 and DeepSeek V4.1 Flash in P3 (each 1/84; 1.2 %; 0.2–6.4). In the other 29 cells, no basic error was observed (0/84; upper bound 4.4 %). Under the stricter measure that also counts minor imprecision, 60 readings were affected; the highest rates were observed for DeepSeek V4.1 Flash in P1 (17/84; 20.2 %) and GPT-5.6 Terra in P1 (9/84; 10.7 %), and 17 cells had none.

These figures apply to claims within the detector's scope; derived claims were not checked systematically.

3.4 What the flags were

Table 3 and Figure 4 give the classification.

In Set 1, 49 of the 50 P0 flags were basic factual errors (98 %; 89.5–99.6). Of the 50 P1–P4 flags, 42 were basic errors (84 %; 71.5–91.7), 3 minor imprecisions, 3 detector false positives and 2 ambiguous. The manual review confirms that most sampled flags in the mixed-model population represent basic factual errors, but does not directly validate the reading-level error rates in Table 1.

In Set 2, the picture is different. Of 92 unique flags, 6 were basic errors (6.5 %; 3.0–13.5), 63 minor imprecisions, 21 detector false positives and 2 derived errors. For the best-performing models, the automatic flag rates in P1–P4 thus considerably overstate the frequency of basic errors. The typical minor imprecision is a sentence such as "Mars conjunct Jupiter in Gemini", where one of the two bodies is in the adjacent sign.

The 30 unflagged readings contained no basic error that the detector had missed (0/30; upper bound 11.4 %). They did contain 4 minor imprecisions, 1 ambiguous sentence and 4 derived errors.

The reviewed material shows three distinct ways in which a reading can come to contradict the chart, and they should not be pooled.

1. **Source-data error.** The input itself is wrong or misleading. Two of the derived errors can be traced to the engine output used in P3 and P4: the "lacking element" was computed from the seven traditional planets and the Ascendant only, so a reading said there were no planets in water signs when Uranus was in one; and the priority list named a single final dispositor when another planet in its own sign also disposed of itself. The models reported what they were given. Both defects have since been corrected in the engine (Appendix G); the readings were not regenerated.
2. **Grounding error.** The correct fact is in the input, but the text contradicts it: the wrong sign, house or aspect type for a named body. This is what the detector measures and what the table reduces.
3. **Derivation error.** The input is correct, but the text draws an incorrect inference from it, such as calling an aspect the tightest when another is tighter, or converting 0.32° to 0°32′.

Whether a statement counts as an error can also depend on convention. Aspects wider than 6° are within orb under the moiety rule used here only for pairs of traditional bodies (up to 14.75° between the Sun and the Moon) and are outside it under the tighter orbs many modern practitioners use. We did not count the use of such aspects as an error.

3.5 Amount of checkable content and error rate

If models with low error rates simply made fewer checkable statements, the comparison would be misleading. The data do not show this (Figure 5). Across models, the number of checkable claims per reading was negatively associated with the flagged share of claims (Spearman $\rho = -0.64$ in P0, $n = 21$ models; $\rho = -0.60$ in P1, $n = 22$). All eight proprietary models made more checkable claims per reading in P1 (by 5–14 claims). The pro-

portion of claims flagged fell for seven of the eight; for Claude Fable 5.2 it was about 0.1 % in both conditions (1 of 1,036 claims in P0, 2 of 1,670 in P1). Reading length did not differ materially between the two conditions (mean 577 words in P0 and 574 in P1; per-model results in Supplementary Table S6). These associations are descriptive. In P3 and P4, readings contained fewer checkable claims than in P1 for most models, which should be borne in mind when their flagged shares are compared.

3.6 Row identifiers and richer input

P2. Asking the model to cite table rows did not change the flagged share reliably. It was lower than in P1 for 10 models, higher for 7 and equal for 4. The largest change was an improvement for Llama 3.2 3B (from 58/84 to 33/84), which did not remain below 0.05 after Holm adjustment in the birth-level test (adjusted $p = 0.059$).

P3. Additional derived chart data did not consistently improve the accuracy of basic chart claims. Among the 16 models with both conditions, the flagged share was lower in P3 than in P1 for 11 and higher for 5. After Holm adjustment, only the reduction in the share of flagged readings for DeepSeek V4.1 Flash remained below 0.05 (18/84 to 2/84; adjusted $p = 0.008$). The largest increase among complete cells, for Phi-4 14B (34/84 to 55/84), did not (adjusted $p = 0.092$); LFM3 1.7B rose more in the paired comparison (33.3 percentage points), but on 27 pairs only. The effect on derived claims could not be established with the present detector, which does not check them.

As a preliminary observation, all four derived errors in the 30 unflagged readings occurred in P3 or P4 (4 of 11 readings) and none in P1 or P2 (0 of 19). Two of the four can be traced to the engine output itself (Section 3.4). Richer input gives a model more derived statements to make and therefore more opportunities for such errors; the sample is too small to estimate a rate.

P4 (supplementary). Adding the proprietary glossary and rules to the P3 input changed the flagged share in no consistent direction (in paired readings lower for 6 models, higher for 5, equal for 4); no comparison approached significance after adjustment.

3.7 Unknown birth time

For the 12 unknown-time births (36 readings per cell), the instruction prohibited references to houses and angles. None of the eight proprietary models, the two Gemma 4 models, DeepSeek V4.1 Flash, Qwen 3.5 9B or Granite 4.2 8B stated a house or an angle in any condition. Seven smaller open-weight models did so without a table, from 11 of 36 P0 readings (Ministral 3B, Mistral Small 3.2 24B) to 26 of 36 (Llama 3.2 3B), and LFM3 1.7B did so once in its 11 completed P0 readings. With the table, which omits houses and angles for these births, such statements became rare (0–6 of 36 in P1), but Llama 3.2 3B and Gemma 3 4B produced them again with the richer input (12 and 3 of 36 in P3). These counts were not reviewed manually.

3.8 Multilingual check

Table 4 gives the citation-consistency result for P2 in eight languages. For GPT-6.1 Sol and GPT-6.2 Sol, no sentence named a sign outside its cited rows in any language; Claude Sonnet 5.5 had 3 such sentences out of 1,369 (0.2 %) and Gemma 4 31B had 6 out of 1,082 (0.6 %). Mistral Small 3.2 24B had 36 of 1,321 (2.7 %), from 0 % in French to 6.0 % in Slovak. Gemma 4 26B had 84 of 959 (8.8 %), most of which reflected a single systematic error: citing the Ascendant row for a statement about the Midheaven; its rate was similar across languages (5.8–13.1 %).

Within the limits of this check, we found no evidence that citation consistency declined outside English for the three proprietary models. Per-language cells contain 20 readings, so small differences between languages are not interpretable. Sentences that name a sign without citing any row are not in the denominator; they were numerous in German for three models, so coverage in German is incomplete (Supplementary Table S8).

Table 4. Exploratory check of citation consistency: sentences naming a sign that is not in any cited row / sentences naming a sign and citing a row (P2; 20 readings per cell unless noted).

Model	en	cs	sk	pl	de	es	fr	it	Total
GPT-6.1 Sol	0/191	0/186	0/177	0/169	0/167	0/178	0/164	0/168	0/1,400
GPT-6.2 Sol	0/197	0/187	0/201	0/193	0/180	0/198	0/190	0/190	0/1,536
Claude Sonnet 5.5	0/182	1/159	0/196	1/182	1/121	0/187	0/159	0/183	3/1,369
Gemma 4 31B	0/144	0/149	1/130	0/135[a]	3/106[b]	1/144[b]	1/136[b]	0/138[a]	6/1,082
Mistral Small 3.2 24B	3/154	5/131	14/235	3/149	3/106	1/174	0/207	7/165	36/1,321
Gemma 4 26B	14/136	6/103	6/100	13/99	14/127	10/132	13/136	8/126	84/959

[a] 19 readings. [b] 18 readings.

3.9 Cost

For all eight proprietary models, a reading was cheaper with the table than without it, by a factor of about 3 (GPT-6.2 Sol) to 12 (Claude Fable 5.2) when the full cost of a condition, including failed and repeated requests, is divided by the number of unique completed readings (Supplementary Table S7; Figure 6), even though the P1 input is three to four times longer. The savings came from reduced use of reasoning tokens, not from shorter readings: GPT-6.1 Sol, for example, used an average of 2,526 reasoning tokens out of 3,166 output tokens per P0 request and 53 of 710 in P1. Claude Fable 5.2, the only model that was rarely flagged without a table, cost USD 0.48 per unique completed P0 reading against USD 0.04 in P1. For open-weight models that do not reason internally, the table increased costs slightly, and P3 and P4 cost roughly 2–11 times more than P1. The total estimated generation cost of the main experiment was about USD 126.

4 Discussion

H1 is supported for the observed outcome. Supplying a computed chart table substantially reduced the share of readings flagged by the detector. The flagged share fell for 20 of 21 models, in most cases by more than 40 percentage points, and the P0 flags were almost all real errors in the reviewed sample. The practical implication is straightforward: for nearly every model tested, a reading written from birth data alone describes a chart that differs from the reference chart, usually in the houses, the angles or the aspects.

H2 is supported in part. "Rarely" was not given a threshold in advance, so the hypothesis is assessed descriptively. With a table, several models reached near-zero observed error rates: after manual review of all detector flags, basic factual errors were identified in 3 of 2,688 readings. The 11 models were selected for review because their flagged rates were low, so the figure does not generalise to all 22. The smaller open-weight models continued to contradict the table in a large share of readings, so a provider who uses a small model for cost reasons should expect errors even with a correct chart in the prompt.

H3 is not supported by these data. Additional derived chart data did not consistently improve the accuracy of basic chart claims. The four derived errors found in the richer conditions are a preliminary observation that warrants a follow-up study using a detector or rating protocol designed for derived claims.

One model without a table. Claude Fable 5.2 was flagged in 1 of 82 readings from birth data alone, and we found no evidence of unintended chart data in the requests. This result covers 28 births between 1940 and 2009, a single model version and one week, at about twelve times the cost of supplying a table. It does show that the P0 error rate is a property of particular models at a particular time and not a fixed limit; the results provide a dated benchmark against which future model generations can be evaluated.

Cost. Supplying the table also made a reading cheaper for all eight proprietary models, by a factor of about 3 to 12, although the input is three to four times longer. The reduction in reasoning-token usage is consistent with the models no longer needing to derive the chart themselves. For these models, a longer, computed prompt was

less expensive and, in most cases, substantially more accurate; for open-weight models that do not reason internally the table added a small cost.

Derived-data consistency. While all tested positions, house placements and aspects were independently reproduced, manual review identified two issues in the engine's derived chart data (the lacking element and the final dispositor). Both have since been corrected in the Twelvera engine. A grounded reading inherits the errors of its source, and a study that treats the source as ground truth cannot detect them by construction; we found these two only because the readings were read in full. "Ground truth" in this paper therefore applies to positions, houses and aspect types, not to every derived field.

5 Limitations

- **Sample size and coverage.** Forty births, 28 with a known birth time, limit both precision and coverage. Births before 1940, historical time-zone changes, high latitudes and birth times near sign or house boundaries were not sampled deliberately.
- **Detector scope.** Only signs, houses, angle signs and aspect types are checked, in English, by pattern matching. Sensitivity was not estimated; in the secondary review of 18 Claude Fable 5.2 P0 readings the detector missed three out-of-orb aspects. Precision differs sharply between samples (84 % of flags were basic errors in the mixed sample, 6.5 % for the best models), so the automatic flag rates in Table 1 are only roughly comparable across models; for the best models, only the manually reviewed figures should be used.
- **Aspects without a table.** In P0 a model must choose its own orbs, so an aspect that a reading names and the reference chart lacks may reflect a wider orb convention and not a wrong position; the detector cannot tell the two apart.
- **Manual review.** The author was the only human reviewer, without enforced blinding. P0 flags were reviewed only in a random sample.
- **Conventions.** Placidus houses, the tropical zodiac, the true node and one orb scheme were fixed. Results may differ under other conventions.
- **Models and time.** Hosted models change. Results describe the versions served on 24–26 November 2026 under default settings, with unequal output limits, through one hosting provider per open-weight model, except Gemma 4 31B, which was partly served by a second (Appendix D). Missing cells and truncated readings reduce comparability.
- **One prompt.** A single system instruction was used. P1 adds both the table and the instruction to use only the listed factors, so their effects cannot be separated, and P3 changes the format as well as the amount of input. Other wording, a request to compute step by step, or access to tools could change P0 results substantially.
- **Supplementary parts.** P4 uses proprietary product-specific interpretation rules and is therefore reported as a supplementary experiment; all primary conditions (P0–P3) are documented and released. The multilingual check measures citation consistency only.

6 Conclusion

Providing a computed natal chart together with an explicit instruction to use it substantially improved the factual consistency of language-model readings. Across 22 models and more than 11,000 completed readings, the share of flagged readings fell for 20 of the 21 models for which a direct comparison was possible. Among 2,688 known-time readings written by 11 selected models with computed chart data, manual review of all detector flags identified three readings with a basic factual error. Smaller open-weight models continued to misstate the table they had been given.

