

PRAXIS: An English-Review Benchmark for Pragmatic-Form Bias in Suggestion Mining

Aniket Upadhyay^{1†}, Aakash Trivedi^{1†}, Sparsh Bajoria², Shreyanshu Ranjan²
Dhruv Kumar¹, Pratik Narang¹

¹ Department of Computer Science & Information Systems,
Birla Institute of Technology and Science, Pilani, India
p20241007@pilani.bits-pilani.ac.in, f20191076P@alumni.bits-pilani.ac.in,
dhruv.kumar@pilani.bits-pilani.ac.in, pratik.narang@pilani.bits-pilani.ac.in

² Kalinga Institute of Industrial Technology, Bhubaneswar, India
2329214@kiit.ac.in, 2305013@kiit.ac.in

Abstract

Suggestion mining has been treated as binary classification since its introduction more than a decade ago, but the underlying linguistic object is not binary. A single review can express the same operational recommendation as a direct imperative (*add a vegetarian entrée*), an optative wish, an interrogative request, or a hedged conditional. These utterances share illocutionary force but differ sharply in surface pragmatic form, and they make very different demands on a model: lexical cue-spotting suffices for the first, indirect speech-act interpretation for the rest. No prior benchmark measures these distinctions: existing resources live in one to five domains and label suggestions at one binary level.

We introduce **PRAXIS**¹, an English consumer-review benchmark that reframes suggestion mining as the measurement of *pragmatic-form bias* rather than a leaderboard. PRAXIS releases (i) a silver corpus of 174,860 reviews across 68 industries, labeled by an LLM annotator (Claude Opus 4.7), (ii) a 15,000-row human-labeled gold evaluation set, with reliability estimated by a three-annotator panel on a 1,000-row subset (Tier 1 Fleiss $\kappa = 0.77$), and (iii) a benchmark of four model families scored against the human gold with bootstrap 95% confidence intervals. Opus-vs-human agreement is reported as a corpus-validation statistic (Tier 1 Cohen’s $\kappa = 0.71$), not as inter-annotator agreement.

Our headline finding is structural rather than competitive. Instruction-tuned LLMs lead encoders on binary macro-F1 (deepseek-v4-pro 0.882 vs. RoBERTa-large 0.780), but the ranking reverses in one Tier 1 form cell, the conditional. RoBERTa-large recalls 85.2% of

conditional suggestions; the strongest zero-shot LLM recalls 46.7% and the other six recall 8.0–43.1%. This is a form-classification gap, not a detection gap: the models identify these rows as suggestions at the binary level but mis-route them at the form classifier. A matched-supervision QLoRA run on Qwen2.5-7B, trained on the same silver labels as the encoders, recovers encoder-level conditional recall (0.904). The benchmark surfaces a structural error mode that flat binary scores cannot detect. PRAXIS is a measurement instrument, not a leaderboard.

1 Introduction

A user writing a restaurant review can communicate the same operational recommendation in several pragmatically distinct ways. *Add more vegetarian options*, *the menu would benefit from more vegetarian options*, *could you add a vegetarian dish?*, and *the food was disappointing* differ along two axes: the first three share an illocutionary force (a directive aimed at the restaurant) but diverge in surface form (imperative, conditional, interrogative); the fourth shares no force at all. A business owner sifting customer feedback at scale needs the first three collapsed (all are actionable suggestions) and the fourth separated (opinion, not directive). The literature has, since its inception, posed the problem as binary suggestion classification (Ramanand et al., 2010; Negi and Buitelaar, 2015), and existing benchmarks treat it that way (Negi et al., 2018, 2019). We argue that this framing hides the empirical question that matters most.

We use one real gold row as a running example (gold_plus_03361, Amazon, Automotive):

[could be a bit stronger] but handy

Gold: hedged conditional. Zero-shot instruction-tuned LLMs route it to modal_deontic or

[†]Equal contribution.

¹<https://huggingface.co/datasets/au-hf/praxis>.

optative. A Qwen2.5-7B model LoRA-tuned on the same silver labels as our encoders routes it to conditional. Binary detectors flag the row as a suggestion either way; a form-aware queue does not. A domain $\times$ form gallery of further real reviews is in Appendix V.

