Supplementary Information

A Multi-Model, Multi-Domain Benchmark of Large Language Model Agreement with Humans and with Each Other in Sentiment Classification

Aneesh K Sajan — Independent Researcher, Seattle, WA, USA — asajannow@gmail.com — ORCID 0009-0002-5704-957X

Supplementary Tables

Supplementary Table S1. Dataset selection evidence

Application of the five inclusion criteria (IC1–IC5) to candidate datasets; the five included datasets pass all criteria.

Dataset	IC1 Annotation	IC2 Public	IC3 Size	IC4 Labels	IC5 Citations	Decision
Financial PhraseBank [12]	✓ 16 experts	✓	✓ 4,846	✓ 3-class	✓	**Include**
SST-2 [14]	✓ Human judges	✓	✓ 11,855	✓ Binary	✓	**Include**
TweetEval sentiment [13]	✓ Crowd-sourced	✓	✓ 12,284	✓ 3-class	✓	**Include**
Twitter Financial News [TFNS]	✓ Crowd-sourced	✓	✓ 11,932	✓ 3-class	✓	**Include**
VADER validation [8]	✓ Screened raters	✓	✓ 23,703	✓ Convertible	✓	**Include**
IMDb [25]	✗ Star rating proxy	✓	✓	✓	✓	Exclude (IC1)
Yelp Polarity [26]	✗ Star rating proxy	✓	✓	✓	✓	Exclude (IC1)
Amazon Reviews [27]	✗ Star rating proxy	✓	✓	✓	✓	Exclude (IC1)
Sentiment140 [28]	✗ Emoticon proxy	✓	✓	✓	✓	Exclude (IC1)
SemEval-2014 ABSA [29]	✓ Expert	✓	✓	✗ Aspect-level	✓	Exclude (IC4)
SentFiN 1.0	✓ Expert	✗ Limited	✓	✓	✓	Exclude (IC2)

Supplementary Table S2. Sample and word-length statistics

Per-dataset sample sizes, class proportions, and word-length distribution for the experimental samples.

ID	Dataset	Full N	Sample n	Pos%	Neg%	Neu%	Word len (mean/median)	Word len (Q1/Q3/max)	Annotation
D1	Financial PhraseBank	2,264	2,000	25.4	13.4	61.3	22.4 / 21	15 / 28 / 81	16 finance experts
D2	SST-2	872	872	50.9	49.1	0.0	19.5 / 19	13 / 26 / 47	3 human judges
D3	TweetEval sentiment	12,284	2,000	19.3	31.5	49.3	14.9 / 15	11 / 19 / 32	Crowdsourced (AMT)
D4	TFNS	11,931	2,000	18.5	15.6	65.9	12.2 / 12	9 / 15 / 33	Crowdsourced
D5	VADER validation	23,703	2,000	51.2	46.9	2.0	17.3 / 16	11 / 23 / 146	10–20 screened raters
	Total	**51,054**	**8,872**						

Supplementary Table S3. Dataset provenance and sampling details

ID	Dataset	Citation	Source	License	Sampling	Notes / key challenge
D1	Financial PhraseBank	Malo et al. (2014) [12]	HuggingFace `gtfintechlab/financial_phrasebank_sentences_allagree`	CC BY-NC-SA 3.0	All-agree configuration (100% annotator consensus, 2,264 sentences); sample 2,000. Sample labels: neutral 61.3%, positive 25.4%, negative 13.4%.	Investor-perspective framing: models must classify e.g. “Cargo traffic fell 1% year-on-year” as negative, attending to the directional signal (“fell”) rather than the absolute quantity.
D2	SST-2	Socher et al. (2013) [14]	HuggingFace `stanfordnlp/sst2` (validation split)	Research use	Full validation split (872 sentences).	Strictly binary — no neutral gold labels. Any “neutral” prediction is a misclassification; we exploit this to measure spurious neutral generation.

ID	Dataset	Citation	Source	License	Sampling	Notes / key challenge
D3	TweetEval sentiment	Barbieri et al. (2020) [13]	HuggingFace `cardiffnlp/tweet_eval`	CC 3.0	Sample 2,000 from the test split (12,284 tweets).	Three-class crowd-sourced annotation. Informal language, user handles (`@user`), hashtags, and emoji.
D4	TFNS	Zeroshot (2022)	HuggingFace `zeroshot/twitter-financial-news-sentiment`	MIT	Sample 2,000 from the combined train+validation split.	Bullish/Bearish/Neutral labels mapped to positive/negative/neutral. Short analyst-headline style with stock tickers.
D5	VADER validation	Hutto & Gilbert (2014) [8]	`github.com/cjhutto/vaderSentiment`	MIT	Sample 2,000 from the four sub-corpora (tweets, NYT editorials, Amazon reviews, movie snippets).	Threshold: ≥ 0.05 = positive, ≤ -0.05 = negative, else neutral. Neutral class very small (2.0%) by design.