Computed chart data also reduced the cost of a reading about three- to twelvefold for the eight proprietary models tested. Additional derived information did not consistently improve the accuracy of basic chart statements and can introduce errors of its own.

These findings support a practical principle for AI-assisted astrology: separate deterministic chart computation from natural-language interpretation. Doing so improves factual grounding, lowers the cost of generation for reasoning models and gives generated readings a verifiable foundation.

Statements

Conflict of interest. The author develops Twelvera, a commercial astrology application that uses language models with a computed chart, and the reference charts were produced by its engine. The charts were checked independently against the Swiss Ephemeris (Section 2.2).

Use of AI tools. An AI assistant (Claude, Anthropic) assisted with the analysis code, an initial classification of detector flags that the author then re-examined, and the preparation and language editing of this manuscript; two further AI assistants (GPT-6.2 Sol and GPT-6 Astra, OpenAI) read a sample of readings as a secondary check (Section 3.2). Models from the same vendor are among those evaluated. The author verified all results and is responsible for the content.

Data and materials. Births, prompts for P0–P3, the original model readings for P0–P3, reading-level detector results, the detector, analysis scripts, aggregated results and manual-review records are available at https://github.com/etruhla/twelvera-research (folder `llm-natal-chart-accuracy`). The P4 input and the P4 readings are not released.

Ethics and consent. No personal data were used; the births are synthetic. No ethics committee approval was sought.

Funding. No external funding. The author paid the costs of model access.

Author contributions. The author is the sole author and is responsible for the design, the data, the manual review and the text.

References

Chen, J., & Liu, Q. (2026). *BaziQA-Benchmark: Evaluating symbolic and temporally compositional reasoning in large language models* (arXiv:2602.12889). arXiv. https://doi.org/10.48550/arXiv.2602.12889 [preprint]

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., Hwang, J., Sanyal, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems, 36*. https://doi.org/10.52202/075280-3081

Fatemi, B., Kazemi, M., Tsitsulin, A., Malkan, K., Yim, J., Palowitch, J., Seo, S., Halcrow, J., & Perozzi, B. (2025). Test of time: A benchmark for evaluating LLMs on temporal reasoning. *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb7295a8bc613b375726659c2ecd6f14-Abstract-Conference.html

Han, H., & Liu, J. (2026). *Compute the interactions, not the attributes: What a fact layer must supply in the Chinese sexagenary system* (arXiv:2611.13597). arXiv. https://doi.org/10.48550/arXiv.2611.13597

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 9332–9346). https://doi.org/10.18653/v1/2020.emnlp-main.750

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Lilly, W. (1647). *Christian astrology*. London.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). https://doi.org/10.18653/v1/2020.acl-main.173

McRitchie, K. (2023). The marriage of astrology and AI: A model of alignment with human values and intentions. *Correlation, 36*(1), 43–49. https://philpapers.org/archive/MCRTMO.pdf

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). https://doi.org/10.18653/v1/2023.emnlp-main.741

Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1173–1186). https://doi.org/10.18653/v1/2020.emnlp-main.89

Qian, J., Wang, H., Li, Z., Li, S., & Yan, X. (2023). Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 9285–9298). https://doi.org/10.18653/v1/2023.acl-long.516

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems, 36*. https://doi.org/10.52202/075280-2997

Truong, T. H., Lau, J. H., & Qi, J. (2026). *GPSBench: Do large language models understand GPS coordinates?* (arXiv:2602.16105). arXiv. https://doi.org/10.48550/arXiv.2602.16105 [preprint]

Wang, Y., & Zhao, Y. (2024). TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 6389–6415). https://doi.org/10.18653/v1/2024.findings-acl.382

Tables

Cells give k/n (per cent; Wilson 95 % interval). Intervals for single proportions treat the 84 readings in a cell as independent although they come from 28 births.

Table 1. Automatic detector, all 22 models: known-time readings with at least one flagged claim

Flags are detector output, not verified errors. Empty cells indicate that the condition was not run or yielded no completed known-time reading. Proprietary models are listed first, followed by open-weight models ordered by their P1 rate.

Model	P0	P1	P2	P3	P4
Claude Fable 5.2	1/82 (1.2; 0.2–6.6)	2/84 (2.4; 0.7–8.3)	1/84 (1.2; 0.2–6.4)		
GPT-6.1 Sol	7/84 (8.3; 4.1–16.2)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)		
GPT-6.2 Sol	16/84 (19.0; 12.1–28.7)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)		
GPT-6.1 Terra	37/84 (44.0; 33.9–54.7)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)		
Claude Opus 5.5	42/84 (50.0; 39.5–60.5)	2/84 (2.4; 0.7–8.3)	0/84 (0.0; 0.0–4.4)		
Claude Sonnet 5.5	49/84 (58.3; 47.7–68.3)	3/84 (3.6; 1.2–10.0)	1/84 (1.2; 0.2–6.4)	1/84 (1.2; 0.2–6.4)	1/84 (1.2; 0.2–6.4)
GPT-6 Luna	82/84 (97.6; 91.7–99.3)	3/84 (3.6; 1.2–10.0)	3/84 (3.6; 1.2–10.0)		
GPT-5.6 Terra	81/84 (96.4; 90.0–98.8)	9/84 (10.7; 5.7–19.1)	3/84 (3.6; 1.2–10.0)	1/84 (1.2; 0.2–6.4)	5/84 (6.0; 2.6–13.2)
Gemma 4 26B	84/84 (100.0; 95.6–100.0)	2/84 (2.4; 0.7–8.3)	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Gemma 4 31B	84/84 (100.0; 95.6–100.0)	2/84 (2.4; 0.7–8.3)	6/84 (7.1; 3.3–14.7)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Mistral Small 3.2 24B	81/84 (96.4; 90.0–98.8)	10/84 (11.9; 6.6–20.5)	14/84 (16.7; 10.2–26.1)	11/84 (13.1; 7.5–21.9)	5/84 (6.0; 2.6–13.2)
LFM3 1.7B	37/40 (92.5; 80.1–97.4)	9/49 (18.4; 10.0–31.4)		14/42 (33.3; 21.0–48.4)	21/58 (36.2; 25.1–49.1)
Granite 4.2 8B	80/84 (95.2; 88.4–98.1)	17/84 (20.2; 13.0–30.0)	5/23 (21.7; 9.7–41.9)	7/84 (8.3; 4.1–16.2)	4/84 (4.8; 1.9–11.6)
DeepSeek V4.1 Flash		18/84 (21.4; 14.0–31.3)	8/84 (9.5; 4.9–17.7)	2/84 (2.4; 0.7–8.3)	3/84 (3.6; 1.2–10.0)
Qwen 3.5 9B	75/84 (89.3; 80.9–94.3)	18/84 (21.4; 14.0–31.3)	22/84 (26.2; 18.0–36.5)	5/84 (6.0; 2.6–13.2)	8/83 (9.6; 5.0–17.9)
Gemma 3 4B	84/84 (100.0; 95.6–100.0)	27/84 (32.1; 23.1–42.7)	22/84 (26.2; 18.0–36.5)	24/84 (28.6; 20.0–39.0)	33/84 (39.3; 29.5–50.0)
Llama 3.1 8B	84/84 (100.0; 95.6–100.0)	30/84 (35.7; 26.3–46.4)	22/84 (26.2; 18.0–36.5)	15/84 (17.9; 11.1–27.4)	11/84 (13.1; 7.5–21.9)
gpt-oss 20B	83/84 (98.8; 93.6–99.8)	32/84 (38.1; 28.4–48.8)	36/84 (42.9; 32.8–53.5)	31/84 (36.9; 27.4–47.6)	38/84 (45.2; 35.0–55.9)
Phi-4 14B	84/84 (100.0; 95.6–100.0)	34/84 (40.5; 30.6–51.2)	36/84 (42.9; 32.8–53.5)	55/84 (65.5; 54.8–74.8)	
Ministral 3B	84/84 (100.0; 95.6–100.0)	41/84 (48.8; 38.4–59.3)	40/84 (47.6; 37.3–58.2)	38/84 (45.2; 35.0–55.9)	34/84 (40.5; 30.6–51.2)
Ministral 8B	84/84 (100.0; 95.6–100.0)	42/84 (50.0; 39.5–60.5)	46/84 (54.8; 44.1–65.0)	43/84 (51.2; 40.7–61.6)	33/84 (39.3; 29.5–50.0)
Llama 3.2 3B	84/84 (100.0; 95.6–100.0)	58/84 (69.0; 58.5–77.9)	33/84 (39.3; 29.5–50.0)	64/84 (76.2; 66.1–84.0)	59/84 (70.2; 59.8–79.0)

Note. DeepSeek V4.1 Flash completed 6 of 120 P0 readings (none with a known time) and LFM3 1.7B completed no P2 reading; Phi-4 14B could not accommodate the P4 input. Cells with n below 84 lost readings to truncation (Fable P0 82; LFM3 40–58; Granite P2 23; Qwen P4 83).

Table 2. Manually verified models (11 models, conditions P1–P4): known-time readings after manual review of every flag

Every one of the 94 detector flags for these models was checked against the chart table and classified (Table 3). "Automatic flag" repeats the corresponding values from Table 1. "Basic error" = a sign, house or aspect type explicitly given to the wrong body, or an aspect absent from the table. "Basic or minor" adds loose wording such as one sign given for a conjunction that straddles a sign boundary. Errors outside the detector's scope (derived claims) are not counted here. This table counts known-time readings; Table 3 counts individual flags, including those in readings for births without a known time, so a model can have a basic error in Table 3 and none here (three of the six basic errors in Set 2, among them the one for Claude Opus 5.5, are in unknown-time readings).

Model	Cond.	Automatic flag	Basic error confirmed from a detector flag	Basic or minor
Claude Fable 5.2	P1	2/84 (2.4; 0.7–8.3)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Claude Fable 5.2	P2	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	1/84 (1.2; 0.2–6.4)
Claude Opus 5.5	P1	2/84 (2.4; 0.7–8.3)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Claude Opus 5.5	P2	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Claude Sonnet 5.5	P1	3/84 (3.6; 1.2–10.0)	0/84 (0.0; 0.0–4.4)	2/84 (2.4; 0.7–8.3)
Claude Sonnet 5.5	P2	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	1/84 (1.2; 0.2–6.4)
Claude Sonnet 5.5	P3	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	1/84 (1.2; 0.2–6.4)
Claude Sonnet 5.5	P4	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.2 Sol	P1	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.2 Sol	P2	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.1 Sol	P1	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.1 Sol	P2	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.1 Terra	P1	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6.1 Terra	P2	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-6 Luna	P1	3/84 (3.6; 1.2–10.0)	0/84 (0.0; 0.0–4.4)	3/84 (3.6; 1.2–10.0)
GPT-6 Luna	P2	3/84 (3.6; 1.2–10.0)	0/84 (0.0; 0.0–4.4)	3/84 (3.6; 1.2–10.0)
Gemma 4 26B	P1	2/84 (2.4; 0.7–8.3)	0/84 (0.0; 0.0–4.4)	2/84 (2.4; 0.7–8.3)
Gemma 4 26B	P2	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	1/84 (1.2; 0.2–6.4)
Gemma 4 26B	P3	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Gemma 4 26B	P4	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Gemma 4 31B	P1	2/84 (2.4; 0.7–8.3)	0/84 (0.0; 0.0–4.4)	1/84 (1.2; 0.2–6.4)
Gemma 4 31B	P2	6/84 (7.1; 3.3–14.7)	1/84 (1.2; 0.2–6.4)	6/84 (7.1; 3.3–14.7)
Gemma 4 31B	P3	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
Gemma 4 31B	P4	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-5.6 Terra	P1	9/84 (10.7; 5.7–19.1)	1/84 (1.2; 0.2–6.4)	9/84 (10.7; 5.7–19.1)
GPT-5.6 Terra	P2	3/84 (3.6; 1.2–10.0)	0/84 (0.0; 0.0–4.4)	3/84 (3.6; 1.2–10.0)
GPT-5.6 Terra	P3	1/84 (1.2; 0.2–6.4)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
GPT-5.6 Terra	P4	5/84 (6.0; 2.6–13.2)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)
DeepSeek V4.1 Flash	P1	18/84 (21.4; 14.0–31.3)	0/84 (0.0; 0.0–4.4)	17/84 (20.2; 13.0–30.0)
DeepSeek V4.1 Flash	P2	8/84 (9.5; 4.9–17.7)	0/84 (0.0; 0.0–4.4)	8/84 (9.5; 4.9–17.7)
DeepSeek V4.1 Flash	P3	2/84 (2.4; 0.7–8.3)	1/84 (1.2; 0.2–6.4)	2/84 (2.4; 0.7–8.3)
DeepSeek V4.1 Flash	P4	3/84 (3.6; 1.2–10.0)	0/84 (0.0; 0.0–4.4)	0/84 (0.0; 0.0–4.4)

Pooled over the 32 cells: basic error in 3/2688 (0.1; 0.0–0.3) readings; basic or minor in 60/2688 (2.2; 1.7–2.9). Pooling mixes models and conditions and is provided only to indicate the overall scale.

