The Podesta Corpus as an AI Classification Benchmark: A Methodology Note

Methodology Note

Fookes Software Ltd
Charmey, Switzerland
www.aid4mail.com

Table of Contents

The Podesta Corpus as an AI Classification Benchmark: A Methodology Note

1. Introduction

The WikiLeaks Podesta release is sometimes proposed as a benchmarking corpus for AI-based email classification in eDiscovery and digital-forensics workflows. It is large, public, topically dense, and free of the artifacts that synthetic corpora introduce. Used as background data—a known-unresponsive volume against which a separately constructed responsive set is measured—it works well. Used on its own as a basis for measuring classifier accuracy on subjective themes, it does not.

This note documents a structured experiment that exposed why. Across one month of the corpus (March 2016, 1,083 emails after deduplication and removal of bulk mail), we ran two operations: a multi-session theme-identification pass through Claude Code Opus 4.7 with reconciliation, and a binary prompt classification of one selected theme—Damage Control and Crisis Communications—across seven model instances at single-run settings, six of them under Aid4Mail at temperature 0 and one a separate Claude Code run at default temperature. The seven runs produced widely divergent results, with no defensible procedure for collapsing them into a single ground-truth label set. Forty-five percent of the emails any model flagged were flagged by exactly one model. Models also disagreed in distinctive ways that have implications for ensemble approaches and for how practitioners should read inter-rater agreement on this kind of task.

The remainder of this note documents the experiment, presents the patterns that led to our negative conclusion about Podesta-only benchmarking, and describes the corpus design we now recommend instead—Podesta retained as background unresponsive data, paired with a separately constructed responsive set built around distinct, externally verifiable forensic categories.

2. Background

The corpus. The WikiLeaks Podesta release comprises emails from the personal Gmail account of John Podesta, chair of Hillary Clinton's 2016 presidential campaign, published in October 2016. The full release runs to roughly 50,000 messages spanning multiple years. For this experiment we worked exclusively with the March 2016 slice of personal correspondence, isolated by applying the Aid4Mail filter query `Type:Personal AND NOT Type:Duplicate` to the raw mailbox; this removed duplicates, automated notifications, and bulk mail, leaving 1,083 emails. The slice was chosen as a tractable single-month window with sufficient volume for inter-model agreement analysis but bounded enough to permit manual review where needed.

Why the corpus is attractive as a benchmark candidate. Several properties make Podesta appealing for AI classifier evaluation. It is large enough for statistically meaningful

evaluation, naturalistic in language and structure (no synthetic fingerprints), publicly available and stable (no licensing or distribution constraints), topically dense (campaign decision-making, media handling, donor coordination, internal deliberation), and well-known enough that practitioners can reason about the domain without specialized briefing.

Why it is hard. The same properties that make it attractive also limit its evaluative use. The content is substantively subjective—what counts as "damage control," "candid internal assessment," or "concern about optics" depends on interpretive choices that are not reducible to ground truth without external adjudication, and the corpus does not come with such adjudication. The selection process itself was retrieval-shaped; emails were leaked because they were in one specific custodian's account, not sampled to be representative of any defined population. Many themes of forensic interest (fraud, threats, regulatory violations) are also absent or rare, so the corpus does not exercise the categories that production AI classifiers are typically asked to detect.

The question this experiment addresses. Granting that the corpus is naturalistic and tractable, can it support a reliable benchmark for AI classifier performance on subjective themes by way of inter-model agreement? Or does the absence of external ground truth make any such procedure circular? The experiment was designed to make this question empirically answerable.

3. Methodology

The experiment consisted of two operations on the same corpus, executed independently.

Theme identification. Three sessions of Claude Code Opus 4.7 at xhigh reasoning effort were given the full 1,083-email corpus and asked to identify and count themes that could plausibly support classification. The first two sessions ran independently with no shared context. The third session was given both prior outputs and asked to reconcile them into a single inventory. The reconciled inventory was treated as the working theme set for the subsequent prompt-classification step. Each email was permitted up to four theme labels, so the operation is multi-label rather than winner-takes-all; an email could carry "Damage Control and Crisis Communications" alongside, for example, "Message Discipline and Talking Points" or "Concern About Optics." The reconciled inventory contained 41 themes spanning roughly 75% of the corpus, with 25% remaining "(untagged)"—a figure we return to in the findings.

