All in the Family? Lineage, Size, and Agreement in LLM Content Coding

Craig Singleton

The Ethics Observatory · ORCID 0009-0003-6737-0436 · csingleton3@toromail.csudh.edu

First report · October 2026

Abstract

Researchers increasingly use large language models (LLMs) as "independent" content coders. We examined agreement and repeatability using 65 frozen institutional documents, a 91-code instrument and 16 models from six vendors. Fifteen models were retested; each model's agreement with itself provided a comparison reference.

First, observed retest agreement ranged from .90 to .71 under the historical pooled nominal method. Second, among selected non-small tiers, raw same-vendor agreement averaged .783 and cross-vendor agreement .694. Dividing pair agreement by the geometric mean of the two retests gave descriptive means of .926 (95% document-bootstrap interval .911–.940) within vendors and .842 (.822–.860) across vendors. These ratios are not validated attenuation corrections. Third, an early Haiku run whose prompt was not saved agreed with two later runs at about .59, while the later runs agreed at .826. Fourth, API-versus-tool point differences were small on 25 documents, but intervals did not establish equivalence. Fifth, ethical-principle agreement patterns resembled our earlier study of the same corpus; selected cross-family pairs diverged more on interpretive principles. Sixth, output adherence did not vary monotonically with tier.

We recommend reporting retests for all LLM coders, saving exact prompts and settings, building panels across vendors, and avoiding treating same-vendor models as independent replications. This exploratory study measures agreement, not accuracy; family, route and incompletely recorded settings remain partly confounded.

1. Meet the Parents: Introduction

Two LLM coders agreed with one another at pooled nominal Krippendorff's alpha = .89. However, both models came from the same vendor company. Each also agreed with a second run of itself nearly to the same extent (.88 and .90). By this measure, the two models were as similar to each other as each was to its own repeat. Content analysis relies upon agreement between coders as an indicator of the reliability of a coding scheme (Krippendorff, 2018). Agreement between independently working coders is valuable, but does not establish that their errors are independent. Models from one vendor may share training data, architecture and tuning; benchmark studies find correlated errors associated with providers and architectures (Goel et al., 2025; Kim et al., 2025). Thus, similar results could reflect the document being analyzed, shared ancestry, or both. An alternative theory is that highly capable models converge regardless of who produces them. Below both theories lies another quantity: how well does a model reproduce its own codes? A model's self-

consistency affects the interpretation of every pair it belongs to. The conventional .80 reliability threshold is useful context, but the pooled nominal index here does not validate all 91 mixed-scale codes together.

We had 16 models from six vendor families code 65 documents using the same frozen instrument. Models were selected across vendor tiers, up to four per family. Of these 16 models we ran 15 of them at least twice. “Size” in this paper refers to the vendor’s tier, not a known parameter count. We ask three questions: how consistent each model is with itself; whether same-vendor models are more similar than cross-vendor models once each pair is compared with the models' own retests; and whether model tier affects how well models follow the instrument's output rules.

This work makes five contributions. It reports the observed test–retest agreement of 15 LLM coders on a full research instrument, with bootstrap intervals, and compares same-vendor with cross-vendor agreement relative to those retests. It also provides a direct test of the interface confound for two models, a comparison with our prior study of three vendor families, and a deviation log that preserves raw replies and records every transformation used for scoring.

This is a first report. It provides information about the experimental design and some of the primary data so that those interested may review it. A final report will provide a summary of the findings for all fields of interest. Additionally, a final report will include an analysis of deviations and a fuller comparison of interfaces. The study was exploratory and was not preregistered.

2. Keeping Up with the Joneses: Related Work

In Kim et al. (2025), the authors assessed over 350 models and reported a high degree of error correlation, with some of it associated with shared providers and architectures. Larger and more accurate models also had correlated errors across providers. This is an example of capability convergence, and represents a competing explanation for why models produced by different vendors exhibit similar patterns of behavior. In Goel et al. (2025), the authors proposed Chance Adjusted Probabilistic Agreement (CAPA) for assessing model similarity and demonstrated that LLM judges favor models with similar mistakes. In Kohli (2026), nine judges from seven families provided about two to two-and-a-half effective votes on the tested tasks; family diversity alone did not restore independence. In Awuni et al. (2026), four open-weight families at two scales showed a same-family preference after candidate-quality controls. In Najera et al. (2026), disagreement among four models coding 1,260 public comments was treated as diagnostic of interpretive complexity and directed human review. These studies already examine model similarity, scale and disagreement. Our contribution is to combine vendor-tier comparisons, retests and adherence checks on frozen full-document coding packets. The prior study conducted by Singleton (2026) used one principal configuration from each of three vendors on the same 65 documents and concluded that cross-family disagreement was diagnostic. It did not include a full within-family tier design or test-retest estimates. The current design adds both, while retaining limitations on causal attribution. The corpus and instrument are shared with that earlier study; the new model runs are separate measurements.

3. Downton Abbey: Data and Study Design

The dataset consists of 65 institutional documents concerning AI ethics and governance, which are divided into five categories — Government, Legal and Regulatory (15); Journal (14); Institutional (14); Public Forum (12); Corporate (10). Every coder receives each of the 65 documents as one frozen package including both the complete instrument and the full source text. The packet content is identical across coders; harness and launch instructions can differ, and some early launch prompts were not saved.

Instrument. Version 2.14 of the DPPRA protocol (Discourse-Policy-Practice-Resources-Accountability) supplies the coding instrument. The earlier study calls the wider workflow the Intercoder Stress Test (ICST); this study uses its coding pass, without the adversarial counter-reading or theoretical lens-swap exercises. Each document is to return 91 codes and 16 narrative fields in one JSON object, with no surrounding prose. Codes include 0–5 ratings for ten ethical principles (Equity, Accountability, Transparency, Privacy, Autonomy, Integrity, Non-Maleficence, Mechanical Beneficence, Ethical Beneficence, Justice), along with categorical and binary fields. Nominal and ordinal agreement are analyzed separately.

Table 1. Models, interfaces and repeat runs

Family	Models in vendor tier order	Interface	Repeat runs
Anthropic	Claude Fable 5.1, Opus 5.5, Sonnet 5.5, Haiku 4.5	Claude Code sub-agents	all four (Haiku twice)
OpenAI	GPT-6 Astra, GPT-6.1 Sol, GPT-6 Luna (medium effort; Luna also at high)	Codex app	all three (Sol twice)
Google	Gemini 3.1 Pro (preview), 3.8 Flash, 3.5 Flash-Lite	Vertex API, direct	all three
Qwen	Qwen 3.7 Max, Plus, Flash	OpenRouter, direct; not pinned, served by Alibaba	Max, Plus
DeepSeek	DeepSeek v4 Pro, v4 Flash	OpenRouter, host pinned to Alibaba	both
Moonshot	Kimi K2.6	OpenRouter, host pinned to Moonshot AI	yes

Note. Every main arm and full retest attempted the same 65 packets; saved-prompt and route differences are documented below. An earlier selected Astra repeat covered 20 documents; Table A1 uses its full retest. Sonnet and Sol were also run through OpenRouter on 25 documents (Section 4e). "Capable" in the tables and figures means the selected non-small tiers; accuracy or matched capability was not measured. Table A7 records model identifiers and settings.

The interface confound. The families were not operated via the same interface. Claude models operated as Claude Code sub-agents, one new agent per document. OpenAI models operated in the Codex app. Gemini models were called directly through the Vertex API. Qwen, DeepSeek and Kimi were called through OpenRouter. The host was requested as pinned for DeepSeek and Kimi, but three early provider histories remain ambiguous, one each for DeepSeek Pro, DeepSeek Flash and Kimi. Qwen was not pinned; observed final reply events name Alibaba. Family and route are therefore partly confounded in the main design, and we tested route sensitivity directly for two models (Section 4e). Exact Claude served snapshots and inherited thinking settings were not universally recorded.