What PRAXIS measures, and what it does not. We define **pragmatic-form bias** as a systematic difference in per-form recall across the six Tier-1 communicative forms, *conditional on a correct binary decision*. It is a performance-slice construct in the sense of Ribeiro et al. (2020) and Bean et al. (2026), not a social or fairness claim. PRAXIS does not measure general pragmatic competence, multilingual directive systems, or genres other than English consumer reviews.

The argument turns on two observations. First, a flat binary score cannot distinguish two models that agree at the binary level but disagree at the pragmatic-form level: a model that reaches 85% binary recall by mapping *please add*, *add*, and *add more* to the positive class is empirically very different from one that reaches the same recall by treating wishes as suggestions but conditional statements as counterfactual evaluation. Both will report similar headline numbers and silently differ on a substantial fraction of an operational queue. Second, suggestion forms are not distributed uniformly: on the natural-distribution sampling bucket of our gold, 8.7% of human-labeled positives are direct imperatives, with the remainder spread across optatives (36.8%), modal/deontic (27.3%), conditionals (27.2%), and effectively zero interrogatives or comparatives. The distribution is skewed; a benchmark that does not resolve form will be dominated by whichever form a model happens to lexicalize.

This paper builds the instrument needed to measure the resulting behavior. Our thesis is that pragmatic form, not raw scale, is the missing empirical axis in suggestion mining. We do not lead with a leaderboard. Form-stratified slicing reveals a structural error mode that flat binary benchmarks cannot detect: the encoder advantage in per-form recall over zero-shot instruction-tuned LLMs is concentrated almost entirely in the conditional form cell.

Contributions. (i) A 174,860-review silver corpus spanning 68 business categories grouped into four macro-domains, drawn from public review platforms (Ni et al., 2019). (ii) A three-tier pragmatic annotation scheme grounded in speech-act

theory (Austin, 1975; Searle, 1969, 1975; Levinson, 1983): communicative form (six classes), affective/illocutionary modifiers (four multi-label tags), and operational specificity (three-point ordinal). (iii) A 15,000-row human-labeled gold evaluation set (PRAXIS-gold); a single human annotator labeled every row from scratch, blind to silver labels, and reliability is estimated by a three-annotator panel on a stratified 1,000-row subset (subset Fleiss κ on Tier 1 is 0.77). (iv) A benchmark of four model families (classical, encoder fine-tunes, small open LLMs, and instruction-tuned LLMs served over APIs) with bootstrap 95% confidence intervals on every reported cell, scored against the human gold, plus a matched-supervision LoRA test showing that the encoder-LLM conditional gap is a training/convention effect rather than an architecture effect. Encoders are trained on silver-174K minus the 15K gold rows.

2 Related Work

Suggestion mining. The task was introduced by Ramanand et al. (2010) as wish detection in product reviews and operationalized as binary suggestion classification by Negi and Buitelaar (2015) and Brun and Hagège (2013). Negi et al. (2018) released an open-domain dataset spanning five domains and announced the forthcoming shared task. SemEval-2019 Task 9 (Negi et al., 2019) then provided the first shared benchmark, with roughly 9,000 posts drawn from a software-suggestions forum and a hotel-review subdomain. PRAXIS differs from this line on three axes: scale (174,860 silver and 15,000 gold rows vs. benchmarks under 10,000), category coverage (68 business categories grouped into four review-platform macro-domains), and task structure (a three-tier pragmatic typology rather than a single binary flag). We do not claim that scale alone is a contribution; we claim that the combination of scale, category-level spread, and tiered structure is what makes pragmatic-form-bias measurement possible.

Behavioral / slice-based evaluation. PRAXIS’s stratified per-form, per-modifier, and per-industry slicing is methodologically adjacent to CheckList (Ribeiro et al., 2020), which argues that aggregate accuracy can hide systematic failure modes that surface only when behavior is examined on linguistically-defined sub-populations. Our per-form recall analysis (§6.2) is a CheckList-style slice test against a pragmatic typology, with slices

produced by human gold annotation against a typology rather than by templated tests.

Speech acts and pragmatic typology. The distinction between locutionary form and illocutionary force (Austin, 1975; Searle, 1969) grounds the three-tier scheme. The interrogative and conditional classes correspond to canonical indirect speech acts (Searle, 1975; Levinson, 1983); the comparative form encodes directive intent through conversational implicature (Grice, 1975). The Tier 2 modifier set picks up surface markers correlated with politeness and affect (Brown and Levinson, 1987). The closest computational traditions are dialogue-act tagging (Stolcke et al., 2000) and request detection (Lampert et al., 2010), both on conversational data.