Table 3. Manual classification of detector flags (claim level)

Set 1 is a random sample from all 22 models (population: 9,072 flags in P0 and 2,172 in P1–P4, counted over all readings including unknown-time ones); its P1–P4 half is dominated by small open-weight models. Set 2 is the complete set of P1–P4 flags for the 11 models in Table 2. Two flags occur in both sets. Precision figures apply only to the set from which they were obtained and must not be transferred between sets.

Category	Set 1, P0 (random 50)	Set 1, P1–P4 (random 50)	Set 2, P1–P4 (all 94)
Basic factual error	49	42	6
Minor imprecision	0	3	63
Derived error (outside detector scope)	0	0	2
Detector false positive	1	3	21
Ambiguous	0	2	0
Duplicate flag	0	0	2
Share of flags classified as basic errors	49/50 (98.0; 89.5–99.6)	42/50 (84.0; 71.5–91.7)	6/92 (6.5; 3.0–13.5) of 92 unique flags
Share of flags classified as basic errors or minor imprecisions	49/50 (98.0; 89.5–99.6)	45/50 (90.0; 78.6–95.7)	69/92 (75.0; 65.3–82.7) of 92 unique flags

Set 2 by model

Model	Flags	Basic	Minor	Derived	False positive	Duplicate
Claude Fable 5.2	3	0	1	0	2	0
Claude Opus 5.5	3	1	0	0	2	0
Claude Sonnet 5.5	7	0	5	0	2	0
GPT-6 Luna	7	0	6	0	0	1
GPT-5.6 Terra	20	1	11	0	8	0
Gemma 4 26B	5	1	4	0	0	0
Gemma 4 31B	10	1	8	0	1	0
DeepSeek V4.1 Flash	39	2	28	2	6	1

Unflagged readings read in full (30 random known-time readings by the 11 models, 21 distinct births)

Finding	P1/P2 (n = 19)	P3/P4 (n = 11)	All
No error found	16	5	21
Minor imprecision	2	2	4
Ambiguous	1	0	1
Derived error (outside detector scope)	0	4	4
Basic error missed by the detector	0	0	0

Finding no missed basic errors in 30 readings gives a Wilson upper bound of 11.4 % for the share of unflagged readings that contain one. This sample is not a sensitivity estimate: it was drawn from unflagged readings only, and derived claims are outside the detector's scope by design.

Figures

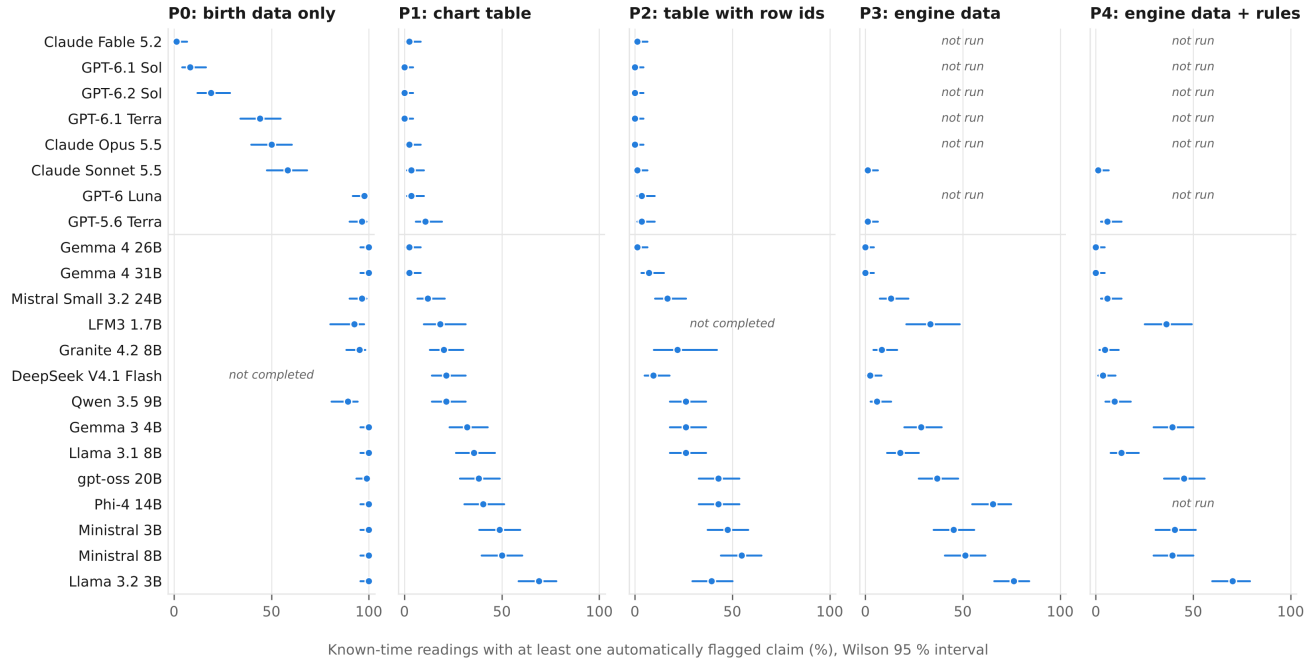


Figure 1. Known-time readings with at least one automatically flagged claim, by model and condition. Points show observed shares and bars show Wilson 95 % intervals (n = 84 unless noted in Table 1). Proprietary models are above the line; open-weight models are below.

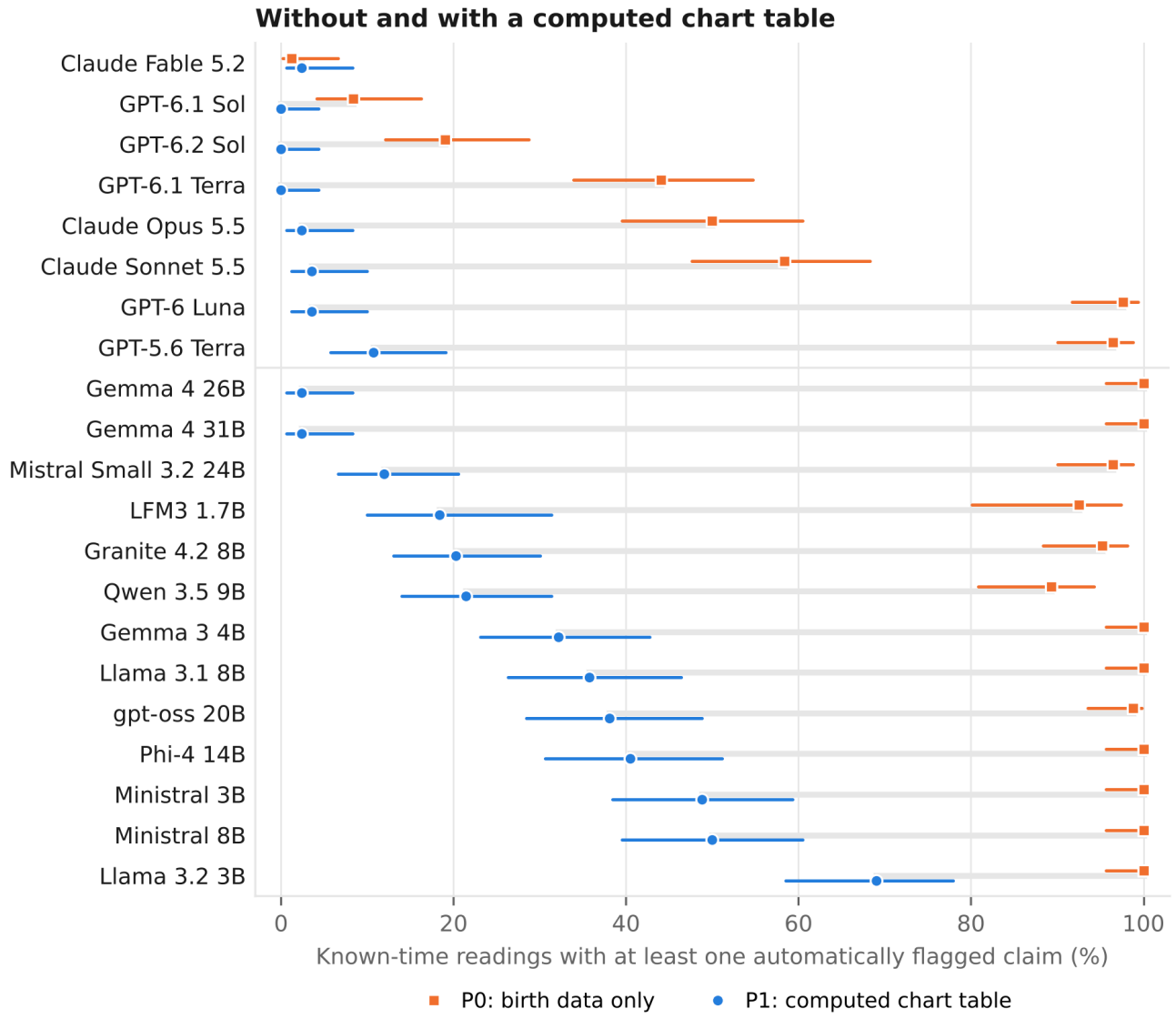


Figure 2. Flagged share without a chart table (P0) and with it (P1), with Wilson 95 % intervals.

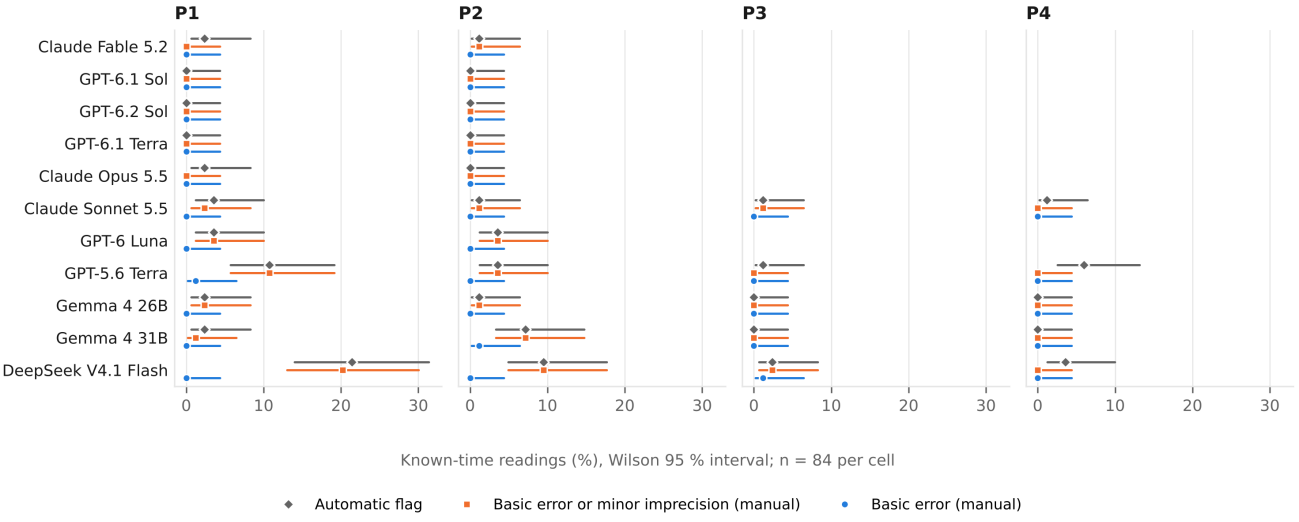


Figure 3. Eleven models with manually reviewed detector flags: readings flagged by the detector, readings with a basic error or minor imprecision after manual review, and readings with a basic error. Wilson 95 % intervals, n = 84 per cell.

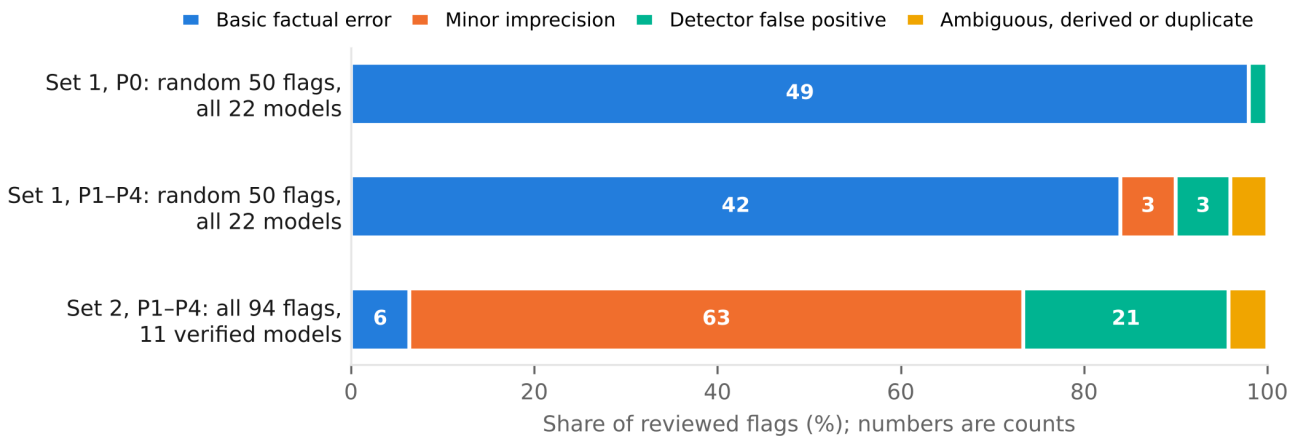


Figure 4. Manual classification of detector flags in the three reviewed sets. Numbers are counts.