Prompts. The first Opus, Sonnet and Haiku runs were launched before the sub-agent prompt was saved to a file. Every later Claude run, including both Fable runs and all Claude retests, used a saved prompt template. Section 4b shows why this matters.

Scoring. Raw replies were never edited. Historical eligibility requires a parsed JSON object with the mandatory meta, codes and narrative blocks, at least 91 codes, at least 16 narrative fields and integer codes. To preserve the original analysis, "strict" below means this count-based rule, including replies whose only recorded format deviation was one outer code fence, removed in a logged scoring copy. Lenient scoring additionally accepts a reply missing only its final closing brace, restored in a logged copy. This historical rule is not exact instruction compliance: extra keys, copied schema annotations and some range errors can pass. Bare-JSON and fence-only exact-contract checks are reported separately in Table A8.

Agreement. Agreement is measured as Krippendorff's α pooled over fields using the script employed in the prior study, which reproduces that study's ordinal .725 and nominal .706. The headline nominal pool in this run contains 32 observed categorical keys, including 26 canonical nominal fields. Each field–value combination is a distinct category in the pooled chance calculation. The historical helper treats 8/9 as missing for every field, although assessment_object_code specifies 88/99. Table A6 gives the field list and sensitivities using canonical keys, documented ranges and field-specific missing values, plus an alternative within-field chance calculation. Those choices change the coefficient's magnitude. The ordinal 0–5 pool is separate, and the nominal pool is not a coefficient over all 91 codes. We present historical strict nominal α first, followed by ordinal α, lenient values and document counts in the appendix.

Ceilings. A model's observed test-retest agreement is used as a reference, called its "ceiling" in the original run tables. It is not a theoretical maximum. To compare pairs whose members differ in self-consistency, we divide a pair's α by the geometric mean of its members' retest coefficients:

$$\textit{retest-relative agreement}(A, B) = \alpha(A, B) \div \sqrt{\left(\alpha(A, A') \times \alpha(B, B') \right)}$$

This has the algebraic form of the classical correction for attenuation of correlations (Spearman, 1904). We use it descriptively; it is not a validated attenuation correction for Krippendorff's α, a noise-free agreement estimate or an effective coder count. A value near 1 means pair agreement is near the geometric mean of the two observed retests. "Normalised" below refers to this descriptive index.

Intervals. The 95% intervals come from 4,000 bootstrap resamples of the 65 documents. Each resample recomputes retests and pairs together, retaining their shared-document dependence. Route intervals resample the 24 or 25 matched documents for each comparison. Group means weight model pairs equally; Anthropic supplies three of the seven capable same-family pairs. These are exploratory, pointwise intervals without multiplicity or selection adjustment. They reflect document sampling, not which models were chosen, and most retests rest on two runs.

4. Full House: Reliability Results

Figure 1 shows the agreement between every pair of models. Same-vendor blocks are visible along the diagonal, most clearly for Anthropic and OpenAI. The rest of this section asks how much of that pattern survives once each model's agreement with itself is taken into account.

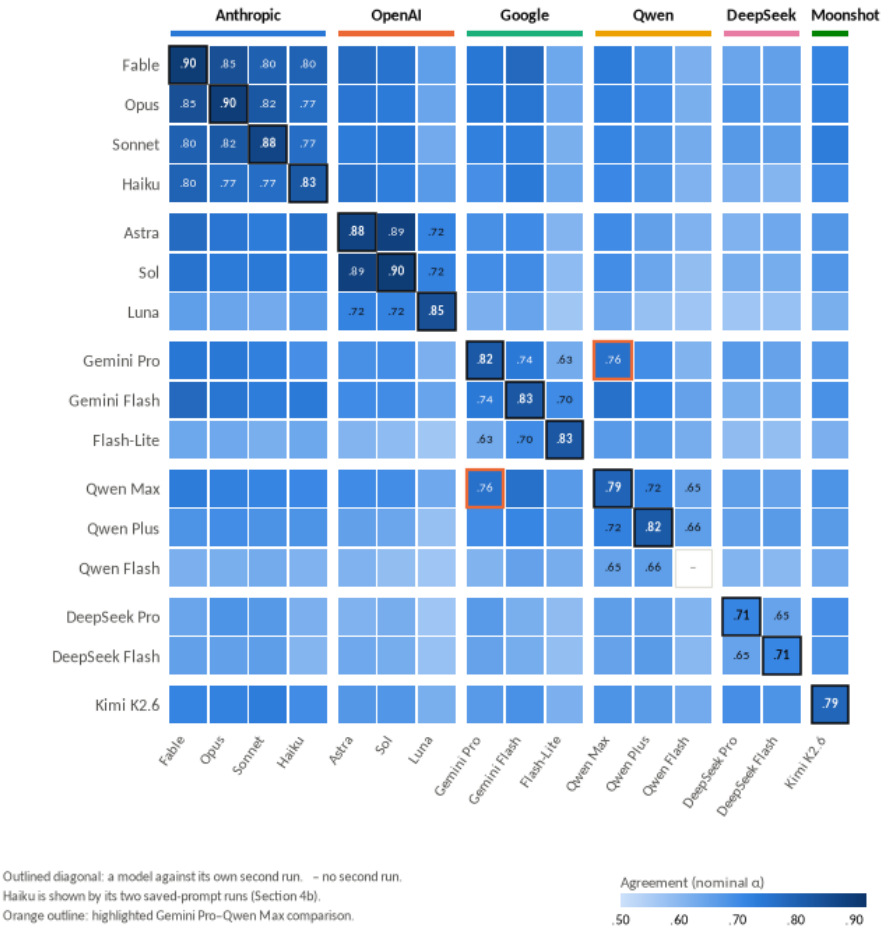


Figure 1. Agreement between every pair of models. Nominal α, strict scoring, first run of each model (for Haiku, its first saved-prompt run), grouped by vendor. Outlined diagonal cells show a model against its own second run.

4a. Groundhog Day: Test–Retest Agreement

The ceiling scores show no model exactly reproduces itself. Figure 2 shows that Fable had the highest observed retest agreement of .903, followed by Sol at .902 and Opus at .898, while the two DeepSeek models had the lowest at .715 and .713. Observed retests varied by vendor without a monotone tier ordering. OpenAI and Anthropic were able to achieve ceilings in the .88-.90 range for their capable models, the three Gemini models achieved ceilings in the .82-.83 range, and Qwen and Kimi achieved ceilings in the .79-.82 range. Within Google, the smallest model (Flash-Lite, .827) was as consistent as the largest (Pro, .824). Where a model was retested more than once, the values were close (Sol: .902, .884, and .878 between its two repeats). A pairwise α therefore means

different things in relation to different retests: .80 is farther below retests near .90 than retests near .80. This comparison does not separate systematic disagreement from sampling variation.