Supplementary Table S4. Accuracy by model and dataset

Model	D1 FPB	D2 SST-2	D3 TweetEval	D4 TFNS	D5 VADER	Mean
Claude Opus 4.7	**0.961**	0.937	0.649	0.688	0.794	**0.806**
GPT-4o	0.914	0.843	**0.704**	**0.765**	0.679	0.781
GPT-5.5	0.959	**0.951**	0.641	0.677	**0.824**	**0.810**
Gemini 2.5 Pro	0.532	0.299	0.315	0.334	0.374	0.371
Llama 3.1 8B	0.707	0.779	0.662	0.722	0.636	0.701
VADER (baseline)	0.564	0.557	0.533	0.469	0.601	0.545
RoBERTa (supervised baseline)	0.709	0.690	0.692	0.702	0.611	0.681

Supplementary Table S5. Bootstrap 95% confidence intervals for mean Macro F1

1,000 resamples, seed = 42.

Model	Mean Macro F1	95% CI
Claude Opus 4.7	0.775	[0.767, 0.783]
GPT-5.5	0.776	[0.767, 0.783]
GPT-4o	0.757	[0.748, 0.765]
Llama 3.1 8B	0.659	[0.649, 0.669]
VADER (baseline)	0.500	[0.489, 0.512]
Gemini 2.5 Pro	0.335	[0.325, 0.346]

Supplementary Table S6. Per-class F1 by model (averaged across all datasets)

Neutral F1 is averaged over the four datasets with a gold neutral class (D1, D3, D4, D5); SST-2 contributes only positive and negative F1.

Model	Neg F1	Neu F1	Pos F1	Macro F1
Claude Opus 4.7	0.840	0.601	0.813	**0.775**
GPT-4o	0.827	0.636	0.756	0.757
GPT-5.5	0.836	0.598	0.821	**0.776**
Gemini 2.5 Pro	0.286	0.380	0.345	0.335
Llama 3.1 8B	0.755	0.572	0.594	0.659
VADER (baseline)	0.459	0.472	0.547	0.501
RoBERTa (supervised baseline)	0.645	0.592	0.619	0.631

Supplementary Table S7. Spurious neutral rate on D2 SST-2

Percentage of predictions = neutral; SST-2 gold has no neutral class, so any neutral prediction is an error.

Claude Opus 4.7	GPT-4o	GPT-5.5	Gemini 2.5 Pro	Llama 3.1 8B	VADER
3.4%	10.6%	**1.1%**	32.2%	18.6%	18.1%

Supplementary Table S8. Pairwise κ by dataset for high-agreement pairs

Pair	D1 FPB	D2 SST-2	D3 TweetEval	D4 TFNS	D5 VADER
Claude × GPT-5.5	**0.952**	**0.914**	0.832	0.886	0.862
Claude × GPT-4o	0.806	0.782	0.725	0.737	0.751
GPT-4o × GPT-5.5	0.794	0.757	0.692	0.710	0.725
Claude × Llama	0.326	0.682	0.717	0.613	0.669
GPT-4o × Llama	0.415	0.760	0.698	0.715	0.743

Supplementary Table S9. Multi-model consensus and plurality rates

Rows where all 6 models produced valid predictions.

Dataset	n	Full consensus (6/6)	Plurality ≥5/6
D1 FPB	1,999	486 (24.3%)	1,187 (59.4%)
D2 SST-2	869	101 (11.6%)	487 (56.0%)
D3 TweetEval	1,935	281 (14.5%)	1,042 (53.9%)

Dataset	n	Full consensus (6/6)	Plurality ≥5/6
D4 TFNS	1,996	280 (14.0%)	1,051 (52.7%)
D5 VADER	1,989	322 (16.2%)	1,144 (57.5%)

Supplementary Table S10. Intra-model (run-to-run) Cohen's κ vs. inter-model reference

Claude Opus 4.7 and GPT-5.5 re-run twice each on a stratified 500-sentence subset under identical prompt and decoding settings. (Twenty GPT-5.5 responses exceeded the output token budget on long inputs and are excluded, hence n = 480 for that model.)