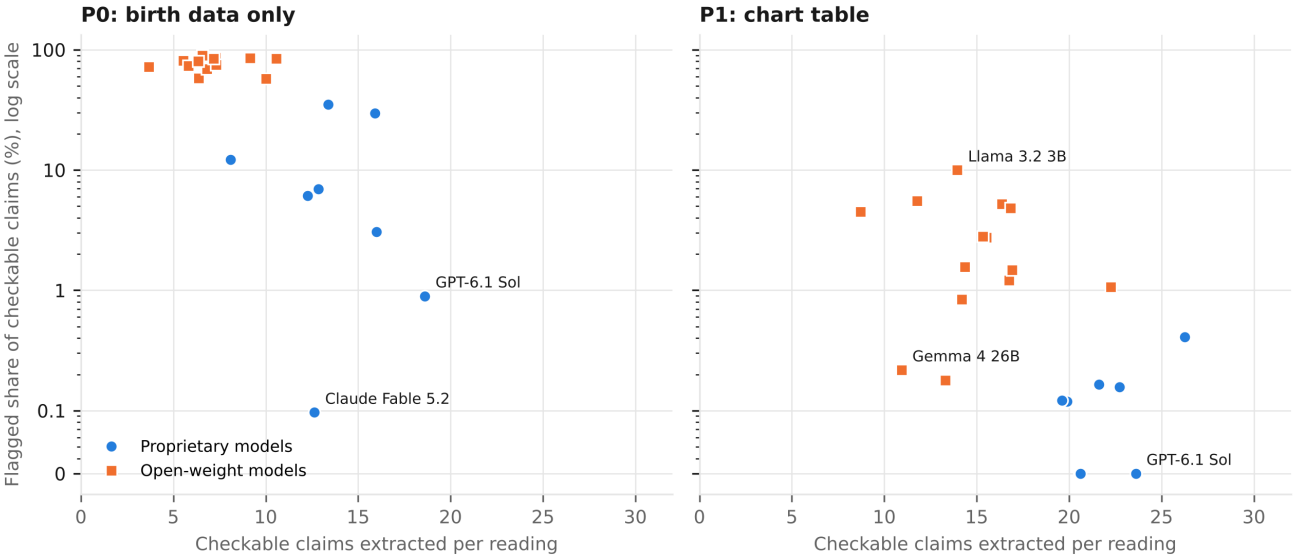


Figure 5. Checkable claims per reading against the flagged share of claims, per model. Logarithmic vertical axis; models with no flagged claim are plotted at the position marked 0.

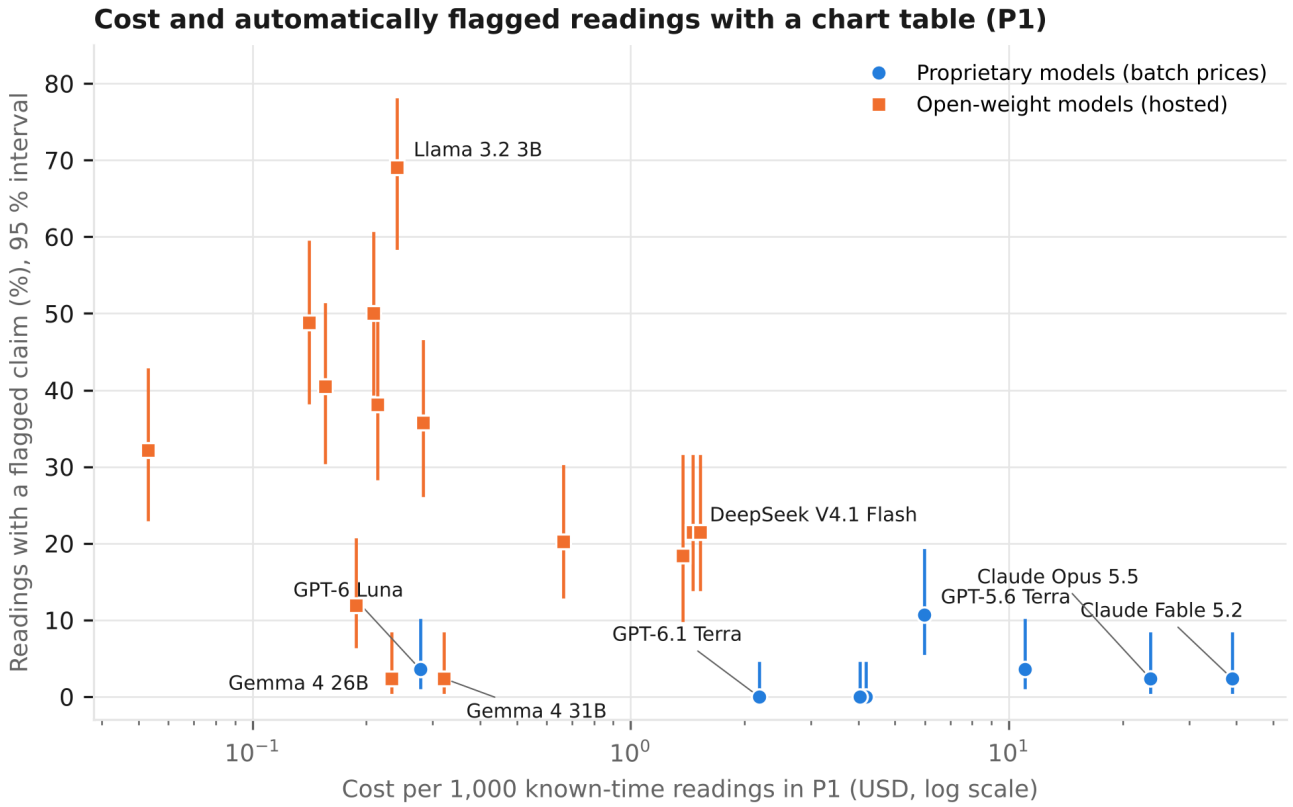


Figure 6. Cost per 1,000 completed known-time readings with a chart table (P1) against the share of readings with a flagged claim (Wilson 95 % intervals). Logarithmic horizontal axis. Flags are detector output, not verified errors.

Supplementary tables

Tables S1–S9 are computed from the per-reading results. In Table S6, "Flags/100 claims" is the share of checkable claims flagged by the detector, not a manually confirmed error rate. In Table S7, "Completed" counts completed requests, including resubmissions of the same reading for four models in P0 (Claude Fable 5.2, Claude Opus 5.5, Claude Sonnet 5.5 and GPT-6 Luna; 12,056 requests for 11,640 planned readings), so its total exceeds the 11,121 readings analysed (Table S5). The costs per reading quoted in the text divide the total cost of a condition by its unique completed readings. Model names follow the per-reading analysis and differ slightly from the main text (for example "Ministral 3 8B" for Ministral 8B, "Phi-4" for Phi-4 14B). "n/i" marks a relative reduction that is not interpretable because the baseline is below 5 %.

Table S1. Paired comparison of P0 and P1 (known birth time)

Pairs = same birth and repetition index in both conditions. Difference in percentage points with 95 % interval from a cluster bootstrap over births (4000 resamples). Tests are in Table S9.

Model	Pairs	Before	After	Difference (pp, 95 % CI)	Relative reduction	Discordant (only before / only after)
Claude Fable 5.2	82	1.2 %	1.2 %	+0.0 (-3.6 to +3.6)	n/i	1 / 1
Claude Opus 5.5	84	50.0 %	2.4 %	-47.6 (-59.5 to -35.7)	95.2 %	41 / 1
Claude Sonnet 5.5	84	58.3 %	3.6 %	-54.8 (-67.9 to -40.5)	93.9 %	48 / 2
GPT-6.2 Sol	84	19.0 %	0.0 %	-19.0 (-28.6 to -10.7)	100.0 %	16 / 0
GPT-6.1 Sol	84	8.3 %	0.0 %	-8.3 (-15.5 to -2.4)	100.0 %	7 / 0
GPT-6.1 Terra	84	44.0 %	0.0 %	-44.0 (-54.8 to -33.3)	100.0 %	37 / 0
GPT-5.6 Terra	84	96.4 %	10.7 %	-85.7 (-94.0 to -75.0)	88.9 %	72 / 0
GPT-6 Luna	84	97.6 %	3.6 %	-94.0 (-98.8 to -86.9)	96.3 %	79 / 0
Gemma 4 31B	84	100.0 %	2.4 %	-97.6 (-100.0 to -94.0)	97.6 %	82 / 0
Gemma 4 26B A4B	84	100.0 %	2.4 %	-97.6 (-100.0 to -94.0)	97.6 %	82 / 0
Gemma 3 4B	84	100.0 %	32.1 %	-67.9 (-77.4 to -57.1)	67.9 %	57 / 0
Mistral Small 3.2 24B	84	96.4 %	11.9 %	-84.5 (-94.0 to -72.6)	87.7 %	71 / 0
Ministral 3 8B	84	100.0 %	50.0 %	-50.0 (-61.9 to -38.1)	50.0 %	42 / 0
Ministral 3 3B	84	100.0 %	48.8 %	-51.2 (-59.5 to -42.9)	51.2 %	43 / 0
Llama 3.1 8B	84	100.0 %	35.7 %	-64.3 (-73.8 to -54.8)	64.3 %	54 / 0
Llama 3.2 3B	84	100.0 %	69.0 %	-31.0 (-40.5 to -21.4)	31.0 %	26 / 0
Qwen 3.5 9B	84	89.3 %	21.4 %	-67.9 (-77.4 to -57.1)	76.0 %	58 / 1
gpt-oss-20b	84	98.8 %	38.1 %	-60.7 (-67.9 to -51.2)	61.4 %	52 / 1
Phi-4	84	100.0 %	40.5 %	-59.5 (-70.2 to -48.8)	59.5 %	50 / 0
Granite 4.2 8B	84	95.2 %	20.2 %	-75.0 (-84.5 to -65.5)	78.8 %	63 / 0
LFM3 1.7B	19	89.5 %	21.1 %	-68.4 (-88.9 to -47.4)	76.5 %	13 / 0

Table S2. Paired comparison of P1 and P2

Pairs = same birth and repetition index in both conditions. Difference in percentage points with 95 % interval from a cluster bootstrap over births (4000 resamples). Tests are in Table S9.

Model	Pairs	Before	After	Difference (pp, 95 % CI)	Relative reduction	Discordant (only before / only after)
Claude Fable 5.2	84	2.4 %	1.2 %	-1.2 (-4.8 to +2.4)	n/i	2 / 1
Claude Opus 5.5	84	2.4 %	0.0 %	-2.4 (-6.0 to +0.0)	n/i	2 / 0
Claude Sonnet 5.5	84	3.6 %	1.2 %	-2.4 (-6.0 to +0.0)	n/i	2 / 0
GPT-6.2 Sol	84	0.0 %	0.0 %	+0.0 (+0.0 to +0.0)	n/i	0 / 0
GPT-6.1 Sol	84	0.0 %	0.0 %	+0.0 (+0.0 to +0.0)	n/i	0 / 0
GPT-6.1 Terra	84	0.0 %	0.0 %	+0.0 (+0.0 to +0.0)	n/i	0 / 0
GPT-5.6 Terra	84	10.7 %	3.6 %	-7.1 (-14.3 to -1.2)	66.7 %	6 / 0
GPT-6 Luna	84	3.6 %	3.6 %	+0.0 (-6.0 to +4.8)	n/i	3 / 3
Gemma 4 31B	84	2.4 %	7.1 %	+4.8 (-1.2 to +11.9)	n/i	1 / 5
Gemma 4 26B A4B	84	2.4 %	1.2 %	-1.2 (-3.6 to +0.0)	n/i	2 / 1
Gemma 3 4B	84	32.1 %	26.2 %	-6.0 (-21.4 to +10.7)	18.5 %	18 / 13
Mistral Small 3.2 24B	84	11.9 %	16.7 %	+4.8 (-7.1 to +15.5)	-40.0 %	8 / 12
Ministral 3 8B	84	50.0 %	54.8 %	+4.8 (-11.9 to +20.2)	-9.5 %	15 / 19
Ministral 3 3B	84	48.8 %	47.6 %	-1.2 (-13.1 to +10.7)	2.4 %	20 / 19
Llama 3.1 8B	84	35.7 %	26.2 %	-9.5 (-20.2 to +1.2)	26.7 %	23 / 15
Llama 3.2 3B	84	69.0 %	39.3 %	-29.8 (-46.4 to -13.1)	43.1 %	32 / 7
Qwen 3.5 9B	84	21.4 %	26.2 %	+4.8 (-9.5 to +20.2)	-22.2 %	16 / 20
gpt-oss-20b	84	38.1 %	42.9 %	+4.8 (-7.1 to +16.7)	-12.5 %	17 / 21
Phi-4	84	40.5 %	42.9 %	+2.4 (-13.1 to +16.7)	-5.9 %	18 / 20
Granite 4.2 8B	23	4.3 %	21.7 %	+17.4 (+4.3 to +33.3)	n/i	0 / 4
DeepSeek V4.1 Flash	84	21.4 %	9.5 %	-11.9 (-23.8 to -1.2)	55.6 %	16 / 6

Table S3. Paired comparison of P1 and P3

Pairs = same birth and repetition index in both conditions. Difference in percentage points with 95 % interval from a cluster bootstrap over births (4000 resamples). Tests are in Table S9.