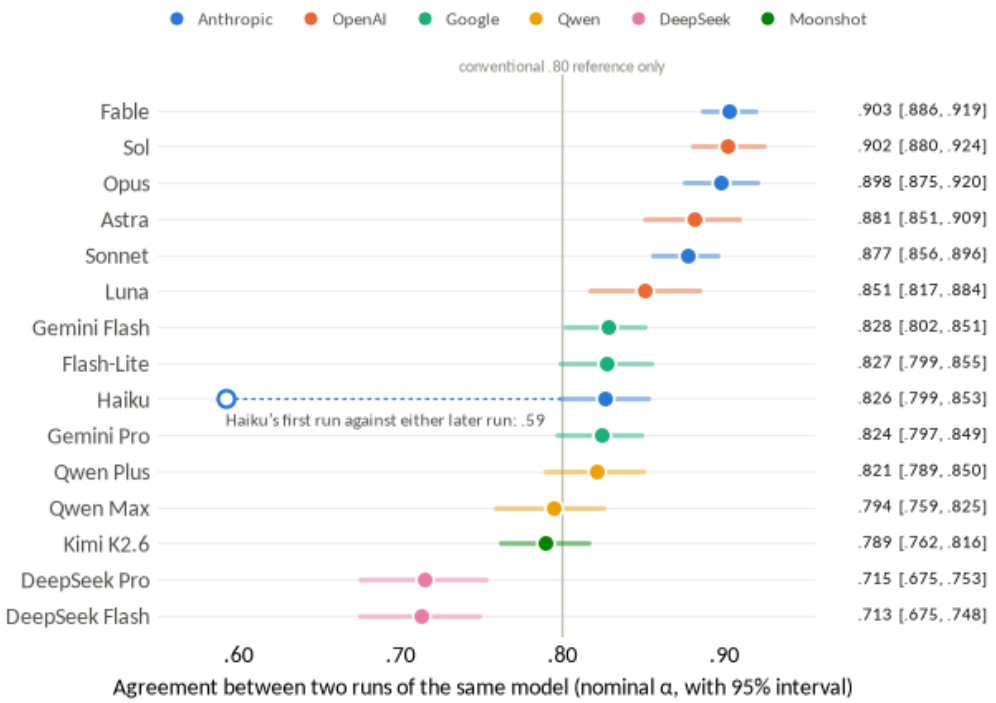


Figure 2. How well each model agrees with itself. Nominal α between two runs of the same model on the same documents, with 95% bootstrap intervals. The hollow marker is Haiku's first run against its later runs (Section 4b).

4b. Twins (1988): Same Model, Different Answers—The Haiku Retest

Haiku was run three times. The first run was made before prompts were saved and returned 50 valid outputs of 65. Two later runs used a saved prompt, one in the cloud and one on the local machine, and each returned 65 of 65. The two later runs agree with each other at .826 [.799, .853]. The first run agrees with them at only .592 and .595, based on the 50 documents that validated in all three runs (Figure 3).

The first run is also the odd one out against the models compared in Figure 3. On the same 50 documents it agrees with Opus at .595, with Sonnet at .602 and with Fable at .576, while the two later runs agree with those models at .76 to .81. The extra disagreement is not confined to one part of the instrument: cost-benefit fields carry 69% of the first run's nominal disagreements and 71% between the later runs, while supplying about 53% of the paired nominal units. Agreement was .615, .593 and .570 for the shortest, middle and longest packet thirds; these descriptive comparisons do not rule out a length effect. Prompt, route and incompletely recorded settings are entangled. Because the first prompt was not saved, these runs cannot identify the cause of the difference.

Two things follow. First, we take Haiku's two saved-prompt runs as its arm and its ceiling, and we report the first run separately. Had we used the first run, Haiku would have looked like a model that cannot reproduce its own codes; the saved-prompt pair has retest agreement comparable to the

Gemini and Qwen models, although its routes differ. Second, a low retest value cannot be read as instability in the model unless the two runs are known to have been made the same way. The first Opus and Sonnet runs were launched the same way as the first Haiku run. Their observed retests (.898 and .877) are high, but also compare an unsaved-prompt run with a saved-prompt run; they do not establish that launch differences had no effect.

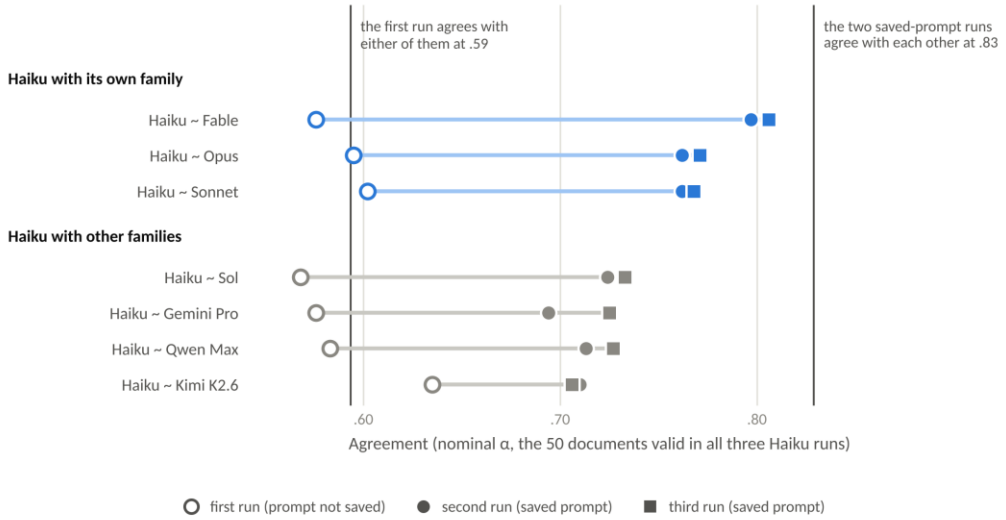


Figure 3. Three runs of Haiku. Agreement of each Haiku run with other models, on the 50 documents valid in all three runs. Vertical lines mark the agreement between Haiku's own runs.

4c. Hum Saath-Saath Hain: Within-Family Agreement

Astra and Sol agree with each other at .890, and their own ceilings are .881 and .902. Normalised, the pair is at 1.00 (Figure 4): on this index, pair agreement is near the geometric mean of their observed retests. For the capable models of each family the normalised values are high throughout: 1.00 for OpenAI, .90 to .94 among Fable, Opus and Sonnet, .90 for Gemini, .90 for Qwen and .91 for DeepSeek. The seven capable same-family pairs average .926 [.911, .940].

Pairs that include a family's smallest model differ by vendor. Haiku sits almost as close to its three larger siblings (.90 to .93, mean .909) as they sit to each other (.926). Luna sits further from Astra and Sol (.83 and .82, against 1.00 between those two), and Flash-Lite further from Gemini Flash and Pro (.85 and .77, against .90). Even so, each small model is closer to its own family on average. Its normalised agreement with its capable siblings exceeds its agreement with capable models of other families by .059 [.038, .078] for Haiku, .083 [.054, .111] for Luna and .053 [.026, .079] for Flash-Lite.