Model	Intra-model κ	95% CI	Label flips	Flip rate
Claude Opus 4.7	0.985	[0.970, 0.997]	5 / 500	1.0%
GPT-5.5	0.994	[0.984, 1.000]	2 / 480	0.4%
Claude × GPT-5.5 (inter-model)	*0.889*	—	—	—

Supplementary Table S11. Macro F1 by domain/register category (excluding Gemini 2.5 Pro)

Model	Formal financial (D1)	Formal general (D2)	Informal financial (D4)	Informal general (D3)	Cross-domain (D5)
Claude Opus 4.7	**0.958**	0.953	0.679	0.652	0.633
GPT-4o	0.906	0.887	**0.733**	**0.698**	0.560
GPT-5.5	0.956	**0.956**	0.671	0.641	**0.655**
Llama 3.1 8B	0.565	0.854	0.674	0.664	0.537
VADER (baseline)	0.479	0.595	0.427	0.531	0.472

Supplementary Table S12. Model-selection decision rule

If your situation is…	…then prefer	Rationale
Formal financial text (earnings, analyst reports), accuracy-first	Claude Opus 4.7 or GPT-5.5	F1 > 0.95 on FPB; κ = 0.889 with each other → consistent results either way
Informal financial text (analyst tweets, financial social media)	GPT-4o	best on TFNS (F1 = 0.733) and informal registers generally
On-premise deployment (privacy or data-control needs)	Llama 3.1 8B	open and runs locally, ~85% of frontier accuracy
High-stakes *individual* decisions	2–3 model ensemble + human review for non-plurality items	41–47% of items lack a 5-model majority; ensembling stabilises borderline cases [36, 42, 43]
Considering a reasoning-capable model in batch	validate on your task first	Gemini 2.5 Pro fails here regardless of thinking budget; do not assume parity

Supplementary Table S13. McNemar's test on per-sentence correctness, all 15 model pairs

Pooled across datasets; continuity-corrected; Bonferroni-adjusted threshold $\alpha = 0.05/15 = 0.0033$.

Pair	b (A right, B wrong)	c (A wrong, B right)	$\chi^2(1)$	p	Significant?
Claude × GPT-5.5	266	283	0.5	0.50	no
Claude × GPT-4o	697	552	16.6	4.6×10□□	yes
GPT-4o × GPT-5.5	607	772	19.5	1.0×10□□	yes
Claude × Llama	1,420	559	373.7	<10□□^{2}	yes
GPT-4o × Llama	1,113	397	338.6	<10□□□	yes
GPT-5.5 × Llama	1,475	591	377.4	<10□□^{3}	yes
Claude × VADER	2,899	716	1,317.0	<10□^{2}□□	yes
GPT-4o × VADER	2,886	847	1,112.6	<10□^{2}□^{3}	yes
GPT-5.5 × VADER	2,833	629	1,401.9	<10□^{3}□□	yes
Llama × VADER	2,535	1,226	454.9	<10□^{1}□□	yes
Claude × Gemini	4,202	570	2,762.8	≈0	yes
GPT-4o × Gemini	4,121	633	2,557.7	≈0	yes
GPT-5.5 × Gemini	4,210	558	2,795.7	≈0	yes
Llama × Gemini	3,629	877	1,679.5	≈0	yes
VADER × Gemini	2,917	1,468	478.2	<10□^{1}□□	yes

All differences are significant at the Bonferroni-corrected threshold except Claude Opus 4.7 vs. GPT-5.5, confirming the two leaders are statistically tied while all other pairwise gaps are real.

Supplementary Table S14. Sample representativeness: subset vs. full source pool

Each experimental sample is compared to its full source pool on text length. Stratified sampling preserves label proportions (Supplementary Table S2); the sample also matches the full pool on mean/median/SD word length and mean character length. SST-2 is the complete validation split, so no sampling error applies.

Dataset	n (sample / full)	Word length mean (sample / full)	Median (sample / full)	SD (sample / full)	Char length mean (sample / full)
D1 FPB	2,000 / 2,264	22.3 / 22.4	21 / 21	10.0 / 10.1	121.5 / 122.0
D2 SST-2	872 / 872	19.5 / 19.5	19 / 19	8.8 / 8.8	105.8 / 105.8
D3 TweetEval	2,000 / 12,284	14.8 / 14.9	15 / 15	5.6 / 5.6	90.9 / 91.3
D4 TFNS	2,000 / 11,931	12.3 / 12.2	11 / 12	4.7 / 4.7	87.1 / 86.0
D5 VADER	2,000 / 23,703	17.2 / 17.3	16 / 16	8.7 / 9.0	100.4 / 100.7

Supplementary Table S15. Extended qualitative disagreement examples

Representative cases with the full model prediction set. Prevalence across the pooled 8,872 sentences (four strong LLMs = Claude, GPT-4o, GPT-5.5, Llama): all four disagree with gold, 986 cases; all four agree with each other yet differ from gold, 865 cases; Claude vs. GPT-5.5 disagree, 626 cases; cashtag/ticker sentences with a model split, 303 cases; negation sentences with a model split, 214 cases.