Model	Pairs	Before	After	Difference (pp, 95 % CI)	Relative reduction	Discordant (only before / only after)
Claude Sonnet 5.5	84	3.6 %	1.2 %	-2.4 (-6.0 to +0.0)	n/i	2 / 0
GPT-5.6 Terra	84	10.7 %	1.2 %	-9.5 (-20.2 to -1.2)	88.9 %	9 / 1
Gemma 4 31B	84	2.4 %	0.0 %	-2.4 (-6.0 to +0.0)	n/i	2 / 0
Gemma 4 26B A4B	84	2.4 %	0.0 %	-2.4 (-6.0 to +0.0)	n/i	2 / 0
Gemma 3 4B	84	32.1 %	28.6 %	-3.6 (-14.3 to +7.1)	11.1 %	18 / 15
Mistral Small 3.2 24B	84	11.9 %	13.1 %	+1.2 (-11.9 to +11.9)	-10.0 %	8 / 9
Ministral 3 8B	84	50.0 %	51.2 %	+1.2 (-17.9 to +19.0)	-2.4 %	23 / 24
Ministral 3 3B	84	48.8 %	45.2 %	-3.6 (-16.7 to +9.5)	7.3 %	22 / 19

Model	Pairs	Before	After	Difference (pp, 95 % CI)	Relative reduction	Discordant (only before / only after)
Llama 3.1 8B	84	35.7 %	17.9 %	-17.9 (-28.6 to -7.1)	50.0 %	24 / 9
Llama 3.2 3B	84	69.0 %	76.2 %	+7.1 (-8.3 to +21.4)	-10.3 %	15 / 21
Qwen 3.5 9B	84	21.4 %	5.9 %	-15.5 (-26.2 to -3.6)	72.2 %	18 / 5
gpt-oss-20b	84	38.1 %	36.9 %	-1.2 (-14.3 to +13.1)	3.1 %	18 / 17
Phi-4	84	40.5 %	65.5 %	+25.0 (+9.5 to +39.3)	-61.8 %	13 / 34
Granite 4.2 8B	84	20.2 %	8.3 %	-11.9 (-23.8 to +0.0)	58.8 %	16 / 6
LFM3 1.7B	27	7.4 %	40.7 %	+33.3 (+16.0 to +51.7)	-450.0 %	0 / 9
DeepSeek V4.1 Flash	84	21.4 %	2.4 %	-19.0 (-28.6 to -9.5)	88.9 %	18 / 2

Table S4. Paired comparison of P3 and P4

Pairs = same birth and repetition index in both conditions. Difference in percentage points with 95 % interval from a cluster bootstrap over births (4000 resamples). Tests are in Table S9.

Model	Pairs	Before	After	Difference (pp, 95 % CI)	Relative reduction	Discordant (only before / only after)
Claude Sonnet 5.5	84	1.2 %	1.2 %	+0.0 (-3.6 to +3.6)	n/i	1 / 1
GPT-5.6 Terra	84	1.2 %	5.9 %	+4.8 (+1.2 to +9.5)	n/i	1 / 5
Gemma 4 31B	84	0.0 %	0.0 %	+0.0 (+0.0 to +0.0)	n/i	0 / 0
Gemma 4 26B A4B	84	0.0 %	0.0 %	+0.0 (+0.0 to +0.0)	n/i	0 / 0
Gemma 3 4B	84	28.6 %	39.3 %	+10.7 (-1.2 to +21.4)	-37.5 %	15 / 24
Mistral Small 3.2 24B	84	13.1 %	5.9 %	-7.1 (-14.3 to +0.0)	54.5 %	9 / 3
Ministral 3 8B	84	51.2 %	39.3 %	-11.9 (-27.4 to +3.6)	23.3 %	24 / 14
Ministral 3 3B	84	45.2 %	40.5 %	-4.8 (-17.9 to +9.5)	10.5 %	19 / 15
Llama 3.1 8B	84	17.9 %	13.1 %	-4.8 (-15.5 to +6.0)	26.7 %	15 / 11
Llama 3.2 3B	84	76.2 %	70.2 %	-6.0 (-20.2 to +8.3)	7.8 %	18 / 13
Qwen 3.5 9B	83	6.0 %	9.6 %	+3.6 (-6.0 to +12.2)	-60.0 %	5 / 8
gpt-oss-20b	84	36.9 %	45.2 %	+8.3 (-8.3 to +25.0)	-22.6 %	17 / 24
Granite 4.2 8B	84	8.3 %	4.8 %	-3.6 (-9.5 to +2.4)	42.9 %	6 / 3
LFM3 1.7B	30	36.7 %	36.7 %	+0.0 (-20.6 to +20.7)	0.0 %	7 / 7
DeepSeek V4.1 Flash	84	2.4 %	3.6 %	+1.2 (-2.4 to +4.8)	n/i	1 / 2

Table S5. Requested and completed readings per model and condition

All 40 births; 120 requests per model and condition (28 births with known time × 3, 12 without time × 3). Completed = generation finished with text (after the documented reruns).

Model	Type	P0	P1	P2	P3	P4
Claude Fable 5.2	proprietary	118/120	120/120	120/120	–	–
Claude Opus 5.5	proprietary	120/120	120/120	120/120	–	–
Claude Sonnet 5.5	proprietary	120/120	120/120	120/120	120/120	120/120

Model	Type	P0	P1	P2	P3	P4
GPT-6.2 Sol	proprietary	120/120	120/120	120/120	–	–
GPT-6.1 Sol	proprietary	120/120	120/120	120/120	–	–
GPT-6.1 Terra	proprietary	120/120	120/120	120/120	–	–
GPT-5.6 Terra	proprietary	120/120	120/120	120/120	120/120	120/120
GPT-6 Luna	proprietary	120/120	120/120	120/120	–	–
Gemma 4 31B	open-weight	120/120	120/120	120/120	120/120	120/120
Gemma 4 26B A4B	open-weight	120/120	120/120	120/120	120/120	120/120
Gemma 3 4B	open-weight	120/120	120/120	120/120	120/120	120/120
Mistral Small 3.2 24B	open-weight	120/120	120/120	120/120	120/120	120/120
Ministral 3 8B	open-weight	120/120	120/120	120/120	120/120	120/120
Ministral 3 3B	open-weight	120/120	120/120	120/120	120/120	120/120
Llama 3.1 8B	open-weight	120/120	120/120	120/120	120/120	120/120
Llama 3.2 3B	open-weight	120/120	120/120	120/120	120/120	120/120
Qwen 3.5 9B	open-weight	120/120	120/120	120/120	120/120	119/120
gpt-oss-20b	open-weight	120/120	120/120	120/120	120/120	120/120
Phi-4	open-weight	120/120	120/120	120/120	120/120	–
Granite 4.2 8B	open-weight	120/120	120/120	35/120	120/120	120/120
LFM3 1.7B	open-weight	51/120	80/120	0/120	63/120	89/120
DeepSeek V4.1 Flash	open-weight	6/120	120/120	120/120	120/120	120/120

Table S6. Reading length, checkable claims and claim-level flagged rate, P0 and P1

Claims = sign, Ascendant/Midheaven sign, house and aspect statements found by the extractor. Flags per 100 claims with 95 % Wilson interval (claims treated as independent; within-text clustering not modelled).

Model	Words P0	Words P1	Claims/text P0	Claims/text P1	Flags/100 claims P0	Flags/100 claims P1
Claude Fable 5.2	568	585	12.6	19.9	0.1 (0.0–0.5)	0.1 (0.0–0.4)
Claude Opus 5.5	567	574	12.3	19.6	6.1 (4.8–7.7)	0.1 (0.0–0.4)
Claude Sonnet 5.5	587	575	8.1	21.6	12.2 (10.0–14.9)	0.2 (0.1–0.5)
GPT-6.2 Sol	505	527	16.0	20.6	3.1 (2.3–4.1)	0.0 (0–0.2)
GPT-6.1 Sol	528	546	18.6	23.6	0.9 (0.5–1.5)	0.0 (0–0.2)
GPT-6.1 Terra	488	521	12.9	20.6	6.9 (5.6–8.6)	0.0 (0–0.2)
GPT-5.6 Terra	606	669	15.9	26.3	29.5 (27.2–32.0)	0.4 (0.2–0.8)
GPT-6 Luna	521	546	13.4	22.7	35.0 (32.2–37.8)	0.2 (0.1–0.5)
Gemma 4 31B	481	472	10.0	13.3	57.6 (54.3–60.9)	0.2 (0.0–0.6)
Gemma 4 26B A4B	501	501	6.4	10.9	58.1 (53.9–62.2)	0.2 (0.1–0.8)
Gemma 3 4B	707	579	7.3	8.7	86.4 (83.5–88.9)	4.5 (3.2–6.3)
Mistral Small 3.2 24B	569	546	6.8	14.2	69.6 (65.7–73.2)	0.8 (0.5–1.5)
Ministral 3 8B	762	716	10.6	16.8	84.1 (81.6–86.4)	4.8 (3.8–6.1)
Ministral 3 3B	664	753	6.6	16.4	89.5 (86.7–91.8)	5.2 (4.2–6.5)
Llama 3.1 8B	494	527	7.3	15.5	75.2 (71.7–78.5)	2.8 (2.0–3.8)
Llama 3.2 3B	521	522	6.4	14.0	80.7 (77.2–83.8)	10.0 (8.4–11.8)
Qwen 3.5 9B	475	470	3.7	14.4	71.8 (66.6–76.6)	1.6 (1.0–2.4)
gpt-oss-20b	750	743	7.2	15.3	84.4 (81.3–87.1)	2.8 (2.0–3.8)
Phi-4	675	614	9.2	11.8	85.3 (82.6–87.6)	5.6 (4.3–7.2)
Granite 4.2 8B	652	598	5.8	16.9	73.8 (69.7–77.5)	1.5 (1.0–2.2)
LFM3 1.7B	426	500	5.5	16.8	81.0 (75.3–85.6)	1.2 (0.7–2.2)
DeepSeek V4.1 Flash	–	503	–	22.3	–	1.1 (0.7–1.6)

Table S7. Tokens and cost per model and condition

Means per request. Reasoning tokens are reported separately only where the API reports them (OpenAI and OpenRouter); Anthropic counts thinking in output tokens. Cost includes failed and repeated requests; closed models at list batch prices valid on 24–26 Nov 2026 (−50 %), cached input not discounted (upper bound); open-weight models at billed OpenRouter cost.

Model	Condition	Requests (of which reruns)	Completed	Input	Cached	Output	Reasoning	USD total	USD on failed	USD per completed request
Claude Fable 5.2	P0	276 (156)	144	271	0	10241	0	56.831	27.984	0.39466
Claude Fable 5.2	P1	120 (0)	120	970	0	1753	0	4.674	0.0	0.03895
Claude Fable 5.2	P2	120 (0)	120	1155	0	2467	0	6.475	0.0	0.05396
Claude Opus 5.5	P0	240 (120)	238	271	0	4443	0	10.793	0.161	0.04535
Claude Opus 5.5	P1	120 (0)	120	970	0	2173	0	2.841	0.0	0.02367
Claude Opus 5.5	P2	120 (0)	120	1155	0	2266	0	2.997	0.0	0.02497
Claude Sonnet 5.5	P0	247 (127)	201	271	0	7109	0	8.846	2.133	0.04401
Claude Sonnet 5.5	P1	120 (0)	120	970	0	2011	0	1.323	0.0	0.01102
Claude Sonnet 5.5	P2	120 (0)	120	1155	0	2715	0	1.767	0.0	0.01473
Claude Sonnet 5.5	P3	120 (0)	120	9859	0	1910	0	2.329	0.0	0.01941
Claude Sonnet 5.5	P4	120 (0)	120	22446	0	1908	0	3.838	0.0	0.03198
DeepSeek V4.1 Flash	P0	120 (0)	6	201	0	15925	15886	0.806	0.769	0.13434
DeepSeek V4.1 Flash	P1	120 (0)	120	661	385	3386	2733	0.175	0.0	0.00146
DeepSeek V4.1 Flash	P2	120 (0)	120	811	484	6729	5897	0.345	0.0	0.00287
DeepSeek V4.1 Flash	P3	120 (0)	120	6341	3923	2437	1771	0.165	0.0	0.00138
DeepSeek V4.1 Flash	P4	120 (0)	120	14727	9578	2141	1474	0.199	0.0	0.00166
Gemma 3 4B	P0	120 (0)	120	202	158	895	0	0.006	0.0	5e-05
Gemma 3 4B	P1	120 (0)	120	688	512	717	0	0.006	0.0	5e-05
Gemma 3 4B	P2	120 (0)	120	875	607	792	0	0.007	0.0	6e-05
Gemma 3 4B	P3	120 (0)	120	6223	901	743	0	0.032	0.0	0.00027
Gemma 3 4B	P4	120 (0)	120	14045	2918	841	0	0.065	0.0	0.00055
Gemma 4 26B A4B	P0	120 (0)	120	206	126	601	0	0.023	0.0	0.00019
Gemma 4 26B A4B	P1	120 (0)	120	692	319	600	0	0.028	0.0	0.00023
Gemma 4 26B A4B	P2	120 (0)	120	879	387	644	0	0.031	0.0	0.00026
Gemma 4 26B A4B	P3	120 (0)	120	6227	3468	616	0	0.076	0.0	0.00063
Gemma 4 26B A4B	P4	120 (0)	120	14049	6739	629	0	0.151	0.0	0.00126