Among themes that the reconciled inventory flagged with sufficient volume to support classification testing—Damage Control and Crisis Communications, Candid Internal Assessment, Offer of Campaign Service, Coalition Management, Surrogate Coordination—Damage Control was selected for the prompt experiment. It is conceptually clear in name (recognizable to any practitioner), narrow enough to invite sharp disagreement at the borderline, and large enough at 34 reconciled-inventory matches to support analysis without being unwieldy.

Prompt classification. The following prompt was issued to seven model instances:

```
You are an experienced eDiscovery practitioner reviewing political-campaign emails.
Reply 'Responsive' when the email shows campaign participants engaged in damage
control or crisis communications. Reply 'Unresponsive' when the email only
forwards, summarizes, monitors, or comments on outside coverage, public reaction,
or political developments, with no decision about how the campaign will respond.
Reply 'Unresponsive' for proactive messaging, routine campaign planning,
scheduling, logistics, acknowledgments, and brief replies or administrative
messages. Reply 'INCONCLUSIVE' when the email indicates the campaign is reacting to
a specific risk event but the intended response is not identifiable. Otherwise,
reply 'Unresponsive'.
```

The prompt is a single line with no markdown formatting, conforming to Aid4Mail's prompt-format constraints. The seven models were Claude Opus 4.7 (run via Claude Code at xhigh reasoning effort), Gemini 3.1 Flash-Lite (preview), Grok 4.1 Fast, Gemma 4 26B Think, Magistral 24B, Ministral 3 14B, and Mistral Small 3.2 24B. The first three were called via their respective cloud APIs; the latter four were run locally via Ollama at Q4_K_M quantization.

All Aid4Mail runs used temperature 0 for deterministic output. Aid4Mail v6 served as the test harness for the six non-Claude-Code models, issuing the prompt per email, parsing the verdicts, and writing per-email results to a result set that fed the analysis. Claude Opus 4.7 was run separately under Claude Code with the same prompt and the same per-email output specification; this run used Claude Code's default temperature setting (1.0) and is therefore not deterministic. Single-run comparison across the seven models is unaffected by the difference: every model contributed exactly one verdict per email, and the analysis treats those verdicts symmetrically. The non-determinism becomes relevant only at the theme-identification step, which is precisely why that step was run three times with a reconciliation pass; cross-session variation under non-deterministic settings was the operating assumption, not a complication.

Hardware. The local-model runs were executed on a workstation running Windows 11 Pro with an AMD Ryzen 9 9950X3D processor, 192 GB DDR5 memory, and an NVIDIA GeForce RTX 5090 (32 GB VRAM). Cloud-API runs used the providers' default infrastructure.

4. Findings

4.1 Per-model verdict counts

The seven models, each given identical input on a single run (Aid4Mail runs at temperature 0; Claude Code at default temperature 1.0), produced strikingly different verdict distributions across the 1,083-email corpus:

Model	Responsive	Inconclusive	Errors	Unresponsive
Claude Opus 4.7 (xhigh effort)	106	27	0	950
Gemini 3.1 Flash-Lite (preview)	79	6	0	998
Grok 4.1 Fast	118	15	0	950
Gemma 4 26B Think	70	12	3	998

Magistral 24B	81	4	0	998
Ministral 3 14B	262	1	0	820
Mistral Small 3.2 24B	119	9	0	955

The Responsive count alone varies by a factor of 3.7 across models—and by a factor of 1.7 even after excluding Ministral 3 14B, which we will return to. This is the headline disagreement at the level of marginal totals; the per-email consensus pattern below is sharper.

4.2 Union and intersection, decomposed

Pooling Responsive verdicts across the seven model runs produces a union of 326 unique emails. The seven-way intersection—emails on which every model agreed—contains 24. These figures count Responsive flags only and represent the cleanest measure of classifier agreement on the same task.

Adding the theme-analysis output as an eighth signal complicates the picture. The theme inventory contributes 2 emails not flagged Responsive by any prompt model, extending the eight-output Responsive union to 328. The intersection shrinks to 14, since the theme inventory's 34 matches cap the maximum possible eight-way intersection. The full per-email matrix used for this analysis contains 351 rows; the 23 rows beyond the 328 Responsive-union emails received only Inconclusive or Error verdicts and no Responsive flag from any source.

Throughout the rest of this note we report the seven-prompt union (326) and intersection (24) as the primary agreement figures. The eight-output figures (328 / 14) are reported here for completeness but conflate two different operations and should not be treated as the headline numbers.