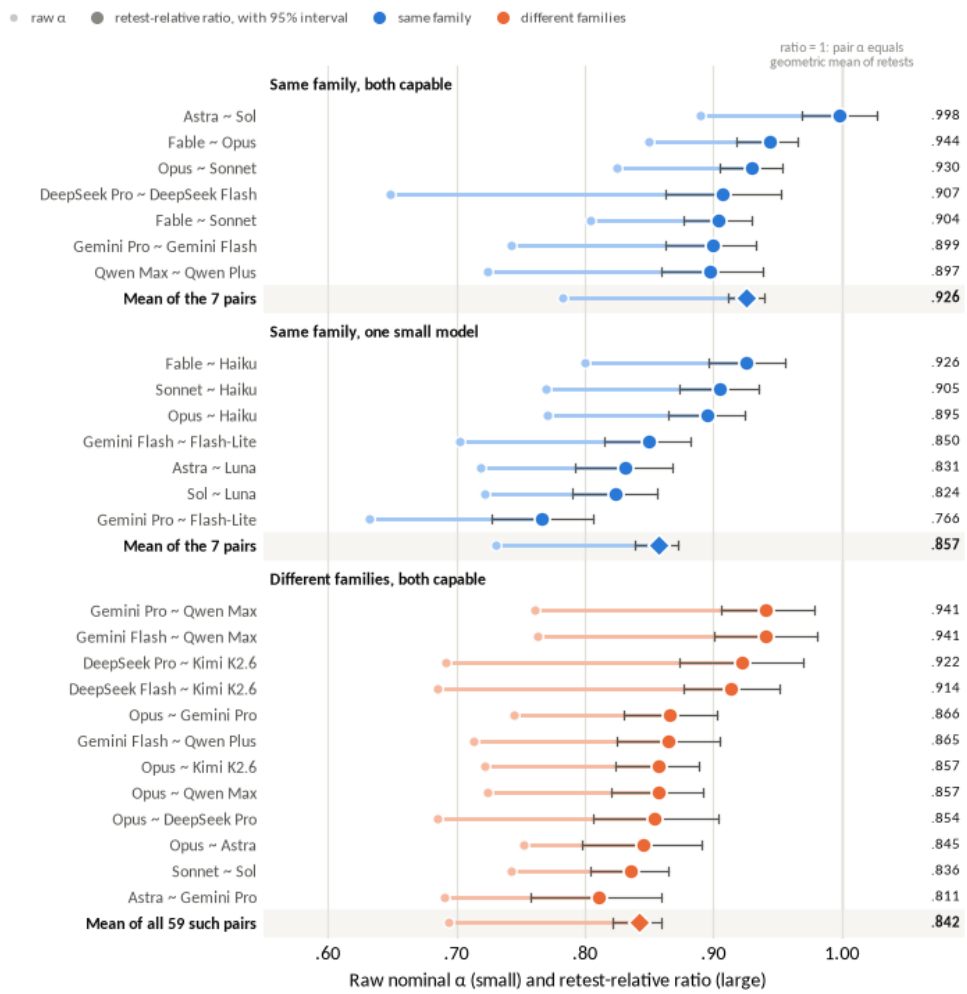


Figure 4. Raw and retest-relative agreement. Small markers: historical pooled nominal α. Large markers: dimensionless ratios dividing that value by the geometric mean of the two observed retest coefficients, with 95% pointwise intervals. Diamonds are unweighted means over model pairs.

4d. Reply 1988: Variation in Family Effects Across Vendors

Raw same-vendor agreement varies widely across vendors. Among the capable tiers it ranges from .890 for OpenAI (Astra ~ Sol) and .825 to .850 for Anthropic (Opus ~ Sonnet, Fable ~ Opus) down to .743 for Google (Gemini Pro ~ Flash), .724 for Qwen (Max ~ Plus) and .648 for DeepSeek (Pro ~ Flash). Read alone, these figures suggest a strong family effect for OpenAI and Anthropic and a weak one for Google, Qwen and DeepSeek.

The retest-relative index changes that reading. The vendors with lower sibling agreement are also the vendors whose models agree less with their own repeats. The OpenAI and Anthropic retests lie between .877 and .903, while Gemini's lie at .824 and .828, Qwen's at .794 and .821, and DeepSeek's at .715 and .713. Once each pair is divided by the geometric mean of its two retests, the range across the five vendors narrows from .648–.890 to .90–1.00: Gemini, Qwen and DeepSeek siblings sit at .90 to .91 of their retest level, within Anthropic's range of .90 to .94. On this descriptive index, vendors differ much more in how consistently their models reproduce their own codes than in how

closely their models resemble one another. This does not establish that lower self-consistency explains the raw differences completely, because the ratio compounds three estimates and is not a validated attenuation correction.

4e. Parasite: Cross-Family Agreement and Disagreement

Capable models from different vendors agree at .615 to .781 raw and at an average of .842 [.822, .860] of their ceilings (see Figure 4). If capability convergence alone accounts for agreement, a same-vendor advantage might narrow when pairs are compared relative to their retests. This is not what occurs; capable same-vendor pairs average .926, and the difference of .084 [.068, .101] holds across resamples of the documents. The observed advantage is consistent with a family association, but route, protocol and unmeasured capability remain alternative explanations. The separation is in the averages and not in every pair: the 59 capable cross-vendor pairs run from .75 to .94, and five of them sit above the lowest capable same-vendor pair.

Finding candidates. There was an agreement of .761 raw between Gemini Pro and Qwen Max. Raw agreements between Gemini Pro and its sibling were lower at .743. Normalized, Qwen Max agreed with Gemini Pro at .941 [.906, .978] and with Gemini Flash at .941 [.901, .981], as close as siblings. We report this only as an exploratory candidate, with pointwise intervals unadjusted for selecting these pairs: these are two pairs among 59, all three models were run using standard API interfaces, and we cannot explain why models from separate vendors were so similar. DeepSeek Pro and Flash also normalise high with Kimi (.92 and .91), but DeepSeek's ceilings are the lowest in the study and its normalised values are the least certain.

Principle. Our prior study found the narrowest rating gaps on privacy and transparency and wider gaps on both beneficence principles, integrity and justice; the OpenAI coder rated all ten principles higher than Claude. The current analysis of the earlier raw outputs also shows higher agreement on the concrete principles (privacy, transparency) than on the interpretive group (Mechanical Beneficence, Ethical Beneficence, integrity, justice). These patterns resemble this study (Figure 5). In a comparable three-family panel (Opus, Astra, Gemini Pro), the ten ordinal α values correlated with the earlier ranking at Spearman's rho = .90. Concrete-principle agreement averaged .75 in each study. Interpretive-principle agreement increased from .41 to .55; justice rose from .63 to .78 while remaining in the fixed interpretive group. The other four principles are outside this grouped comparison. Table A5's earlier per-principle α values are recomputed from the earlier raw outputs, rather than quoted from the SAMYRAD paper.

This study adds retests and within-family comparisons. Across the 15 retests, the mean for concrete principles was .87 and for interpretive principles .80. Among the ten cross-family pairs formed by Opus, Astra, Gemini Pro, Qwen Max and Kimi, retest-relative agreement averaged .87 for concrete principles and .70 for interpretive principles. The four same-vendor pairs (Opus–Sonnet, Astra–Sol, Gemini Pro–Flash, Qwen Max–Plus) averaged .97 and .92. These are descriptive comparisons, not a formal family-by-principle interaction test. The calibration offsets show a similar pattern. Astra and Sol scored higher than Opus on all ten principles, by means of .49 and .48 scale points, compared with .77 in the earlier study. Astra and Sol's signed difference averaged approximately zero across

principles; that does not mean their individual ratings were identical. Higher intensity is not a measure of accuracy.

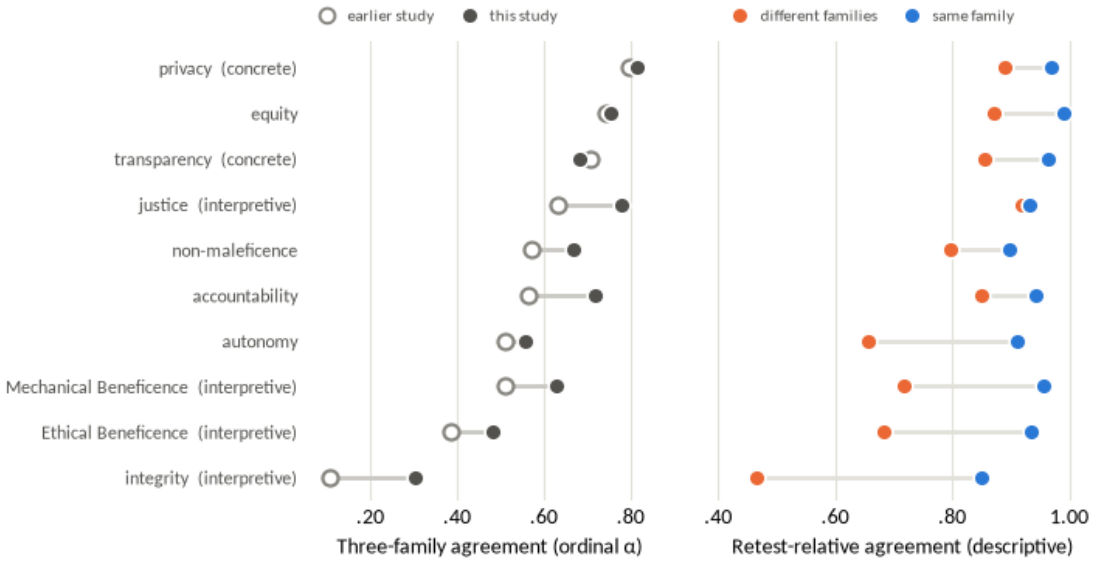


Figure 5. Agreement by ethical principle. Left: ordinal α for a three-family panel in the earlier study (Claude, Codex, Gemini) and this one (Opus, Astra, Gemini Pro). Right: descriptive retest-relative agreement for the ten selected cross-family and four same-family pairs. Group membership is fixed; Table A5 gives the values.