Sentence (excerpt)	Dataset	Gold	Claude	GPT-5.5	Phenomenon
“hilariously inept and ridiculous.”	SST-2	positive	negative	negative	Ambiguous/questionable gold; all four LLMs agree negative
“reign of fire looks as if it was made without much thought — and is best watched that way.”	SST-2	positive	negative	negative	Sarcasm / backhanded phrasing
“my thoughts were focused on the characters.”	SST-2	positive	neutral	neutral	Weak/implicit sentiment; models read as neutral
“or doing last year’s taxes with your ex-wife.”	SST-2	negative	negative	neutral	Leader disagreement; implicit negative simile
“we root for (clara and paul), even like them…”	SST-2	positive	negative	positive	Leader disagreement on mixed sentiment
“you don’t have to know about music to appreciate the film’s easygoing blend…”	SST-2	positive	positive	positive	Negation (“don’t”); weaker Llama flips to neutral/negative
“$GM – GM loses a bull …”	TFNS	negative	negative	negative	Ticker + finance idiom (“loses a bull”)
“Featured #Medical #Marijuana #Stock: … (OTC: $PSIQ) …”	TweetEval	neutral	neutral	positive	Promotional/spam ticker; model split
“@user #Dems mastered how to RIG USA/ system…”	TweetEval	neutral	negative	negative	Political content; gold neutral, models read negative

Supplementary Table S16. Invalid/unresolved output rate by model and dataset

Count of outputs that could not be resolved to a valid label after normalisation and two retries (excluded per-metric via pairwise complete case). The 85 total (0.16%) come entirely from GPT-5.5 and Llama; **Gemini 2.5 Pro produced zero unresolved outputs**, the lowest of any model.

Model	D1 FPB	D2 SST-2	D3 TweetEval	D4 TFNS	D5 VADER	Total
Claude Opus 4.7	0	0	1	0	0	1
GPT-4o	0	0	0	0	0	0
GPT-5.5	1	3	22	4	7	37
Gemini 2.5 Pro	0	0	0	0	0	**0**
Llama 3.1 8B	0	0	43	0	4	47
VADER	0	0	0	0	0	0

Supplementary Table S17. Gemini 2.5 Pro confusion matrices per dataset

Rows are the human gold label; columns are Gemini's predicted label (counts, valid predictions only). On D2 SST-2 the gold labels are strictly binary (no neutral gold). The matrices show a coherent systematic bias — near the class prior on D1 and over-generation of neutral on binary SST-2 — rather than random or dropped outputs.

Dataset	Gold	→ positive	→ negative	→ neutral
D1 FPB	positive	182	80	245
D1 FPB	negative	74	49	144
D1 FPB	neutral	252	141	833
D2 SST-2	positive	154	138	152
D2 SST-2	negative	192	107	129
D3 TweetEval	positive	150	142	94
D3 TweetEval	negative	248	229	152
D3 TweetEval	neutral	369	366	250
D4 TFNS	positive	125	130	114
D4 TFNS	negative	101	92	119
D4 TFNS	neutral	283	585	451
D5 VADER	positive	366	391	267
D5 VADER	negative	287	365	285
D5 VADER	neutral	16	7	16

Supplementary Table S18. Pairwise inter-model Cohen's κ with bootstrap 95% confidence intervals

Averaged across the five datasets; 1,000 bootstrap resamples, seed = 42, for all 21 classifier pairs. Confidence intervals accompany the point estimates in main-text Table 4. RoBERTa (supervised baseline) agrees with the LLMs at Llama's level (κ = 0.47–0.54) and near zero with Gemini. Notably, Claude × GPT-5.5 (0.889 [0.881, 0.898]) and Claude × GPT-4o (0.760 [0.747, 0.773]) have non-overlapping intervals — the leading pair is separable from the next-closest pair.