4.3 Consensus distribution

The most informative cut of the data is how many models agreed on each flagged email. Among the 326 emails in the prompt union:

Models flagging Responsive	Email count	% of prompt union
Exactly 1	147	45.1%
Exactly 2	55	16.9%
Exactly 3	40	12.3%
Exactly 4	28	8.6%
Exactly 5	14	4.3%
Exactly 6	18	5.5%
Exactly 7	24	7.4%

Forty-five percent of the prompt union represents one model's private opinion. Only 7.4%—the 24-email intersection—has unanimous agreement. The distribution does not show noise around a stable signal but rather a fan of independent signals, with a small consensus core

and a long tail of solitary verdicts. This shape is the central empirical finding of the experiment, and the patterns below clarify what is producing it.

4.4 Pairwise Cohen's kappa

Marginal totals and consensus distributions show that models disagree but not how. Pairwise Cohen's kappa, computed over the full 1,083-email corpus (with emails absent from the working matrix treated as the unanimous Unresponsive verdict every model gave them), reveals three distinct patterns:

	Claude	Gemini	Grok	Gemma	Magistral	Ministral	Mistral-S
Claude	—	0.36	0.33	0.42	0.34	0.26	0.36
Gemini	0.36	—	0.69	0.65	0.51	0.34	0.54
Grok	0.33	0.69	—	0.53	0.45	0.42	0.51
Gemma	0.42	0.65	0.53	—	0.51	0.30	0.53
Magistral	0.34	0.51	0.45	0.51	—	0.39	0.67
Ministral	0.26	0.34	0.42	0.30	0.39	—	0.54
Mistral-S	0.36	0.54	0.51	0.53	0.67	0.54	—

Three patterns stand out.

First, **Ministral 3 14B is a distributional outlier.** Its average pairwise kappa (0.38) is the lowest in the matrix. Of its 262 Responsive verdicts, 93—more than a third—are unique to Ministral, with no other model agreeing. Removing Ministral from the union shrinks it from 326 to 233, a 28% reduction driven by one model. Whether this reflects a genuinely different judgment or a near-default Responsive bias is not adjudicable from this data, but the shape of the disagreement is more consistent with weak discrimination on the prompt than with a refined alternative interpretation.

Second, **Claude Opus 4.7 has the lowest agreement among the non-Ministral models.** Its average pairwise kappa (0.36) is the lowest in the matrix excluding Ministral comparisons. Two interpretations are compatible with the data: Claude is making more refined judgments others miss, or Claude has its own idiosyncratic interpretation of the prompt. Without ground truth, the data does not adjudicate between them. Whichever holds, the consequence for ensemble approaches is the same; a majority-vote ensemble would systematically discount Claude's judgments.

Third, **a Mistral-family signal is present but does not include Ministral.** Magistral 24B and Mistral Small 3.2 24B have a pairwise kappa of 0.67, the highest in the matrix. Family resemblance between same-vendor models with shared training lineage is well-documented in the literature; what is notable here is that Ministral 3 14B does not share it. Its kappa with Magistral is 0.39, and with Mistral Small 3.2 24B is 0.54—both lower than the Magistral × Mistral-Small pair, the former materially so.

4.4.1 Per-model uniqueness as a secondary signal

A cleaner expression of the same patterns can be read directly from per-model uniqueness—the number of Responsive flags from each model that no other model corroborated:

Model	Responsive	Unique flags	Unique %
Magistral 24B	81	0	0.0%
Mistral Small 3.2 24B	119	1	0.8%
Gemini 3.1 Flash-Lite	79	3	3.8%
Gemma 4 26B Think	70	4	5.7%
Grok 4.1 Fast	118	13	11.0%
Claude Opus 4.7	106	33	31.1%
Ministral 3 14B	262	93	35.5%

Magistral 24B is the inverse case to Ministral 3 14B. Every Responsive verdict it issued was corroborated by at least one other model. This is consistent with its high pairwise kappa with Mistral Small 3.2 24B and its tight clustering with the rest of the non-Ministral group; Magistral is making the conservative central call on this prompt. Claude Opus 4.7's 31.1% uniqueness rate is the second-highest in the matrix and, combined with the kappa pattern above, is the empirical signature of either a careful judge whose calls are too refined for others to reach or a divergent judge. Both readings remain on the table.