The interface test. Because family and route are partly confounded, we ran Sonnet and Sol once more through OpenRouter, host pinned to the vendor, on a fixed 25-document sample. The sample was drawn within sector before these API calls, from 62 documents historically valid in four original/retest arms, excluding SRC_000024. The pre-run reading rule treated a difference within .04 of the retest reference as small and a deficit exceeding .10 as a signal for investigation; it was not an equivalence test. Table 2 gives the results. Sonnet through the API agrees with its sub-agent runs at .884 and .872, against a retest reference of .911 on the same documents. Sol through the API agrees with its Codex-app runs at .873 to .890, against references of .893–.894. All five point differences are within .04, but one Sonnet interval extends to −.081 (point difference −.038 [−.081, +.000]). Sonnet and Sol agree at .768 through the API and .756 through their own routes, a difference of +.012 [−.026, +.051]. The point differences are small, but the intervals do not establish equivalence or rule out a route effect. The wrapper did change: both models fenced all 25 API replies, which neither did through its vendor's tool. Sonnet's inherited thinking setting was not recorded, so this is not a fully matched settings comparison.

Table 2. The interface test: one model through two routes, 25 documents

Comparison	Docs	Nominal α	Reference on the same documents	Reference α	Difference [95% interval]
Sonnet (API) ~ Sonnet (sub-agent, first run)	25	.884	Sonnet first run ~ retest	.911	−.027 [−.059, +.005]
Sonnet (API) ~ Sonnet (sub-agent, retest)	25	.872	Sonnet first run ~ retest	.911	−.038 [−.081, +.000]
Sol (API) ~ Sol (Codex app, first run)	25	.881	Sol first run ~ retest 1	.893	−.012 [−.043, +.020]
Sol (API) ~ Sol (Codex app, retest 1)	25	.890	Sol first run ~ retest 1	.893	−.003 [−.040, +.032]
Sol (API) ~ Sol (Codex app, retest 2)	24	.873	Sol first run ~ retest 1	.894	−.022 [−.058, +.016]
Sonnet (API) ~ Sol (API)	25	.768	Sonnet ~ Sol, each through its own route	.756	+.012 [−.026, +.051]

Note. Historical strict scoring, including logged fence removal. Differences carry 95% bootstrap intervals over the documents. API = OpenRouter with the host pinned to the vendor; Sol at medium reasoning effort as in the Codex app; no thinking setting was sent for Sonnet.

5. The Rules of the Game: Instruction Adherence

Ten of the 16 models returned an output eligible under the historical count-based rule for all 65 documents on at least one full run, and the six that never did are not simply the small ones (Figure 6). Astra, the largest OpenAI model, returned 51 historically strict-eligible outputs on its first run, fewer than Sol (63) and Luna (58). On 13 of 65 outputs Astra omitted the final closing brace; with that brace restored under lenient scoring it has 64. The failure did not attach to documents: in a full repeat 5 outputs failed the same way, only 1 of them among the original 13.

Small models account for the largest failure counts (Qwen Flash 26 of 65, first-run Haiku 15), but the pattern is not monotone. Haiku went from 15 failures to none in both later runs, and Luna from 7 to none when run at high reasoning effort. Models called through a plain API usually wrapped their output in a code fence that the instrument forbids, and Sonnet and Sol did the same on every reply when they were called that way. Models' accounts of their own work were uneven: some reported a code count that did not match the file, three Sonnet retest replies disclosed that the packet had not been read to the end, and one Fable reply disclosed a revision it had been told not to make. Because invalid outputs drop out of agreement statistics, adherence decides how much evidence each pair rests on: Astra's pairs rest on about 50 documents under strict scoring and 64 under lenient, while the largest strict–lenient α change among the selected 15-model pairs is .016.

One packet (SRC_000024) carried the wrong edition label in its header. Astra's and Sol's original arms halted on the mismatch; the other original main arms coded it. The frozen packet and scoring rules were retained. Setting that packet aside changes no group average of agreement by more than .001.

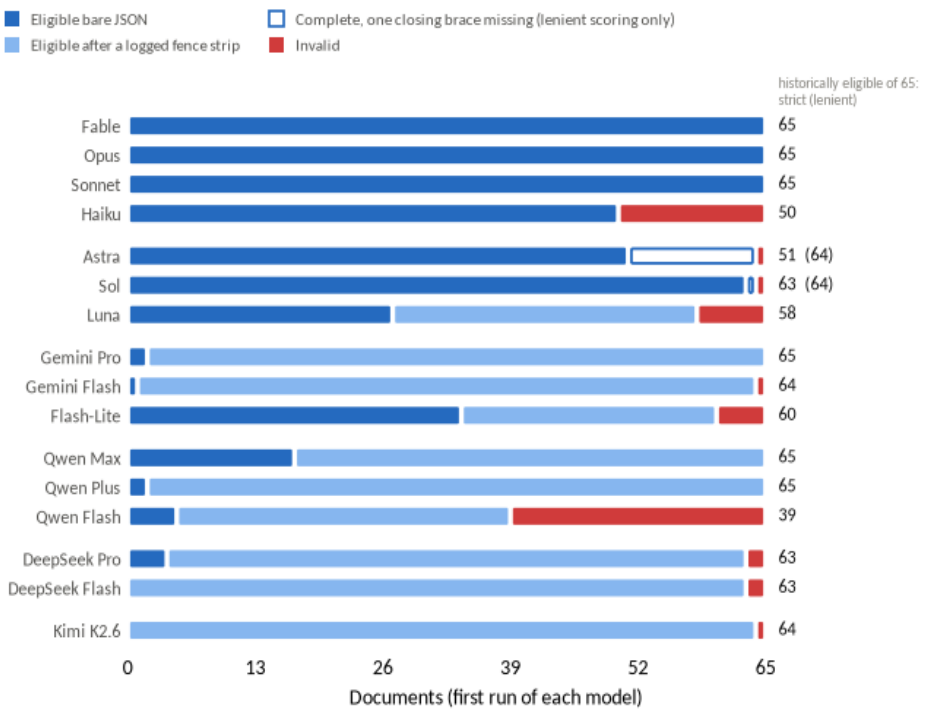


Figure 6. What each model returned. The 65 outputs of each model's first run, classified by historical scoring eligibility. Bare-JSON and fence-stripped replies pass the count-based rule; they need not pass the exact-contract checks in Table A8. Outputs missing one final closing brace count only under lenient scoring.