Pair	κ	95% CI
Claude × GPT-5.5	0.889	[0.881, 0.898]

Pair	κ	95% CI
Claude × GPT-4o	0.760	[0.747, 0.773]
GPT-4o × GPT-5.5	0.736	[0.723, 0.749]
GPT-4o × Llama	0.666	[0.651, 0.681]
Claude × Llama	0.601	[0.588, 0.616]
GPT-5.5 × Llama	0.585	[0.572, 0.599]
RoBERTa × GPT-4o	0.535	[0.518, 0.550]
RoBERTa × Llama	0.518	[0.499, 0.536]
RoBERTa × Claude	0.478	[0.463, 0.492]
RoBERTa × GPT-5.5	0.474	[0.460, 0.489]
GPT-5.5 × VADER	0.317	[0.301, 0.332]
Claude × VADER	0.292	[0.277, 0.308]
RoBERTa × VADER	0.288	[0.272, 0.303]
GPT-4o × VADER	0.283	[0.268, 0.299]
Llama × VADER	0.242	[0.228, 0.257]
GPT-5.5 × Gemini	0.032	[0.017, 0.047]
Claude × Gemini	0.030	[0.015, 0.045]
GPT-4o × Gemini	0.029	[0.014, 0.045]
RoBERTa × Gemini	0.016	[0.001, 0.030]
Gemini × VADER	0.014	[−0.001, 0.029]
Gemini × Llama	0.002	[−0.013, 0.017]

Supplementary Note S1. Diagnostic analysis of the Gemini 2.5 Pro result

Because a frontier model scoring at or below chance invites the hypothesis of a measurement or output-parsing artefact rather than a genuine model deficiency, we provide three diagnostics.

(1) Gemini produced zero unresolved outputs (Supplementary Table S16): all 8,872 of its predictions were successfully normalised to a valid label — the lowest invalid rate of any model, and lower than the well-performing GPT-5.5 (37) and Llama (47). Its near-chance result therefore cannot arise from silently dropped or mishandled predictions.

(2) The confusion matrices are coherent, not random (Supplementary Table S17). On D1 FPB, Gemini's prediction distribution (positive 25%, negative 14%, neutral 61%) almost exactly reproduces the gold class prior (25.4 / 13.4 / 61.3%) while only weakly tracking individual sentences, which is the signature of a model predicting near the marginal distribution — hence $\kappa \approx 0$. On the strictly binary D2 SST-2 it over-generates neutral (32%) and, on polar items, predicts positive slightly more often for gold-negative than gold-positive sentences, producing the below-chance κ (-0.062). This is a systematic behavioural pattern, not parsing noise.

(3) Identical normalisation across providers. The same normalisation pipeline (exact match → synonym mapping → two retries) was applied to every model's raw output; Gemini required no synonym-salvage or retries (zero unresolved), so no provider-specific handling advantaged or disadvantaged it.

Limitations. These diagnostics rule out a measurement or parsing artefact, but not the *serving-pathway* confound: Gemini was the only model evaluated through the Batch API, so its near-chance result cannot be cleanly attributed to the model rather than the Batch-API pathway. In addition, the raw pre-normalisation Batch-API responses were not retained, so verbatim outputs cannot be displayed (the diagnostics above use the normalised predictions in the reproducibility package). A live-API replication with raw outputs retained is needed to separate model from pathway; the main-text claims are scoped to this Batch-API configuration accordingly.