4.5 Inconclusive usage as a secondary signal

The prompt explicitly invites INCONCLUSIVE for cases where a campaign reaction is visible but the intended response is not identifiable. Models varied sharply in how often they invoked it: Claude Opus 4.7 used it 27 times, Grok 4.1 Fast 15, Gemma 4 26B Think 12, Mistral Small 3.2 24B 9, Gemini 3.1 Flash-Lite 6, Magistral 24B 4, and Ministral 3 14B 1. Two readings are consistent with the spread. Models that almost never use INCONCLUSIVE may be either encountering no genuine ambiguity in the corpus or ignoring the instruction. Independent of which holds for which model, INCONCLUSIVE rates function as a prompt-engagement signal that is informative regardless of accuracy.

4.6 Theme analysis vs prompt classification: two operations, not self-disagreement

The 34-email theme inventory and the 106-email Claude Opus 4.7 prompt-classification result answer different questions. Theme analysis was a proactive labeling operation: the model surveyed each email and applied any themes it judged salient, up to four labels per email. Among the 34 Damage-Control matches, 27 carry one or more additional themes—most often Message Discipline and Talking Points (12 emails), Concern About Optics (8), and Speech Drafting and Review (5). Prompt classification was a binary decision under a specific definition, with explicit decision rules for what should count as Responsive, Unresponsive, or Inconclusive.

The two operations differ in three ways: the level of abstraction at which the model is engaging, the inclusiveness of the criterion (the prompt's "engaged in damage control" catches reactive content that the proactive theme tagger may not surface as a primary or even secondary label), and the presence of an explicit Inconclusive option. Of the 34 theme matches, 32 were flagged Responsive by at least one prompt model, and 19 were flagged by Claude Opus 4.7's own prompt run. The theme inventory is a high-precision conservative subset of the broader prompt union. The 19/34 overlap should not be read as Claude disagreeing with itself; it reflects two genuinely different operations producing two reasonable but non-overlapping answers.

4.7 Untagged rate from the reconciled theme analysis

A finding from the theme-identification operation, separate from the prompt experiment, supports the interpretation that subjectivity is intrinsic to this corpus rather than an artifact of any particular classifier. After three sessions of theme identification—two independent and one reconciled—the reconciled inventory left 274 emails, 25.30% of the corpus, classified as "(untagged)." Roughly a quarter of campaign emails resisted confident thematic assignment even after multi-session reconciliation by a strong general-purpose model with up to four labels available per email. This is consistent with the prompt-classification disagreement and is independent evidence that the corpus is not a clean signal for thematic operations.

5. Interpretation

The data supports three propositions.

Ground truth cannot be reliably constructed from this corpus by inter-model aggregation on subjective themes. With 45% of the prompt union representing single-model opinions and only 24 emails receiving unanimous Responsive verdicts, no aggregation procedure produces a stable label set. Manual adjudication of the unanimous-24 set was attempted as a checkpoint during the experiment and surfaced significant residual subjectivity even within that core; emails on which every classifier agreed could nonetheless be argued to belong on either side of the line by a skilled reviewer applying the same prompt. This is the central empirical reason Podesta on its own does not support a defensible accuracy benchmark for subjective themes. It is not that the models are unusable; it is that there is no stable target against which to measure them.

Majority-vote ensembles inherit noise and discount the strongest individual judgments. The combination of Ministral 3 14B's 93 unique Responsive flags and Claude Opus 4.7's 33 unique flags presents a structural problem for naïve consensus methods. A majority-vote ensemble across these seven models would carry forward many of Ministral's uncorroborated flags through votes from one or two other models that happen to overlap, and would systematically discount Claude's distinctive judgments—the model with the lowest pairwise agreement and the second-highest uniqueness rate. If Claude's distinctiveness reflects refined judgment rather than idiosyncrasy (a question this experiment cannot resolve, but other benchmarks in our suite suggest), the ensemble would invert the apparent ranking. The broader point is that consensus methods are not a free accuracy improvement when underlying judges disagree systematically rather than randomly. Whether to use them depends on whether the disagreement is symmetric noise (which averages out) or structured divergence (which compounds).

Inconclusive rates and per-model uniqueness are useful classifier diagnostics independent of accuracy. Even without ground truth, prompt-engagement behavior—whether a model invokes INCONCLUSIVE when invited to—and idiosyncrasy—whether a model's positive flags are corroborated—reveal stable properties of how the model engages with the task. Practitioners selecting a model for production deployment can read these diagnostics without solving the harder problem of measuring accuracy. A model that ignores prompt instructions on a small task will likely ignore them on larger ones; a model that issues many uncorroborated flags on a borderline task is likely to do the same in production review.