6. The Importance of Being Earnest: Discussion and Recommendations

Agreement between LLM coders can reflect the document, each coder's repeatability and similarities associated with family. Typically, researchers read a high coefficient as evidence about the coding scheme. Here, Astra and Sol agree at .89 and Opus and Sonnet at .83 under the historical nominal pool, whereas selected top-tier cross-vendor comparisons are around .69–.76. The panel changed, and so did the evaluation. Those examples do not cover every cross-family pair, and the heterogeneous pooled index should not be treated as an instrument-wide pass or fail against .80. Agreement can be studied without a gold standard, as we do here; correctness and error correlation cannot be inferred from these comparisons. The results give reason to avoid assuming independence from separately named models, while also showing that cross-vendor labels alone do not establish it.

The findings by principle sharpen the earlier study's conclusion that when families have disagreements, these disagreements are diagnostic of what is being measured. Families disagree with each other most on the interpretative principles, while the selected within-family pairs disagree less. Therefore, as long as an investigator uses a same-vendor panel, they can get high levels of agreement precisely at the points where the tested cross-family coders part ways, and they may miss the calibration offset between vendors, where the tested siblings share similar offsets.

Our question complements Kohli's (2026): where Kohli asks how many independent votes a panel provides, we ask how much of the shared signal follows family lines. Kohli found only a small same-

family excess in error correlation (+.047) on a three-label inference task, with the most correlated pairs crossing families. Our larger family gap (.926 versus .842 retest-relative) may reflect the interpretive coding task, more same-family pairs (seven versus two) and adjustment for each model's retest agreement. Consistent with Kohli, some cross-family pairs (Gemini Pro ~ Qwen Max, .94) were as close as siblings: family is a strong average tendency, not a guarantee of either independence or dependence.

Two checks qualify these results. The Haiku result shows that one run made under an unrecorded prompt can disagree with later runs of the same model (.59) by more than selected models of different vendors disagree with each other. The interface test gives small point differences for two models, with intervals too wide to establish equivalence. Neither check isolates a single cause.

For studies that use LLMs as coders, we make six suggestions. First, every coder should be run at least twice and its observed retest agreement reported. Second, the exact prompt and settings of every run should be saved, and retests should be run under the same prompt and settings. Third, raw and retest-relative agreement should be reported together, along with the pooling rule used to compute them. Fourth, panels should be built across vendors, and same-vendor models should not be counted as independent replications. Fifth, cross-vendor pairs should also be checked against their observed retests, because a vendor label is not proof of independence. Finally, researchers should keep raw replies, publish a deviation log and report the interface through which each model was run.

7. ’Tis But a Scratch: Limitations

Interface is partly confounded with family in the main design. Claude models ran as Claude Code sub-agents, OpenAI models in the Codex app, Gemini models through Vertex, and Qwen, DeepSeek and Kimi through OpenRouter. The route test covers two models, 25 documents and one direction, from vendor tools to a plain API. Its point differences were small, but its intervals do not establish equivalence. Sonnet’s sub-agent thinking setting is unknown, and exact Claude served snapshots were not universally documented. Three early DeepSeek/Kimi provider histories remain ambiguous; Qwen was unpinned despite observed Alibaba replies. Existing comparisons cannot settle the confound: capable cross-family pairs sharing OpenRouter have higher retest-relative agreement than other capable cross-family pairs (.875 versus .836) but lower raw α (.671 versus .698), in part reflecting their lower observed retests.

The statistical scope is also limited. Intervals cover documents only, conditional on the selected models, arms and historical eligibility. They do not cover model selection or adjust for multiple exploratory comparisons. Most retests rest on two runs. The descriptive normalised index compounds three estimates and can exceed 1; the Astra–Sol interval reaches 1.03. It is not a validated attenuation correction, and the data do not establish a causal lineage mechanism or independent information counts. Nominal magnitudes depend substantially on field selection, missing-value and chance-pooling rules (Table A6).

A model's runs may not be interchangeable, as Haiku's first run shows. The first Opus and Sonnet runs were also launched without a saved prompt, and several Claude retests crossed a local-to-

cloud boundary. In addition, the Haiku arm was chosen after the results were seen. We use its two saved-prompt runs because they share a documented prompt; using the first run instead lowers the average for pairs with a small model from .857 to .816 and leaves the capable-model results unchanged.

Comparisons rest on unequal overlap. Pairs are compared on the documents both members coded validly, which ranges from 39 to 65 documents. On the 22 documents valid in all 30 runs behind the ceilings, the three group averages are .927, .863 and .852, against .926, .857 and .842 on all documents, so the main pattern does not depend on the larger overlaps.

Finally, the design is narrow. Sizes are not matched across generations: "size" means vendor tier here, and parameter counts and equivalent capabilities are not established. The study uses one corpus, one instrument and 65 English-language documents in an exploratory design with arms added as results came in. The corpus is reused from the earlier study, so comparisons between the studies are not independent replications on new documents, and this paper does not assess coding accuracy or national differences. Qwen Flash has no retest and appears only in Figures 1 and 6.

8. The Joy Luck Club: Conclusion

Two models from one vendor agreed at .89 and were as similar to each other as their observed retests. In this study self-consistency ranged from .90 to .71, capable same-vendor pairs averaged .93 relative to their retests, and capable cross-vendor pairs .84. One run made under an unrecorded prompt agreed with later runs of its own model at .59. These results support a family association in agreement, while leaving route, protocol and settings partly entangled. They do not establish accuracy, a causal lineage effect or an effective coder count. We suggest that reports using LLM coders include each coder's retest agreement, saved prompts and settings, explicit scoring rules, and a panel that crosses vendors. The question in the title therefore has a qualified answer. Models from the same family do tend to code alike, and more alike than models from different families, but the size of that family resemblance can only be judged against how consistently each model agrees with itself. Reliability claims built on LLM coders should report that self-agreement alongside every pairwise coefficient.

Data and Results Availability

Coded outputs, including raw model replies, derived coding files and the deviation log, together with the analysis scripts, are available from the author on request. The 65 source documents are publicly available institutional documents, and their identifiers are listed in the paper. Numerical results supporting this paper will also be published on The Ethics Observatory website (https://craigsingleton.org/) when the final report is completed.

Disclosure

Large language models are involved in this paper in two ways, and both are disclosed. As research instruments and objects of study, their recorded configurations, coding outputs, failures and

departures from the protocol are documented in the methods and results. In manuscript production, Claude models (Anthropic) and Codex (OpenAI) provided editorial and computational assistance. The author dictates drafts; AI assisted with correcting transcription errors, drafting and revision, preparing and executing analysis code, statistical checks, source verification, and suggestions for interpreting exploratory comparisons. Claude models launched the Claude coding arms, and OpenAI models accessed through Codex performed the OpenAI coding arms. Anthropic and OpenAI are among the six vendor families studied. The author directed the study, reviewed the manuscript and analyses, and bears sole responsibility for its content, conclusions and any remaining errors.

References

Awuni, D. A., Achenie, L. E. K., Partey, B. T., Owusu, E. G., & Sowah, N.-N. D. (2026). Who judges matters: Measuring family-conditioned preference in LLM-as-judge panels. arXiv. https://arxiv.org/abs/2609.17857

Goel, S., Strüber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., & Geiping, J. (2025). Great models think alike and this undermines AI oversight. Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, 267, 19621–19678. https://proceedings.mlr.press/v267/goel25b.html

Kim, E. M., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, 267, 30038–30066. https://proceedings.mlr.press/v267/kim25e.html

Kohli, G. (2026). Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. arXiv. https://arxiv.org/abs/2605.29800

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE.

Najera, A., Moon, A., Srinivasan, V., & Veeraraghavan, R. (2026). When models disagree: Rethinking LLM evaluation for public comment analysis. arXiv. https://arxiv.org/abs/2605.29025

Singleton, C. (2026). Coders as standpoint instruments: Reading cross-family LLM disagreement as diagnostic in AI-ethics content analysis [Conference manuscript]. SAMYRAD 2026.