Weak supervision and disagreement. LLM-assisted labeling has become standard at scale (Wang et al., 2021; Gilardi et al., 2023). Subsequent work cautioned that LLM-produced labels require explicit downstream validation against human gold (Pangakis et al., 2023); PRAXIS adopts that prescription literally by reserving the gold for blind human labeling. The disagreement-as-signal literature (Aroyo and Welty, 2015; Plank et al., 2014; Mostafazadeh Davani et al., 2022; Kirk et al., 2024) has argued that genuine pragmatic ambiguity should be preserved. Adjacent benchmarks for sentiment (Maas et al., 2011; McAuley and Leskovec, 2013) and aspect-based sentiment analysis (Pontiki et al., 2016) complement PRAXIS at a higher pragmatic granularity.

3 The PRAXIS Dataset

3.1 Sources and Licensing

PRAXIS draws from two source families spanning 68 industries. Service-side reviews come from the Yelp Open Dataset. Product and digital/media reviews come from the Amazon corpus (McAuley and Leskovec, 2013) as updated by Ni et al. (2019), sampled at 5,000 rows per category (155K Amazon rows). Preprocessing applies a fastText English filter (Joulin et al., 2017), which deduplicates by (review ID, industry). A per-source redistribution matrix is in Appendix A.

3.2 Construction: Two Passes, Strictly Separated

PRAXIS distinguishes two pipelines that prior LLM-assisted labeling work tends to conflate.

Silver corpus (174,860 rows). The full corpus is labeled by a single LLM annotator, Claude Opus 4.7, under the guidelines in Appendix B. A high-recall RoBERTa-large prefilter runs first at 95% recall as a *cheap prefilter* to bound the LLM-labeling pass, so the expensive LLM scores likely positives rather than the whole pool; 71.4% of the prefilter’s positive predictions are confirmed by the LLM annotator; a stratified negative-recovery pass on held-out negatives recovers a further 1.5%. The prefilter never touches the human gold. We release this set as silver: model-produced, suitable for training and analysis, but not the evaluation reference.

Gold evaluation set (PRAXIS-gold; 15,000 rows). A 15,000-row subset is drawn under stratified sampling across (industry, predicted-class), with positive/negative balance held at 7,500/7,500, all 68 industries represented, and low-resource industries included in full. A single human annotator labeled every row from scratch, blind to silver labels. These human labels are PRAXIS-gold and the sole evaluation reference.

Train/eval disjointness. Encoders, and any model trained from silver data (including the LoRA run), are trained on silver-174K minus the 15,000 gold rows. The gold rows are held out of all training by (review_id, industry) and normalized text, preventing the train/test contamination that would otherwise arise from overlap between silver-produced training labels and gold-produced test labels on identical reviews.

Silver/gold relationship. The two label sets are produced by different annotators (Opus 4.7 vs. human), under different processes, on overlapping text. The annotator producing the corpus labels is not the system under test, so a model trained on silver and evaluated against gold is not being scored against its own training oracle. The overlap also lets us validate the silver corpus by measuring Opus-vs-gold agreement (§3.5), which we report explicitly as a silver-validation statistic, not as inter-annotator agreement.

Class balance. The 50/50 positive/negative split in gold is a deliberate evaluation choice. Natural prevalence varies across macro-domains (5.1% product, 5.9% digital/media, 26.8% hybrid retail, 51.2% service); PRAXIS releases a per-row weight eval_weight_natural_binary so any binary metric can be re-weighted to natural prevalence.

3.3 Pragmatic Typology

A suggestion in PRAXIS is a business-directed utterance specifying a concrete operational change. Excluded: opinions, complaints without a directive, customer-to-customer advice, pure product description.

Tier 1 (single-label, six classes). Direct imperative (“add more vegetarian options”); modal/deontic (“you should offer more vegetarian options”); optative/wish (“I wish they offered more vegetarian options”); conditional/hypothetical (“it would be great if they offered more vegetarian options”); interrogative (“could you add a vegetarian option?”); comparative (“the falafel option here is much smaller than at other locations”). When a span carries more than one form, the priority order imperative > modal > optative > conditional > interrogative > comparative determines the label.