Model	Condition	Requests (of which reruns)	Completed	Input	Cached	Output	Reasoning	USD total	USD on failed	USD per completed request
Gemma 4 31B	P0	120 (0)	120	207	0	565	0	0.031	0.0	0.00026
Gemma 4 31B	P1	120 (0)	120	693	0	559	0	0.038	0.0	0.00032
Gemma 4 31B	P2	120 (0)	120	880	0	621	0	0.045	0.0	0.00037
Gemma 4 31B	P3	120 (0)	120	6228	0	603	0	0.134	0.0	0.00111
Gemma 4 31B	P4	120 (0)	120	14049	6010	603	0	0.265	0.0	0.00221
GPT-5.6 Terra	P0	120 (0)	120	181	0	3512	2778	2.55	0.0	0.02125
GPT-5.6 Terra	P1	120 (0)	120	643	0	888	68	0.716	0.0	0.00597
GPT-5.6 Terra	P2	120 (0)	120	761	0	1066	139	0.859	0.0	0.00716
GPT-5.6 Terra	P3	120 (0)	120	5947	3962	930	102	1.383	0.0	0.01153
GPT-5.6 Terra	P4	120 (0)	120	14023	9347	943	96	2.361	0.0	0.01968
GPT-6 Luna	P0	133 (13)	120	181	0	5289	4720	0.197	0.026	0.00164
GPT-6 Luna	P1	120 (0)	120	643	0	985	323	0.033	0.0	0.00028
GPT-6 Luna	P2	120 (0)	120	761	0	1319	507	0.044	0.0	0.00037
GPT-6.1 Sol	P0	120 (0)	120	181	0	3166	2526	1.921	0.0	0.01601
GPT-6.1 Sol	P1	120 (0)	120	643	0	710	53	0.503	0.0	0.00419
GPT-6.1 Sol	P2	120 (0)	120	761	0	874	138	0.615	0.0	0.00513
GPT-6.1 Terra	P0	120 (0)	120	181	0	3398	2830	1.03	0.0	0.00858
GPT-6.1 Terra	P1	120 (0)	120	643	0	746	137	0.262	0.0	0.00219
GPT-6.1 Terra	P2	120 (0)	120	761	0	919	242	0.321	0.0	0.00268
GPT-6.2 Sol	P0	120 (0)	120	181	0	2474	1887	1.506	0.0	0.01255
GPT-6.2 Sol	P1	120 (0)	120	643	0	678	65	0.484	0.0	0.00403
GPT-6.2 Sol	P2	120 (0)	120	761	0	795	98	0.568	0.0	0.00473
Granite 4.2 8B	P0	120 (0)	120	198	166	4807	3979	0.087	0.0	0.00073
Granite 4.2 8B	P1	120 (0)	120	680	397	4278	3393	0.08	0.0	0.00066
Granite 4.2 8B	P2	120 (0)	35	831	488	14803	14590	0.27	0.206	0.00771
Granite 4.2 8B	P3	120 (0)	120	6020	3588	4762	3891	0.11	0.0	0.00091
Granite 4.2 8B	P4	120 (0)	120	14041	8617	4477	3604	0.135	0.0	0.00113
LFM3 1.7B	P0	120 (0)	51	189	0	7400	7127	0.179	0.111	0.00352
LFM3 1.7B	P1	120 (0)	80	686	354	6691	6129	0.165	0.065	0.00206
LFM3 1.7B	P2	120 (0)	0	826	474	8000	7995	0.197	0.197	
LFM3 1.7B	P3	120 (0)	63	6296	3945	7065	6661	0.202	0.106	0.0032
LFM3 1.7B	P4	120 (0)	89	14461	9557	6361	5754	0.223	0.064	0.0025
Llama 3.1 8B	P0	120 (0)	120	192	120	594	0	0.021	0.0	0.00017
Llama 3.1 8B	P1	120 (0)	120	670	129	618	0	0.034	0.0	0.00028
Llama 3.1 8B	P2	120 (0)	120	785	122	616	0	0.037	0.0	0.00031

Model	Condition	Requests (of which reruns)	Completed	Input	Cached	Output	Reasoning	USD total	USD on failed	USD per completed request
Llama 3.1 8B	P3	120 (0)	120	5922	104	637	0	0.173	0.0	0.00144
Llama 3.1 8B	P4	120 (0)	120	13889	140	640	0	0.384	0.0	0.0032
Llama 3.2 3B	P0	120 (0)	120	192	159	633	0	0.026	0.0	0.00022
Llama 3.2 3B	P1	120 (0)	120	670	364	630	0	0.029	0.0	0.00024
Llama 3.2 3B	P2	120 (0)	120	785	465	564	0	0.027	0.0	0.00023
Llama 3.2 3B	P3	120 (0)	120	5922	1991	582	0	0.059	0.0	0.00049
Llama 3.2 3B	P4	120 (0)	120	13889	5830	623	0	0.108	0.0	0.0009
Phi-4	P0	120 (0)	120	187	0	827	0	0.015	0.0	0.00013
Phi-4	P1	120 (0)	120	665	0	785	0	0.019	0.0	0.00016
Phi-4	P2	120 (0)	120	780	0	811	0	0.02	0.0	0.00017
Phi-4	P3	120 (0)	120	5917	0	731	0	0.062	0.0	0.00052
Ministral 3 3B	P0	120 (0)	120	192	127	887	0	0.013	0.0	0.00011
Ministral 3 3B	P1	120 (0)	120	708	471	998	0	0.017	0.0	0.00014
Ministral 3 3B	P2	120 (0)	120	885	591	1198	0	0.02	0.0	0.00017
Ministral 3 3B	P3	120 (0)	120	7084	4390	980	0	0.054	0.0	0.00045
Ministral 3 3B	P4	120 (0)	120	15659	10162	1127	0	0.101	0.0	0.00084
Ministral 3 8B	P0	120 (0)	120	192	128	1047	0	0.022	0.0	0.00019
Ministral 3 8B	P1	120 (0)	120	708	471	984	0	0.025	0.0	0.00021
Ministral 3 8B	P2	120 (0)	120	885	590	1077	0	0.028	0.0	0.00024
Ministral 3 8B	P3	120 (0)	120	7084	4528	957	0	0.079	0.0	0.00065
Ministral 3 8B	P4	120 (0)	120	15659	10443	1032	0	0.144	0.0	0.0012
Mistral Small 3.2 24B	P0	120 (0)	120	192	0	699	0	0.018	0.0	0.00015
Mistral Small 3.2 24B	P1	120 (0)	120	708	0	674	0	0.023	0.0	0.00019
Mistral Small 3.2 24B	P2	120 (0)	120	885	0	739	0	0.026	0.0	0.00021
Mistral Small 3.2 24B	P3	120 (0)	120	7084	0	755	0	0.082	0.0	0.00068
Mistral Small 3.2 24B	P4	120 (0)	120	15659	0	830	0	0.173	0.0	0.00144
gpt-oss-20b	P0	120 (0)	120	232	185	4539	3349	0.077	0.0	0.00064
gpt-oss-20b	P1	120 (0)	120	694	429	1427	459	0.026	0.0	0.00021
gpt-oss-20b	P2	120 (0)	120	812	477	3217	2296	0.056	0.0	0.00047

Model	Condition	Requests (of which reruns)	Completed	Input	Cached	Output	Reasoning	USD total	USD on failed	USD per completed request
gpt-oss-20b	P3	120 (0)	120	5998	2048	1688	680	0.046	0.0	0.00038
gpt-oss-20b	P4	120 (0)	120	14074	5749	1821	818	0.07	0.0	0.00058
Qwen 3.5 9B	P0	120 (0)	120	210	0	6022	5431	0.183	0.0	0.00153
Qwen 3.5 9B	P1	120 (0)	120	737	197	5815	5205	0.183	0.0	0.00153
Qwen 3.5 9B	P2	120 (0)	120	878	216	7847	6987	0.246	0.0	0.00205
Qwen 3.5 9B	P3	120 (0)	120	6399	1978	5467	4882	0.241	0.0	0.00201
Qwen 3.5 9B	P4	120 (0)	119	14340	4440	4965	4405	0.319	0.0	0.00268

Table S8. Multilingual check by language

10 births with known time × 2 repetitions per language. Counted: sentences that mention a zodiac sign and carry at least one row citation; an inconsistency means the sign belongs to none of the cited rows. Sentences without a citation were not evaluated (count shown). This measures citation consistency, not overall factual accuracy.

Model	Language	Interpretations	Sentences checked	Inconsistent (95 % CI)	Uncited sentences with a sign	Citations	Citations to non-existent rows
Claude Sonnet 5.5	en	20	182	0/182 (0.0 %; 0.0–2.1)	8	810	0
Claude Sonnet 5.5	cs	20	159	1/159 (0.6 %; 0.1–3.5)	12	686	0
Claude Sonnet 5.5	sk	20	196	0/196 (0.0 %; 0.0–1.9)	10	677	0
Claude Sonnet 5.5	pl	20	182	1/182 (0.5 %; 0.1–3.0)	7	659	0
Claude Sonnet 5.5	de	20	121	1/121 (0.8 %; 0.1–4.5)	69	706	0
Claude Sonnet 5.5	es	20	187	0/187 (0.0 %; 0.0–2.0)	3	670	0
Claude Sonnet 5.5	fr	20	159	0/159 (0.0 %; 0.0–2.4)	15	715	0
Claude Sonnet 5.5	it	20	183	0/183 (0.0 %; 0.0–2.1)	4	692	0
GPT-6.1 Sol	en	20	191	0/191 (0.0 %; 0.0–2.0)	0	804	0
GPT-6.1 Sol	cs	20	186	0/186 (0.0 %; 0.0–2.0)	0	754	0
GPT-6.1 Sol	sk	20	177	0/177 (0.0 %; 0.0–2.1)	0	697	0
GPT-6.1 Sol	pl	20	169	0/169 (0.0 %; 0.0–2.2)	0	760	0
GPT-6.1 Sol	de	20	167	0/167 (0.0 %; 0.0–2.2)	0	707	0
GPT-6.1 Sol	es	20	178	0/178 (0.0 %; 0.0–2.1)	0	666	0
GPT-6.1 Sol	fr	20	164	0/164 (0.0 %; 0.0–2.3)	0	637	0
GPT-6.1 Sol	it	20	168	0/168 (0.0 %; 0.0–2.2)	0	630	0
GPT-6.2 Sol	en	20	197	0/197 (0.0 %; 0.0–1.9)	0	785	0

Model	Language	Interpretations	Sentences checked	Inconsistent (95 % CI)	Uncited sentences with a sign	Citations	Citations to non-existent rows
GPT-6.2 Sol	cs	20	187	0/187 (0.0 %; 0.0–2.0)	0	776	0
GPT-6.2 Sol	sk	20	201	0/201 (0.0 %; 0.0–1.9)	0	829	0
GPT-6.2 Sol	pl	20	193	0/193 (0.0 %; 0.0–2.0)	0	819	0
GPT-6.2 Sol	de	20	180	0/180 (0.0 %; 0.0–2.1)	0	781	0
GPT-6.2 Sol	es	20	198	0/198 (0.0 %; 0.0–1.9)	0	729	0
GPT-6.2 Sol	fr	20	190	0/190 (0.0 %; 0.0–2.0)	0	709	0
GPT-6.2 Sol	it	20	190	0/190 (0.0 %; 0.0–2.0)	0	762	0
Gemma 4 26B A4B	en	20	136	14/136 (10.3 %; 6.2–16.5)	4	699	0
Gemma 4 26B A4B	cs	20	103	6/103 (5.8 %; 2.7–12.1)	1	577	0
Gemma 4 26B A4B	sk	20	100	6/100 (6.0 %; 2.8–12.5)	2	542	0
Gemma 4 26B A4B	pl	20	99	13/99 (13.1 %; 7.8–21.2)	0	482	0
Gemma 4 26B A4B	de	20	127	14/127 (11.0 %; 6.7–17.7)	8	600	0
Gemma 4 26B A4B	es	20	132	10/132 (7.6 %; 4.2–13.4)	0	580	1
Gemma 4 26B A4B	fr	20	136	13/136 (9.6 %; 5.7–15.7)	1	532	0
Gemma 4 26B A4B	it	20	126	8/126 (6.3 %; 3.3–12.0)	0	646	1
Gemma 4 31B	en	20	144	0/144 (0.0 %; 0.0–2.6)	0	688	0
Gemma 4 31B	cs	20	149	0/149 (0.0 %; 0.0–2.5)	1	476	0
Gemma 4 31B	sk	20	130	1/130 (0.8 %; 0.1–4.2)	3	483	0
Gemma 4 31B	pl	19	135	0/135 (0.0 %; 0.0–2.8)	1	505	0
Gemma 4 31B	de	18	106	3/106 (2.8 %; 1.0–8.0)	21	536	0
Gemma 4 31B	es	18	144	1/144 (0.7 %; 0.1–3.8)	0	542	0
Gemma 4 31B	fr	18	136	1/136 (0.7 %; 0.1–4.0)	0	517	0
Gemma 4 31B	it	19	138	0/138 (0.0 %; 0.0–2.7)	0	602	0
Mistral Small 3.2 24B	en	20	154	3/154 (1.9 %; 0.7–5.6)	4	725	0
Mistral Small 3.2 24B	cs	20	131	5/131 (3.8 %; 1.6–8.6)	5	585	0
Mistral Small 3.2 24B	sk	20	235	14/235 (6.0 %; 3.6–9.8)	3	547	0
Mistral Small 3.2 24B	pl	20	149	3/149 (2.0 %; 0.7–5.8)	15	620	0
Mistral Small 3.2 24B	de	20	106	3/106 (2.8 %; 1.0–8.0)	68	553	0