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101. https://doi.org/10.2307/1412159

Appendix: values behind the figures

Table A1. Test–retest ceilings (values behind Figure 2)

Model	Docs	Nominal α (strict) [95% interval]	Ordinal α	Exact agreement	Nominal α (lenient; docs)
Claude Fable 5.1	65	.903 [.886, .919]	.947	.849	.903 (65)
GPT-6.1 Sol	63	.902 [.880, .924]	.936	.875	.903 (64)
Claude Opus 5.5	65	.898 [.875, .920]	.957	.873	.898 (65)
GPT-6 Astra	47	.881 [.851, .909]	.939	.881	.881 (64)
Claude Sonnet 5.5	64	.877 [.856, .896]	.946	.847	.877 (64)
GPT-6 Luna	53	.851 [.817, .884]	.911	.804	.851 (53)
Gemini 3.8 Flash	64	.828 [.802, .851]	.908	.791	.828 (64)
Gemini 3.5 Flash-Lite	57	.827 [.799, .855]	.916	.768	.827 (57)
Claude Haiku 4.5 (saved-prompt runs)	65	.826 [.799, .853]	.888	.758	.826 (65)
Gemini 3.1 Pro	65	.824 [.797, .849]	.888	.744	.824 (65)
Qwen 3.7 Plus	65	.821 [.789, .850]	.893	.732	.821 (65)
Qwen 3.7 Max	64	.794 [.759, .825]	.866	.735	.794 (64)
Kimi K2.6	64	.789 [.762, .816]	.884	.710	.789 (64)
DeepSeek v4 Pro	58	.715 [.675, .753]	.844	.642	.715 (58)
DeepSeek v4 Flash	61	.713 [.675, .748]	.843	.636	.713 (61)
Haiku's first run, reported separately					
Haiku first run ~ second run	50	.592	.676	.493	.592 (50)
Haiku first run ~ third run	50	.595	.682	.504	.595 (50)

Note. Ordered by strict nominal α. Astra's ceiling uses its full repeat; Sol's uses its first repeat; Haiku's uses its two saved-prompt runs. Exact agreement is the share of codes with identical substantive values in both runs.

Table A2. Within-family pairs (values behind Figure 4)

Pair	Docs	Nominal α (strict)	Nominal α (lenient; docs)	Ordinal α	Retest-relative [95% interval]
Both capable					
Astra ~ Sol	50	.890	.890 (64)	.931	.998 [.969, 1.027]
Fable ~ Opus	65	.850	.850 (65)	.920	.944 [.918, .966]
Opus ~ Sonnet	65	.825	.825 (65)	.921	.930 [.905, .954]
DeepSeek Pro ~ DeepSeek Flash	61	.648	.648 (61)	.816	.907 [.863, .952]
Fable ~ Sonnet	65	.804	.804 (65)	.893	.904 [.877, .930]
Gemini Pro ~ Gemini Flash	64	.743	.743 (64)	.836	.899 [.863, .933]
Qwen Max ~ Qwen Plus	65	.724	.724 (65)	.838	.897 [.859, .938]
One small model					
Fable ~ Haiku	65	.800	.800 (65)	.871	.926 [.896, .956]
Sonnet ~ Haiku	65	.770	.770 (65)	.847	.905 [.874, .935]
Opus ~ Haiku	65	.771	.771 (65)	.861	.895 [.865, .924]
Gemini Flash ~ Flash-Lite	59	.703	.703 (59)	.796	.850 [.815, .882]
Astra ~ Luna	45	.719	.733 (57)	.793	.831 [.792, .868]
Sol ~ Luna	57	.722	.722 (57)	.800	.824 [.790, .856]
Gemini Pro ~ Flash-Lite	60	.632	.632 (60)	.793	.766 [.728, .806]
No retest for Qwen Flash: raw only					
Qwen Plus ~ Qwen Flash	39	.660	.660 (39)	.774	—
Qwen Max ~ Qwen Flash	39	.651	.651 (39)	.790	—

Note. Retest-relative (historically "ceiling-normalised") = strict nominal α divided by the geometric mean of the two models' strict retest ceilings (Table A1). A value of .998 is described as 1.00 in the text. Haiku = its first saved-prompt run. Normalised values here are computed from unrounded α; the pipeline's ceiling table, computed from rounded α, differs by .001 for five pairs.

Table A3. Cross-family pairs named in the text or shown in Figure 4

Pair	Docs	Nominal α (strict)	Nominal α (lenient; docs)	Ordinal α	Retest-relative [95% interval]
Gemini Pro ~ Qwen Max	65	.761	.761 (65)	.849	.941 [.906, .978]
Gemini Flash ~ Qwen Max	64	.763	.763 (64)	.854	.941 [.901, .981]
DeepSeek Pro ~ Kimi K2.6	62	.692	.692 (62)	.828	.922 [.874, .970]
DeepSeek Flash ~ Kimi K2.6	62	.685	.685 (62)	.825	.914 [.877, .951]
Opus ~ Gemini Pro	65	.745	.745 (65)	.811	.866 [.830, .903]
Gemini Flash ~ Qwen Plus	64	.713	.713 (64)	.813	.865 [.825, .905]
Opus ~ Kimi K2.6	64	.722	.722 (64)	.842	.857 [.824, .889]
Opus ~ Qwen Max	65	.724	.724 (65)	.826	.857 [.821, .892]
Opus ~ DeepSeek Pro	63	.685	.685 (63)	.790	.854 [.806, .904]
Opus ~ Astra	51	.752	.754 (64)	.795	.845 [.798, .891]
Sonnet ~ Sol	63	.743	.742 (64)	.771	.836 [.804, .865]
Astra ~ Gemini Pro	51	.691	.701 (64)	.773	.811 [.758, .859]

Note. Ordered by normalised value. These are 12 of the 59 capable cross-family pairs; Table A4 gives the averages over all of them.

Table A4. Average agreement by kind of pair

Kind of pair	Pairs	Normalised mean [95% interval]	Raw mean α [95% interval]
Same family, both capable	7	.926 [.911, .940]	.783 [.766, .800]
Same family, one small model	7	.857 [.839, .873]	.731 [.712, .750]
Different families, both capable	59	.842 [.822, .860]	.694 [.673, .714]
Different families, one or both small	32	.778 [.758, .797]	.643 [.622, .664]
Haiku with Fable, Opus, Sonnet	3	.909 [.883, .932]	.780 [.754, .805]
Haiku with capable models of other families	9	.850 [.830, .871]	.694 [.672, .716]
Luna with Astra, Sol	2	.827 [.796, .859]	.721 [.688, .753]
Luna with capable models of other families	10	.744 [.712, .774]	.620 [.590, .650]
Flash-Lite with Gemini Pro, Flash	2	.808 [.777, .839]	.668 [.640, .694]
Flash-Lite with capable models of other families	10	.755 [.726, .785]	.624 [.598, .649]
Differences			
Same family minus different families, both capable		+.084 [+.068, +.101]	+.089 [+.076, +.103]
Same family: both capable minus one small model		+.069 [+.048, +.090]	+.052 [+.034, +.071]

Note. The 15 models with a retest. Small = Haiku, Luna and Flash-Lite; capable = the other selected tiers, without a measured accuracy criterion. Means weight model pairs equally, not vendors equally. Intervals are exploratory pointwise bootstrap intervals over 65 documents; all shared pairs and differences are recomputed together in every resample.