Tier 2 (multi-label, four tags). Frustrated, sarcastic, backhanded (praise-then-suggest), hedged. Sarcastic and backhanded are *diagnostic/soft-label* targets only (§8; Appendix I).

Tier 3 (3-point ordinal). Vague (“do better with the food”) < specific (“add more vegetarian options”) < very specific (“add a tofu bowl and a chickpea wrap to the lunch menu”).

3.4 Dataset Statistics

Table 1 shows the Tier 1 distribution on the gold positive slice. On the stratified-only bucket (closest to natural prevalence), three classes account for 91.3%: optative (36.8%), modal/deontic (27.3%), conditional (27.2%). Direct imperatives are 8.7% (stratified) and 8.6% (balanced), so the “imperatives are a minority” headline is robust to the rare-taxonomy oversample. Interrogatives (3.4%) and comparatives (0.1%) on the balanced column are entirely a product of that supplement and should be read as benchmark inclusion, not prevalence estimates.

Tier 3 is dominated by specific (72.6%; 21.1% vague, 6.3% very specific). Tier 2 modifiers are sparse: 65.7% of positives carry no modifier; 14.8% frustrated, 13.1% hedged, 9.3% backhanded, 1.2% sarcastic. Of the 7,500 positives, 17.6% were flagged borderline; the rate ranges from 4.5% for direct imperatives to 36.2% for optatives, an 8× gap that is itself a finding about pragmatic form (full breakdown in Appendix D). The four macro-domains are unevenly populated on the 15K gold (Product 6,858; Service 4,422; Digital/media

Tier 1 form	Balanced		Stratified-only	
	Count	Share	Count	Share
Optative	2,581	34.4%	2,463	36.8%
Conditional	2,009	26.8%	1,823	27.2%
Modal/deontic	1,997	26.6%	1,826	27.3%
Direct imperative	647	8.6%	581	8.7%
Interrogative	258	3.4%	0	0.0%
Comparative	8	0.1%	0	0.0%
Total positives	7,500	100.0%	6,693	100.0%

Table 1: Tier 1 communicative form distribution on the gold positives. Balanced is the full released positive slice; stratified-only excludes the 807-row rare-taxonomy oversample.

2,086; Hybrid retail 1,634). All 68 industries are represented at $n \geq 66$. Extended counts, sampling-bucket flags, and the class-balance caveat are in Appendix K.

3.5 Inter-Annotator Agreement

We report agreement at three complementary granularities (Table 2), all computed on the same stratified 1,000-row subset of the gold (the only portion of PRAXIS-gold with more than one independent label per row); the remaining 14,000 rows are single-annotated and are not directly multi-annotator validated. (i) Symmetric human IAA on the subset measures whether the annotation scheme is reproducible across annotators. (ii) Bulk-vs-panel hold-one-out agreement on the same subset measures whether the bulk annotator (who produced the 14,000 singly-annotated rows) tracks panel consensus on the rows where consensus is available, and is the most decision-relevant reliability statistic for the single-annotator portion of the gold. (iii) Opus-vs-gold agreement is reported as a silver-corpus validation statistic, not as IAA. The subset is independently labeled by three trained annotators.

Binary identification is strong ($\kappa = 0.80$) and Tier 1 form is substantial (Fleiss $\kappa = 0.77$). Agreement decreases through the higher tiers, which we read as an empirical property of the phenomenon rather than a methodological weakness: frustrated ($\kappa = 0.72$) and hedged ($\kappa = 0.66$) remain substantially recoverable; sarcastic ($\kappa = 0.54$) and backhanded ($\kappa = 0.42$) are lower. Following [Lan-dis and Koch \(1977\)](#)’s interpretation, we position frustrated and hedged as hard-label benchmark targets and release sarcastic and backhanded as soft-label/diagnostic targets. The Opus-vs-gold column tracks at or below both human numbers on every

Tier / Task	Metric	Human IAA	Bulk-vs-panel	Opus-vs-gold
Binary detection	Cohen’s κ	0.80	0.78	0.78
Tier 1 communicative form (6-way)	Fleiss κ	0.77	0.72	0.71
Tier 3 operational specificity	Quad. weighted κ	0.67	0.64	—
Tier 2 – frustrated	Cohen’s κ	0.72	0.68	0.63
Tier 2 – hedged	Cohen’s κ	0.66	0.61	0.58
Tier 2 – sarcastic (diagnostic)	Cohen’s κ	0.54	0.49	0.47
Tier 2 – backhanded (diagnostic)	Cohen’s κ	0.42	0.37	0.39

Table 2: Agreement on the 1,000-row reliability subset. Human IAA estimates task reproducibility; bulk-vs-panel holds out the bulk annotator and compares against the majority of the other two; Opus-vs-gold is an auxiliary silver-validation statistic. Sarcastic and backhanded are diagnostic-only.

comparable task except backhanded, which is why all model evaluation is restricted to the human gold.