Model	Language	Interpretations	Sentences checked	Inconsistent (95 % CI)	Uncited sentences with a sign	Citations	Citations to non-existent rows
Mistral Small 3.2 24B	es	20	174	1/174 (0.6 %; 0.1–3.2)	0	715	0
Mistral Small 3.2 24B	fr	20	207	0/207 (0.0 %; 0.0–1.8)	1	566	0
Mistral Small 3.2 24B	it	20	165	7/165 (4.2 %; 2.1–8.5)	1	563	0

Table S9. Birth-level exact permutation tests of the differences between conditions

Unit of analysis: the birth. For each birth the flagged readings in the second condition minus those in the first are summed over its completed pairs; under the null hypothesis the sign of each per-birth sum is arbitrary, and the two-sided p-value is computed exactly from all sign assignments. Holm adjustment across the models within each comparison.

Comparison	Model	Pairs	Births	Births with fewer flagged readings	Births with more	p	p (Holm)
P0 to P1	Claude Fable 5.2	82	28	1	1	1.000	1.000
P0 to P1	Claude Opus 5.5	84	28	23	0	<0.001	<0.001
P0 to P1	Claude Sonnet 5.5	84	28	21	0	<0.001	<0.001
P0 to P1	GPT-5.6 Terra	84	28	26	0	<0.001	<0.001
P0 to P1	GPT-6 Luna	84	28	28	0	<0.001	<0.001
P0 to P1	GPT-6.1 Sol	84	28	6	0	0.031	0.062
P0 to P1	GPT-6.1 Terra	84	28	22	0	<0.001	<0.001
P0 to P1	GPT-6.2 Sol	84	28	12	0	<0.001	0.002
P0 to P1	Gemma 3 4B	84	28	27	0	<0.001	<0.001
P0 to P1	Gemma 4 26B	84	28	28	0	<0.001	<0.001
P0 to P1	Gemma 4 31B	84	28	28	0	<0.001	<0.001
P0 to P1	Granite 4.2 8B	84	28	28	0	<0.001	<0.001
P0 to P1	LFM3 1.7B	19	17	12	0	<0.001	0.002
P0 to P1	Llama 3.1 8B	84	28	27	0	<0.001	<0.001
P0 to P1	Llama 3.2 3B	84	28	19	0	<0.001	<0.001
P0 to P1	Ministral 3B	84	28	27	0	<0.001	<0.001
P0 to P1	Ministral 8B	84	28	24	0	<0.001	<0.001
P0 to P1	Mistral Small 3.2 24B	84	28	26	0	<0.001	<0.001
P0 to P1	Phi-4 14B	84	28	26	0	<0.001	<0.001
P0 to P1	Qwen 3.5 9B	84	28	27	0	<0.001	<0.001
P0 to P1	gpt-oss 20B	84	28	27	1	<0.001	<0.001
P1 to P2	Claude Fable 5.2	84	28	2	1	1.000	1.000
P1 to P2	Claude Opus 5.5	84	28	2	0	0.500	1.000
P1 to P2	Claude Sonnet 5.5	84	28	2	0	0.500	1.000
P1 to P2	DeepSeek V4.1 Flash	84	28	9	3	0.074	1.000
P1 to P2	GPT-5.6 Terra	84	28	4	0	0.125	1.000
P1 to P2	GPT-6 Luna	84	28	1	2	1.000	1.000
P1 to P2	GPT-6.1 Sol	84	28	0	0	1.000	1.000
P1 to P2	GPT-6.1 Terra	84	28	0	0	1.000	1.000
P1 to P2	GPT-6.2 Sol	84	28	0	0	1.000	1.000
P1 to P2	Gemma 3 4B	84	28	13	8	0.566	1.000
P1 to P2	Gemma 4 26B	84	28	1	0	1.000	1.000

Comparison	Model	Pairs	Births	Births with fewer flagged readings	Births with more	p	p (Holm)
P1 to P2	Gemma 4 31B	84	28	1	4	0.312	1.000
P1 to P2	Granite 4.2 8B	23	17	0	4	0.125	1.000
P1 to P2	Llama 3.1 8B	84	28	11	4	0.174	1.000
P1 to P2	Llama 3.2 3B	84	28	18	6	0.003	0.059
P1 to P2	Ministral 3B	84	28	10	8	1.000	1.000
P1 to P2	Ministral 8B	84	28	6	9	0.667	1.000
P1 to P2	Mistral Small 3.2 24B	84	28	4	8	0.549	1.000
P1 to P2	Phi-4 14B	84	28	9	11	0.873	1.000
P1 to P2	Qwen 3.5 9B	84	28	9	8	0.643	1.000
P1 to P2	gpt-oss 20B	84	28	5	9	0.585	1.000
P1 to P3	Claude Sonnet 5.5	84	28	2	0	0.500	1.000
P1 to P3	DeepSeek V4.1 Flash	84	28	12	0	<0.001	0.008
P1 to P3	GPT-5.6 Terra	84	28	5	1	0.156	1.000
P1 to P3	Gemma 3 4B	84	28	11	9	0.680	1.000
P1 to P3	Gemma 4 26B	84	28	2	0	0.500	1.000
P1 to P3	Gemma 4 31B	84	28	2	0	0.500	1.000
P1 to P3	Granite 4.2 8B	84	28	11	5	0.087	0.960
P1 to P3	LFM3 1.7B	27	19	0	8	0.008	0.109
P1 to P3	Llama 3.1 8B	84	28	13	3	0.009	0.116
P1 to P3	Llama 3.2 3B	84	28	6	12	0.446	1.000
P1 to P3	Ministral 3B	84	28	10	7	0.715	1.000
P1 to P3	Ministral 8B	84	28	9	11	1.000	1.000
P1 to P3	Mistral Small 3.2 24B	84	28	5	8	1.000	1.000
P1 to P3	Phi-4 14B	84	28	4	15	0.006	0.092
P1 to P3	Qwen 3.5 9B	84	28	16	4	0.023	0.274
P1 to P3	gpt-oss 20B	84	28	14	9	1.000	1.000
P3 to P4	Claude Sonnet 5.5	84	28	1	1	1.000	1.000
P3 to P4	DeepSeek V4.1 Flash	84	28	1	2	1.000	1.000
P3 to P4	GPT-5.6 Terra	84	28	0	4	0.125	1.000
P3 to P4	Gemma 3 4B	84	28	3	12	0.124	1.000
P3 to P4	Gemma 4 26B	84	28	0	0	1.000	1.000
P3 to P4	Gemma 4 31B	84	28	0	0	1.000	1.000
P3 to P4	Granite 4.2 8B	84	28	5	2	0.453	1.000
P3 to P4	LFM3 1.7B	30	20	5	5	1.000	1.000
P3 to P4	Llama 3.1 8B	84	28	10	6	0.529	1.000
P3 to P4	Llama 3.2 3B	84	28	14	5	0.543	1.000
P3 to P4	Ministral 3B	84	28	10	8	0.622	1.000
P3 to P4	Ministral 8B	84	28	14	7	0.196	1.000
P3 to P4	Mistral Small 3.2 24B	84	28	9	3	0.146	1.000
P3 to P4	Qwen 3.5 9B	83	28	4	8	0.613	1.000
P3 to P4	gpt-oss 20B	84	28	10	13	0.424	1.000

Appendices and supplementary material

Everything here is taken from the study's scripts and records. Where a detail is not recorded, the entry says so. The materials are prepared so that they can be released with the paper.

Appendix A. Prompts (P0–P3)

System instruction (all conditions; `{lang_name}` is "English" in the main study):

```
You are an experienced Western astrologer writing a natal chart interpretation for a general
reader.
Use the tropical zodiac, Placidus houses and the five Ptolemaic aspects.
Write {lang_name} prose of 400 to 600 words in three sections with short headings: personality,
relationships, work and vocation.
Mention concrete chart factors (planets, signs, houses, aspects) when you rely on them.
Do not write about health, illness, death, pregnancy, legal matters or financial investments.
If the birth time is unknown, say so once and do not use the Ascendant, the Midheaven or
houses.
```

User message, P0:

```
Date of birth: 1955-09-07
Time of birth: 02:42 local time
Place of birth: Seoul, South Korea (latitude 37.57, longitude 126.98)

Please write the interpretation.
```

For an unknown-time birth the second line reads `Time of birth: unknown`.

User message, P1: the three birth lines, then

```
Computed chart (geocentric, tropical, Placidus; computed from the JPL DE442s ephemeris):
<table, Appendix B, without row identifiers>

Please write the interpretation. Use only the chart factors listed above.
```

User message, P2: as P1 with row identifiers, and the closing instruction

```
Please write the interpretation. Use only the chart factors listed above. After every sentence
that mentions a chart factor, put the identifiers of the rows it relies on in square brackets,
e.g. [E3][E14].
```

User message, P3: the three birth lines, then

```
Computed chart data (geocentric, tropical, Placidus; computed from the JPL DE442s ephemeris):
<engine natal output as compact JSON, field interpretationRules removed>

Please write the interpretation. Use only the chart factors listed above.
```

The P3 JSON is the engine's natal output without interpretive text. It contains positions, houses, aspects with strength tiers, dignities, rulers, dispositors, element and modality balance, patterns and a priority list. The exact P0–P3 inputs sent to the models are archived with the study materials, together with every reading as returned. The P3 data were produced on 25 November 2026 by the engine at repository commit 47cb14d6 (ephemeris library go-orbitalis, revision 9650e77ff0c0), before the defects described in Appendix G were fixed; a current engine produces different P3 data, so replication of P3 should use the archived inputs.

P4 uses the same wrapper with the glossary of meanings and interpretation rules added. Its content is proprietary and is not reproduced.

Appendix B. Chart table format (example: birth B01, P2)

```
[E1] Sun: Virgo 13°20′, house 3
[E2] Moon: Taurus 9°05′, house 10
[E3] Mercury: Libra 7°20′, house 3
[E4] Venus: Virgo 14°48′, house 3
[E5] Mars: Virgo 6°33′, house 2
[E6] Jupiter: Leo 18°20′, house 2
[E7] Saturn: Scorpio 16°24′, house 4
[E8] Uranus: Leo 0°40′, house 1
[E9] Neptune: Libra 26°25′, house 4
[E10] Pluto: Leo 27°02′, house 2
[E11] North Node: Sagittarius 22°36′, house 5, retrograde
[E12] Ascendant: Cancer 25°18′
[E13] Midheaven: Aries 10°53′

Aspects (orb in degrees):
[E14] Sun trine Moon, orb 4.25
[E15] Sun conjunction Venus, orb 1.47
[E16] Sun conjunction Mars, orb 6.78
…
[E31] Pluto trine North Node, orb 4.43
```

Degrees are rounded to the minute. Orbs are in decimal degrees. For unknown-time births the house, Ascendant and Midheaven entries are omitted. The table lists every aspect within the permitted orb without a strength tier; rows such as `Moon square Jupiter, orb 9.26` are therefore presented on the same footing as tight aspects. This is a property of the study's input and is discussed in the manuscript (Section 3.4).

Appendix C. Reference chart conventions and orbs

Item	Value
Ephemeris	JPL DE442s, through the engine's own ephemeris library (library revision 9650e77ff0c0, the version pinned by the engine commit in Appendix A)
Frame	Geocentric, apparent, ecliptic of date; tropical zodiac
Time	Local civil time converted to UTC with the IANA time zone database; UTC instant stored per birth
Houses	Placidus
Lunar node	True (osculating) node
Bodies	Sun, Moon, Mercury to Pluto, North Node; Ascendant and Midheaven when time is known
Aspects	Conjunction, sextile, square, trine, opposition
Unknown time	Chart for 12:00 local time; houses, angles, and in the detector Moon claims, are excluded

Orbs. Each body has a moiety (half-orb). An aspect is within orb when the separation from exactness does not exceed the sum of the two moieties. If either body is non-traditional (Uranus, Neptune, Pluto, node, angle) the sum is capped at 6°.