Table A5. Agreement and calibration by principle (values behind Figure 5)

Principle	Earlier study, three families	This study, three families	Retest, mean of 15	Cross-family, normalised	Same-vendor, normalised	Astra – Opus (0–5 points)
privacy	.797	.816	.909	.891	.970	+.45
equity	.745	.754	.900	.872	.990	+.47
transparency	.707	.684	.832	.855	.964	+.20
justice	.634	.779	.868	.919	.931	+.45
non-maleficence	.572	.667	.820	.796	.897	+.20
accountability	.564	.720	.872	.850	.944	+.22
autonomy	.512	.556	.799	.656	.912	+.53
beneficence (mechanical)	.510	.630	.851	.718	.955	+.73
beneficence (ethical)	.386	.484	.801	.682	.936	+1.00
integrity	.108	.302	.669	.467	.851	+.69

Note. Ordinal α. The earlier per-principle coefficients were recomputed from the earlier raw outputs. Three families = Claude, Codex, Gemini in the earlier study and Opus, Astra, Gemini Pro here. Retest = mean over 15 selected retests for that principle. Cross-family = mean of the ten pairs among Opus, Astra, Gemini Pro, Qwen Max and Kimi K2.6, each divided by the geometric mean of its members' retests for that principle. Same-vendor uses Opus–Sonnet, Astra–Sol, Gemini Pro–Flash and Qwen Max–Plus. These per-principle indices retain the original analysis's rounding of component α values before ratios. Last column: mean rating difference, Astra minus Opus.

Supplementary scoring and provenance details

Table A6. Nominal scoring sensitivities on historical eligibility

Scoring method	Same-vendor mean raw / ratio	Cross-vendor mean raw / ratio	Astra–Sol raw / ratio
Historical observed-key pool	.783 / .926	.694 / .842	.890 / .998
Canonical keys, documented ranges, field-specific missing values	.772 / .922	.678 / .833	.883 / .998
Same canonical data, within-field chance diagnostic	.616 / .851	.463 / .679	.780 / .996

Note. Selected non-small pairs and retests match the primary comparison. Sensitivity rows are point estimates, without new bootstrap intervals. The within-field diagnostic weights each field's expected disagreement by its paired units before pooling; it answers a different chance-baseline question and is not an automatic replacement for the historical coefficient. The direction of the same-vendor advantage persists. Ethical ordinal profiles are unchanged by the numeric missing-value correction.

Canonical nominal keys (26): assessment_object, cb_balance, cb_benefit_evidence_label, cb_cost_evidence_label, cb_distribution_direction, cb_primary_cost_bearer, cb_primary_material_beneficiary, cb_primary_value_capturer, dominant_value_type, external_link_boundary, gov_binding_status, gov_compliance_mechanism, gov_enforcement_density, gov_fria_present, gov_opensource_carveout, gov_remedy_present, gov_sanction_present, gov_sector_carveout, gov_standards_displacement, journal_level_of_analysis, journal_method_evidence_label, journal_method_type, journal_source_type, learner_posture, overall_dominant_valence, responsibility_concentration.

The historical observed-key pool also includes six noncanonical keys: cb_accessibility_assistive_code, cb_credential_opportunity_code, gov_ethical_carveout_code, gov_ethics_compliance_code, journal_method_type_code_code, surveillance_monitoring_presence_strength_code_code. The main historical method retained integer values under its existing eligibility rule; the canonical sensitivity removes noncanonical keys and values outside the frozen instrument's ranges. For assessment_object_code the sensitivity treats 88/99 as missing and retains substantive 8/9; other fields retain their documented missing codes. Process and deprecated fields are excluded.

Recorded model configurations

Table A7. Recorded model labels, request identifiers and settings

Model	Recorded identifier or selector	Route and recorded settings
Fable	fable; error log: claude-fable-5-1	Claude Code; thinking setting unrecorded
Opus	opus (recorded label: Opus 5.5)	Claude Code; served snapshot/thinking unrecorded
Sonnet	sonnet (recorded label: Sonnet 5.5)	Claude Code; served snapshot/thinking unrecorded
Haiku	haiku (recorded label: Haiku 4.5)	Claude Code; served snapshot/thinking unrecorded
Astra	gpt-6-astra	Codex app; reasoning medium
Sol	gpt-6.1-sol	Codex app; reasoning medium
Luna	gpt-6-luna	Codex app; reasoning medium
Gemini Pro	gemini-3.1-pro-preview	Vertex; maxOutputTokens 65,536
Gemini Flash	gemini-3.8-flash	Vertex; maxOutputTokens 65,536
Flash-Lite	gemini-3.5-flash-lite	Vertex; maxOutputTokens 65,536
Qwen Max	qwen/qwen3.7-max	OpenRouter; unpinned; observed Alibaba; max_tokens 16,000
Qwen Plus	qwen/qwen3.7-plus	OpenRouter; unpinned; observed Alibaba; max_tokens 16,000
Qwen Flash	qwen/qwen3.7-flash	OpenRouter; unpinned; observed Alibaba; max_tokens 16,000
DeepSeek Pro	deepseek/deepseek-v4-pro	OpenRouter; pin requested: Alibaba; max_tokens 64,000
DeepSeek Flash	deepseek/deepseek-v4-flash	OpenRouter; pin requested: Alibaba; max_tokens 64,000
Kimi K2.6	moonshotai/kimi-k2.6	OpenRouter; pin requested: Moonshot AI; max_tokens 64,000
Sonnet API	anthropic/claude-sonnet-5.5	OpenRouter; Anthropic pinned; max_tokens 64,000; no thinking setting sent
Sol API	openai/gpt-6.1-sol	OpenRouter; OpenAI pinned; max_tokens 64,000; medium reasoning

Note. Identifiers are transcribed from the run manifests and launch records; aliases/labels do not independently verify a served snapshot. All full retests retain the corresponding recorded model identifier. Astra's selected pilot repeat covers 20 documents; its main retest covers 65. Sol has two full retests. Haiku has an original unsaved-prompt arm and two saved-prompt arms; the latter are the primary comparison. Fable's original arm mixes cloud/local Claude Code; its retest is local. Luna high effort is a separate full 65-document arm. The API interface arms have 25 replies each. API scripts did not send temperature or top_p; Gemini thinking settings were not explicitly fixed. Proprietary inherited defaults are not recoverable from the available records. Pin requests are distinguished from returned provider attribution; three early histories remain ambiguous.

Output eligibility and exact contract checks

Table A8. Historical and exact-contract eligibility, replies out of 65

Model / selected arm	Historical first	Exact bare first	Exact fence-only first	Exact fence-only retest
Fable	65	51	51	52
Opus	65	65	65	65
Sonnet	65	63	63	63
Haiku, saved-prompt	65	42	42	49
Astra	51	6	6	8
Sol	63	7	7	5
Luna	58	13	21	23
Gemini Pro	65	0	27	29
Gemini Flash	64	1	29	25
Flash-Lite	60	12	15	18
Qwen Max	65	3	4	2
Qwen Plus	65	1	5	5
DeepSeek Pro	63	4	40	45
DeepSeek Flash	63	0	42	38
Kimi K2.6	64	0	52	55
Qwen Flash	39	0	0	No retest

Note. First means the arm selected for the agreement figures; Haiku therefore uses its first saved-prompt run. Exact checks require the frozen 91-code, 16-narrative and 20-meta key sets, flat blocks, no duplicate keys, JSON types, documented numeric ranges and source/version checks. Fence-only permits removal of an outer fence; it does not repair a brace or change substantive values. These stricter checks are supplementary and do not silently replace the historical agreement cohort. Copied schema annotations can fail an exact key check without omitting a substantive code; other flags include actual shortfalls or invalid values. Exact-contract compliance, received replies and historically eligible replies are different counts. The original Haiku arm has 50 historical-eligible replies, 43 exact bare and 43 exact fence-only; all 65 replies are preserved.

Across the completed original program, 2,280 intended replies in 37 arms were received. This paper selects 16 main model configurations and the retests described above; the complete reply inventory is broader than its headline analytical panel. Invalid returned replies were preserved and were not recoded to improve compliance.