A note on keyword search and TAR. A direct comparison was not part of this experiment. The pattern observed—wide variance in classifier judgment on a subjectively themed task at deterministic settings—is consistent with the known limitations of those methods on subjective content, but we do not make a stronger claim than that from this data alone.

6. Recommended corpus design

The corrective design is straightforward and follows directly from the findings.

The Podesta corpus should be retained as **background unresponsive data**—large-volume, naturalistic, topically dense, and with no externally verifiable responsive content under the categories of forensic interest. Its volume gives a benchmark realistic prevalence dynamics, and its topical density tests classifiers against the kind of content that does not match a target category but might be confused with one.

The responsive set should be constructed **separately**, around distinct, non-overlapping forensic categories: illicit financial activity, corruption and bribery, threats, harassment, phishing and social engineering, data exfiltration, drug trafficking, missing-person communications, and so on. These categories share three properties that the Podesta themes do not. Ground truth is externally verifiable; an email either does or does not document drug-trafficking activity in a way that can be confirmed against external context. Category boundaries can be controlled at corpus-design time; the corpus designer decides what counts as IFA before the test is run, not afterward by adjudication. And the categories exercise the kind of work production AI classifiers are typically deployed to do.

The tradeoff should be stated explicitly. This design measures classifier accuracy on **clearly distinguishable content**. It does not measure performance on nuanced borderline judgment—the very thing the Podesta experiment exposed. That tradeoff is acceptable for a reproducible accuracy benchmark, but readers should not infer that high scores on a clean-categories benchmark imply equivalent performance on subjective forensic review. They are different problems, and the cleanly verifiable benchmark is the easier one.

This is the design adopted for Test 1 in our broader benchmark series, documented separately. The combined approach—Podesta as background, distinct-categories as responsive—produces a benchmark with a defensible ground truth, a realistic 6% prevalence rate, and unambiguous F1, recall, and precision figures for cross-model comparison. Practitioners considering a similar program for their own evaluation work should expect to invest most of the corpus-design effort in the responsive set rather than in the background; Podesta, or a similarly large public corpus, can serve as the unresponsive volume essentially as-is.

7. Reproducibility notes

This experiment can be replicated with modest infrastructure. The full prompt is reproduced verbatim in §3. The corpus slice is the March 2016 segment of the public WikiLeaks Podesta release (1,083 emails). Aid4Mail's prompt-format constraints—single-line text, no markdown formatting—must be observed when porting the prompt to other harnesses, since multi-line prompts and inline markdown are incompatible with parts of the AI configuration pipeline.

For local-model runs, models were sourced from Ollama at Q4_K_M quantization—a mixed-precision format that performs well on classification tasks at a significantly reduced VRAM footprint relative to full-precision builds. The hardware baseline (Ryzen 9 9950X3D, 192 GB DDR5, RTX 5090 32 GB VRAM) is sized comfortably above what is needed for the largest model in the test (Mistral Small 3.2 24B at Q4_K_M); a 24 GB VRAM card with adequate context-length budget is sufficient for any of the offline models in this experiment.

Cloud-API runs used the providers' standard endpoints. Aid4Mail-driven runs used temperature 0; the Claude Code run used Claude Code's default temperature (1.0 at the time of testing). Aid4Mail results are reproducible up to provider-side stochasticity in cloud models, which is non-zero even at temperature 0 but is small enough not to affect the qualitative findings. The Claude Code run would be expected to produce somewhat different verdict counts on rerun, in line with default-temperature behavior; this is the same property that motivated running the theme-identification step three times with a reconciliation pass.

8. Limitations

A small number of caveats apply. The experiment exercised one corpus slice (March 2016) and one theme (Damage Control and Crisis Communications). The patterns observed—Ministral 3 14B isolation, Claude Opus 4.7 as low-agreement, Mistral-family clustering—held on this configuration but have not been retested across other slices or themes. Cross-theme replication would strengthen the finding that single-model uniqueness is intrinsic to subjective themes rather than specific to this prompt.

Each model was queried once per email. Multi-pass evaluation, chain-of-thought elaboration before classification, and self-consistency methods were not tested. Some of the disagreement might compress under richer evaluation protocols, though the structural patterns are unlikely to disappear.