Body	Moiety (°)	Body	Moiety (°)
Sun	8.5	Jupiter	6
Moon	6.25	Saturn	5
Mercury	3.5	Uranus, Neptune, Pluto	2.5
Venus	4	North Node	2
Mars	3.75	Ascendant, Midheaven	5

The values for the seven traditional planets are half of the orbs in the first column of the table in Lilly (1647, p. 107): Saturn 10°, Jupiter 12°, Mars 7°30′, Sun 17°, Venus 8°, Mercury 7° and Moon 12°30′. Lilly gives a second, mostly smaller set in the same table (9°, 9°, 7°, 15°, 7°, 7°, 12°) and says he uses sometimes the one and sometimes the other; the engine uses the first set. The values for the outer planets, the node and the angles are the engine's own convention, not tradition. Under this scheme the widest permitted aspect is Sun–Moon at 14.75°. Strength tiers in P3/P4: PARTILE up to 1°, CLOSE up to 3°, MODERATE up to 6°, WIDE above 6°.

Births. Forty synthetic births generated with seed 20261125: dates 1940-05-29 to 2009-01-05, 25 cities, latitudes −37.81° to 59.33°, 28 with a known time and 12 without. The script draws a UTC instant uniformly between 1940-01-01 and 2011-01-01 in whole minutes and a city uniformly from a fixed list; the first 28 draws keep their time. It was not designed to cover boundary cases.

Independent recomputation. All 40 reference charts were recomputed with the Swiss Ephemeris (version 2.10.03, its own ephemeris files), with local times converted to UTC independently of the engine. The largest differences were 0.27″ for the planets, 0.10″ for the Moon, 0.02″ for the true node and 0.46″ for the Ascendant, Midheaven and Placidus cusps. All 428 signs of bodies (eleven bodies including the true node in 40 charts, without the Moon in the 12 unknown-time charts), all 56 signs of the Ascendant and Midheaven and all 308 house placements were identical, and each of the 901 aspects listed by the engine was reproduced within its orb limit. In the opposite direction, applying the orb rule to the Swiss Ephemeris positions produced the same 901 aspects, with none missing from the engine's list and none listed by the engine only. The recomputation covers what the detector checks; it does not cover the derived fields discussed in Appendix G.

Appendix D. Models and generation parameters

Generation dates: 24–26 November 2026. No temperature, top-p, seed or reasoning-effort parameter was sent; provider defaults applied. No tools were defined in any request.

Proprietary models (vendor batch interfaces, 50 % batch discount).

Name in paper	API identifier	Interface	Output limit (tokens)	Conditions
Claude Fable 5.2	claude-fable-5-2	Anthropic Message Batches	8,000; P0 truncated readings resubmitted at 16,000, then 32,000	P0–P2
Claude Opus 5.5	claude-opus-5-5	same	same	P0–P2
Claude Sonnet 5.5	claude-sonnet-5-5	same	same; P3/P4 16,000	P0–P4
GPT-6.2 Sol	gpt-6.2-sol	OpenAI Batch, Responses endpoint	16,000	P0–P2
GPT-6.1 Sol	gpt-6.1-sol	same	8,000	P0–P2
GPT-6.1 Terra	gpt-6.1-terra	same	16,000	P0–P2
GPT-5.6 Terra	gpt-5.6-terra	same	16,000	P0–P4
GPT-6 Luna	gpt-6-luna	same	8,000; P0 resubmitted at 16,000	P0–P2

The OpenAI responses report a default reasoning effort of "medium".

Dated snapshot identifiers were not requested, so the aliases above resolved to whatever version each vendor served on those days. The model field returned in the responses repeats the alias and gives no dated version; the responses are archived.

Open-weight models (OpenRouter chat completions; one pinned provider per request, `allow_fallbacks: false`, `data_collection: deny`).

Name in paper	OpenRouter identifier	Provider / declared precision	Output limit
Gemma 4 31B	google/gemma-4-31b-it	Novita / bf16 (498 stored responses) and Crusoe / bf16 (234)	16,000

Name in paper	OpenRouter identifier	Provider / declared precision	Output limit
Gemma 4 26B	google/gemma-4-26b-a4b-it	CoreWeave / bf16	16,000
Gemma 3 4B	google/gemma-3-4b-it	Parasail / bf16	8,000
Mistral Small 3.2 24B	mistralai/mistral-small-3.2-24b-instruct	DeepInfra / fp8	8,000
Ministral 8B	mistralai/ministral-8b-2512	Mistral (EU) / not declared	8,000
Ministral 3B	mistralai/ministral-3b-2512	Mistral (EU) / not declared	8,000
Llama 3.1 8B	meta-llama/llama-3.1-8b-instruct	CoreWeave / bf16	8,000
Llama 3.2 3B	meta-llama/llama-3.2-3b-instruct	Parasail / bf16	8,000
Qwen 3.5 9B	qwen/qwen3.5-9b	Parasail / bf16	16,000
gpt-oss 20B	openai/gpt-oss-20b	DeepInfra / bf16 (15 responses on its flex tier)	16,000
Phi-4 14B	microsoft/phi-4	DeepInfra / bf16	8,000 (no P4: input exceeds the 16k context)
Granite 4.2 8B	ibm-granite/granite-4.2-8b	CoreWeave / bf16	16,000
LFM3 1.7B	liquid/lfm-3-1.7b	Liquid / not declared	8,000
DeepSeek V4.1 Flash	deepseek/deepseek-v4.1-flash	DeepInfra / fp8	16,000

Requests that failed with a transient HTTP error were retried up to five times. A reading cut off by the output limit was excluded from analysis. Requests, resubmissions, completed readings, tokens by type and cost per cell are in Supplementary Tables S5 and S7. Anthropic P0 cells are analysed on the 16,000-token run, replaced by the 32,000-token run where the former was cut off.

Appendix E. Automatic detector

The detector is a set of English regular expressions applied to the reading after removing markdown emphasis and row identifiers.

Claim type	Pattern (informally)	Flagged when
Sign	"BODY (is / sits / lies / falls / placed ...) in SIGN"	the reference sign of that body differs
Angle sign	"SIGN rising / Ascendant / Midheaven"	the reference sign of the angle differs
House	"BODY ... in the/your ORDINAL house"	the reference house differs
Aspect	"BODY conjunct / sextile / square / trine / opposite BODY", "BODY–BODY square" etc.	the reference chart has no aspect between the two bodies within orb, or one of another type

Known behaviours established by manual review (Appendix F): the detector attaches a house or sign to the nearest preceding body, so lists and "respectively" constructions produce false positives; it does not recognise qualifications such as "out of sign"; it treats a conjunction "in SIGN" as a sign claim for both bodies, which is the source of most minor-imprecision flags; hypothetical sentences are not excluded. It checks no degrees, orbs, retrograde status or derived claims, and it does not work outside English.

Unknown-time births: house and angle statements are counted as prohibited statements; Moon claims are skipped; other claims are checked against the noon chart.

Appendix F. Manual review protocol

The item-level records of all three reviewed sets, with the original and the revised label for each item, are part of the published materials.

1. Set 1 was drawn with random seed 20261126 from all flags: 50 from P0 (population 9,072) and 50 from P1–P4 (population 2,172). Populations count flags over all readings, including unknown-time births.

2. Set 2 is every P1–P4 flag (94) of the 11 models listed in the manuscript.
3. Thirty unflagged known-time readings of the same 11 models were drawn at random (seed 26112026 for the first eight) and read in full.
4. First classification: AI assistant, with the chart table for the birth at hand. Second pass: the author re-examined all 100 + 94 + 30 items and listed disagreements. Resolution: the disputed items were re-read in full sentences and one written rule was applied to both sets.

Rule.

- *Basic factual error*: a sign, house or aspect type explicitly assigned to the wrong body; an aspect that is not in the table.
- *Minor imprecision*: a group of planets or a conjunction across a sign or house boundary written with one sign or house, regardless of word order; an orb in the wrong notation (0.34° written as 0°34′); "close to the Ascendant" at 18°; "partile" at different degrees.
- *Derived error* (outside detector scope): dispositors, elements, dignities, "tightest aspect".
- *Detector false positive*: the sentence is correct (aspect to a third body; explicit "out of sign"; "respectively"; rulers; a conditional statement that is true in the data).
- *Ambiguous*: the grammar allows the house or sign to attach to either body.
- *Duplicate*: the same sentence flagged twice; counted once.

Overlap and counting. Two flags are in both sets. Duplicates (two in Set 2) are excluded from denominators, giving 92 unique flags. Table 2 counts readings; Table 3 counts flags. A reading with several flags is counted once in Table 2 under its most serious category.

Secondary review of Claude Fable 5.2 readings without a table. Eighteen of the 28 known-time births were drawn at random (seed 20261127) and one completed P0 repetition was drawn for each: B01 r1, B02 r3, B03 r3, B04 r3, B05 r2, B06 r1, B08 r3, B09 r3, B10 r1, B12 r1, B13 r2, B17 r3, B18 r3, B19 r3, B22 r3, B25 r2, B26 r2, B27 r3. Each reading was shown next to the reference chart table for its birth, without the detector's result. Two AI assistants (GPT-6.2 Sol and GPT-6 Astra) were each given this material by the author. The page itself carried the instruction to compare every stated sign, house and aspect with the table, to treat a wrong sign, house or aspect type or an aspect absent from the table as a basic error, and to note anything else found; any further wording used in the two sessions was not archived. Their reports were merged, every reported discrepancy was recomputed from the reference positions, and the author confirmed the retained ones. Six reported wide aspects were in fact listed in the reference chart and were dropped. The one reading flagged by the detector (B18 r2) was not among the 18. Retained: Saturn opposition Uranus (7.5°) and Saturn opposition Pluto (7.1°) in B18 r3, Sun conjunction Neptune (7.0°) in B27 r3, and Venus given at 24° for 22°42′ in B18 r3. One further sentence (B05 r2, a square from Neptune to a Venus–Mars conjunction; Mars–Neptune is listed, Venus–Neptune at 6.2° is not) was left as ambiguous.

What the review cannot support. Set 1 precision applies to the mixed population, Set 2 precision to the 11 models; neither transfers to the other. The 30 unflagged readings bound the share of unflagged readings with a missed basic error (0/30, upper bound 11.4 %) and do not estimate sensitivity.

Appendix G. Errors traced to the reference engine

Two derived statements judged wrong in the review reproduce the engine's P3/P4 output:

1. **Lacking element.** The element balance reported an element as lacking on the basis of the seven traditional planets and the Ascendant. A reading then said the chart had no planets in water although Uranus was in a

water sign. Counting only traditional planets is a defensible convention, but the output did not say so, and the statement as written was false.
2. **Final dispositor.** The priority list named one final dispositor although a second planet in its own sign also disposed of itself.

Both were corrected in the engine after the readings were generated (the output now names outer planets in a lacking element and lists several final dispositors together). The study's readings were not regenerated. These cases are counted as derived errors with the mechanism "source-data error" and are kept in the results.

Appendix H. Literature search

Searched on 26–28 November 2026: arXiv, the ACL Anthology and the open web (combinations of "large language model", "ChatGPT", "astrology", "natal chart", "horoscope", "ephemeris", "planetary position", "hallucination", "benchmark"); Google Scholar (five queries on the same terms, 33 results screened by title and snippet); and the sites of *Correlation* and the *Journal of Scientific Exploration*, whose archives could not be searched exhaustively. Scopus and Web of Science were not queried.

No study was found that measures the factual consistency of language-model readings of Western natal charts against a computed chart. The closest items concern the Chinese Four Pillars tradition, namely the two preprints cited in the Introduction (Chen & Liu, 2026; Han & Liu, 2026). One test that is not peer-reviewed documents its method: the vendor Cosine Astro asked three models for the sign and degree of seven bodies, the Ascendant and the Midheaven for ten births, answered as JSON, and reports 25 scored answers with the Sun sign correct in all and the Moon sign in one (https://cosineastro.com/en/ai/chatgpt-astrology-test, accessed 27 November 2026). It scores positions returned as structured data for one prompt; the present study scores statements in continuous readings and compares input conditions.

Appendix I. Statistical methods

- Wilson score interval for a proportion; Holm adjustment within each family of comparisons (one family per pair of conditions, across models); Spearman rank correlation across models.
- Per-reading analyses: one row per reading; pairs formed by birth and repetition index; cluster bootstrap over births with 4,000 resamples; word counts; tokens by type; cost including failed and repeated requests.
- Tests between conditions take the birth as the unit: the per-birth sum of differences over its completed pairs, an exact two-sided sign-flip permutation test over births, and Holm adjustment within each comparison (Supplementary Table S9).
- Claim counts in the tables, in Figure 5 and in the correlations of Section 3.5 all come from the per-reading extraction.