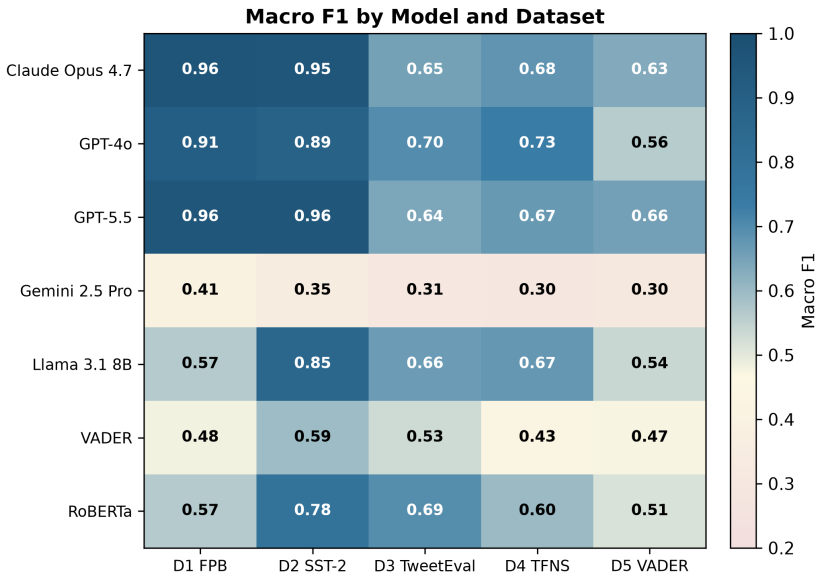

Figure 1: Supplementary Figure S1

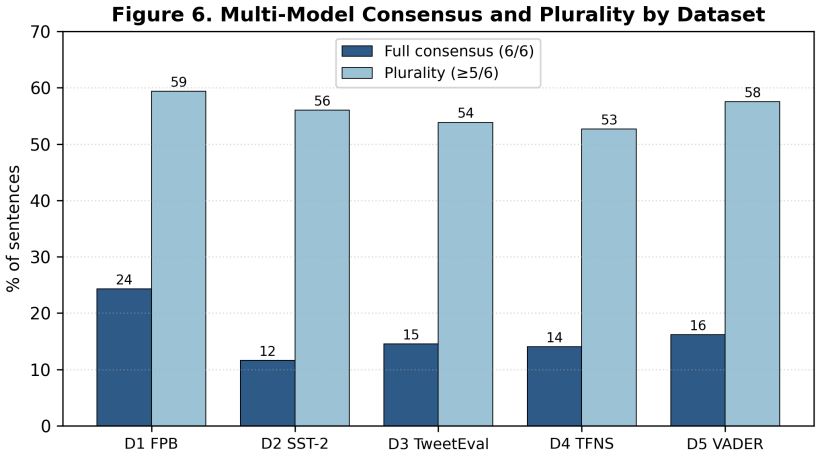

Figure 2: Supplementary Figure S2

Supplementary Figures

Supplementary Fig. S1. Macro F1 heatmap by model and dataset

Heatmap of Macro F1 for each model on each dataset (the values tabulated in main-text Table 2, "Macro F1 by Model and Dataset"). All non-Gemini LLMs peak on the formal datasets (D1, D2).

Supplementary Fig. S2. Full-consensus and plurality agreement rates per dataset

Full six-model consensus never exceeds 25%; plurality (≥5/6) ranges 53–59% (values in Supplementary Table S9).

Supplementary Fig. S3. Domain/register matrix of the five datasets

The five datasets span a 2×2 domain (financial vs. general) × register (formal vs. informal) grid, plus one cross-domain anchor (D5).

	Financial domain	General domain
Formal register	D1 Financial PhraseBank	D2 SST-2 (movie reviews)
Informal register	D4 Twitter Financial News (TFNS)	D3 TweetEval (social media)
Cross-domain	D5 VADER validation (tweets + news + reviews + movies)	

Supplementary Methods

Reproducibility

- All 5 datasets downloaded, sampled (seed = 42, stratified), and validated.
- Model version strings recorded at API call time (June 2026).
- Temperature = 0 for GPT-4o and Llama 3.1 8B; not applicable for Claude Opus 4.7, GPT-5.5, and Gemini 2.5 Pro (Batch API).
- Unresolved predictions: 85 / 53,232 (0.16%), excluded via pairwise complete case.
- Software: scikit-learn $\geq$ 1.3, scipy $\geq$ 1.10, pandas $\geq$ 2.0; supervised baselines via Hugging Face `transformers`.
- Bootstrap 95% CIs: 1,000 resamples, seed = 42 (Supplementary Table S5).
- Intra-model κ baseline: Claude + GPT-5.5 re-run ×2 on a 500-sentence subset (Supplementary Table S10).
- Gemini overthinking control: Batch re-run at minimum thinking budget (128 tokens) on a 500-sentence subset.
- McNemar's test with Bonferroni correction (α = 0.0033 for 15 pairs; Supplementary Table S13).

Companion analysis scripts (released with the reproducibility package)

- `substudy_roberta_baseline.py` — tweet-trained RoBERTa [60] on all five datasets (supervised general-transformer baseline).
- `substudy_prompt_robustness.py` — re-runs all five LLMs on a 500-sentence subset under an alternative minimal prompt, to test prompt sensitivity of the inter-model κ matrix.
- Representativeness check comparing each sample to its full source pool (label proportions, word and character length; Supplementary Table S14).

All scripts, per-sentence predictions, and metrics are available in the public repository cited in the main-text Data Availability statement (https://github.com/aneeshks/Multi-Domain-Sentiment-Benchmark).