Inter-rater agreement was computed without external ground truth. This is the constraint that motivated the experiment, and the conclusion—that ground truth cannot be constructed reliably from this corpus on subjective themes—is itself the answer rather than a workaround.

9. Appendices

Appendix A: Reconciled theme inventory (March 2016 slice, 1,083 emails)

Theme	Count	%
Scheduling and Logistics	134	12.37%
Debate Preparation	89	8.22%
Message Discipline and Talking Points	85	7.85%
Campaign Strategy Deliberation	63	5.82%
Coalition Management	57	5.26%
Media Strategy and Press Planning	54	4.99%
Offer of Campaign Service	52	4.80%
Personal and Family Correspondence	46	4.25%
Forwarded News and FYI	46	4.25%
Journalist Interactions	36	3.32%
Personnel Hiring and Staffing	36	3.32%
Surrogate Coordination	36	3.32%
Opposition Research	35	3.23%
Damage Control and Crisis Communications	34	3.14%
Donor Cultivation and Stewardship	33	3.05%
Candid Internal Assessment	33	3.05%
Endorsement Strategy	30	2.77%
Polling and Voter Analytics	29	2.68%
Allied Organisation Coordination	28	2.59%
Climate and Energy Policy Discussion	26	2.40%
Concern About Optics	25	2.31%
Primary and Caucus Tactics	24	2.22%
Op-Ed and Earned Media Strategy	20	1.85%
Fundraising Event Planning	17	1.57%
Speech Drafting and Review	17	1.57%
Bundler and Finance Network Coordination	16	1.48%
Vetting and Background Review	14	1.29%

Economic Policy Discussion	14	1.29%
White House and Administration Coordination	13	1.20%
Newsletter and Automated Mail	13	1.20%
Lobbyist or Industry Engagement	11	1.02%
Congressional Liaison	11	1.02%
Confidential or Sensitive Information Sharing	10	0.92%
Foreign Policy Discussion	7	0.65%
Social Invitations	6	0.55%
Faction or Interpersonal Tension	5	0.46%
Super PAC Coordination Concerns	2	0.18%
Phishing and Social Engineering	2	0.18%
Email and Records Practices Discussion	2	0.18%
Healthcare Policy Discussion	1	0.09%
Legal Hold or Litigation Awareness	1	0.09%
(untagged)	274	25.30%

Appendix B: Per-email matrix summary statistics

The per-email matrix used for this analysis contains 351 rows, comprising every email that received a non-Unresponsive verdict from any of the eight outputs (seven prompt runs plus the theme-analysis run). Composition:

Subset	Count
Seven-prompt Responsive union	326
Theme-only emails (theme-flagged but no prompt-Responsive)	2
Eight-output Responsive union	328
Inconclusive-only rows (no Responsive, no theme)	22
Error-only rows (no Responsive, no Inconclusive, no theme)	1
Total per-email matrix rows	351
Seven-prompt unanimous Responsive intersection	24
Eight-output unanimous Responsive intersection (incl. theme)	14

Appendix C: Pairwise Cohen's kappa on the filtered subset

For readers interested in the alternative view of inter-rater agreement restricted to the contested 351-email subset (the union of all non-Unresponsive verdicts), we reproduce the kappa matrix on that basis. This is not the standard inter-rater agreement metric and should not be used as a primary measure, but it gives a sharper view of how models disagree on cases where any model thought a Responsive flag was warranted.

	Claude	Gemini	Grok	Gemma	Magistral	Ministral	Mistral-S
Claude	—	0.21	0.12	0.30	0.18	−0.13	0.16
Gemini	0.21	—	0.62	0.58	0.42	0.11	0.42
Grok	0.12	0.62	—	0.42	0.32	0.09	0.34
Gemma	0.30	0.58	0.42	—	0.42	0.09	0.43
Magistral	0.18	0.42	0.32	0.42	—	0.17	0.59
Ministral	−0.13	0.11	0.09	0.09	0.17	—	0.27
Mistral-S	0.16	0.42	0.34	0.43	0.59	0.27	—

Qualitative patterns are preserved on this basis—Ministral 3 14B remains the lowest-agreement model overall, Claude Opus 4.7 the lowest among non-Ministral models, Magistral × Mistral Small the highest pair—but absolute values are markedly lower than on the full-corpus basis, and Ministral × Claude turns slightly negative. The negative kappa is an artifact of computing the statistic on the filtered union rather than on the full set every model rated; on the full corpus the same pair has a positive (though still low) kappa of 0.26.