Critically, bulk-vs-panel agreement is within 3–5 κ points of symmetric IAA on every task (mean gap 0.04). The bulk annotator, who produced the 14,000 singly-annotated gold rows, therefore agrees with panel consensus at close to the rate the panel agrees among itself, across all four tasks (the conditional form is the exception; Appendix J), on the 1,000-row subset where consensus is available. This substantially reduces the likelihood that the released gold is dominated by annotator-specific labelling idiosyncrasies, but it is not a direct multi-annotator check on the 14,000 singly-annotated rows themselves; that limitation is restated in the Limitations section. The reliability claim for PRAXIS-gold is therefore the conjunction of subset panel IAA and the subset bulk-vs-panel hold-one-out, not a claim that all 15,000 rows have been multiply validated. Appendix J expands this interpretation.

Annotation design. 15K gold = 14K bulk-only + 1K triple-annotated panel (the bulk annotator is one of the three). Bulk-vs-panel on that 1K is the reliability link to the 14K. We release per-panelist subset labels and the gold borderline flag as a per-row confidence signal.

4 Benchmark Design

PRAXIS specifies four tasks. Each is evaluated on the 15,000-row gold; Tiers 1–3 are evaluated on the 7,500-row positive slice. Every reported cell includes a 1,000-resample bootstrap 95% CI; for encoders, the bootstrap is pooled across rows and seeds jointly (Appendix C).

- **Task A:** binary suggestion detection. Macro-F1 on $n = 15,000$.
- **Task B:** Tier 1 form (6-way single-label). Per-

form recall on $n = 7,500$ positives. Comparative ($n = 8$) is reported with a † marker and discussed in Appendix N.

- **Task C:** Tier 2 modifiers (multi-label). Macro-F1 and per-label F1. Results are in Appendix I; sarcastic and backhanded are diagnostic-only.
- **Task D:** Tier 3 specificity. Quadratic-weighted κ between predicted and gold levels (Appendix I).

Tasks C and D are discussed in the appendix rather than as main-paper leaderboard cells; they remain part of the released benchmark. Cross-domain transfer is reported in Appendix O.

5 Baselines and Models

We benchmark four families.

Classical baselines. Random (Bernoulli $p=0.5$), majority-zero, rule-cue regex, TF-IDF + logistic regression, TF-IDF + linear SVM. All five are trained on silver-174K minus the 15K gold rows.

Encoder fine-tunes. BERT-base-uncased, RoBERTa-base, DeBERTa-base, RoBERTa-large (Devlin et al., 2019; Liu et al., 2019; He et al., 2023). Three seeds per model ($\{13, 42, 137\}$); reported cells are seed means. Batch size 16, learning rate 2×10^{-5} , AdamW, weight decay 0.01, 10% warmup, three epochs, max sequence length 192, best epoch chosen by gold-dev macro-F1.

Small open LLMs. Qwen2.5-Instruct at 0.5B, 1.5B, 3B, 7B (Qwen Team), scored zero-shot by the first-token log-prob margin between *yes* and *no*, deterministic decoding.

Instruction-tuned LLMs (API-served). We call these models “API-served” because we queried them through a hosted API; the term describes how they were served, not how they are licensed, and most of the weights can be run locally. We score the full set of API-served models that completed

the 15K gold pass: deepseek-v4-pro, deepseek-v4-flash, gpt-oss:120b, gemma4:31b, nemotron-3-nano:30b, ministral-3:14b, and ministral-3:8b. Bootstrap confidence intervals are reported for all complete-coverage models. All are scored zero-shot with deterministic decoding (temperature 0, no sampling, no few-shot exemplars, one call per gold row, no retries on parse failure). Task A uses a one-token yes/no prompt; the joint Tier 1/2/3 task uses a single prompt that elicits a one-line JSON object `{"tier1": . . . , "tier2": [. . . ], "tier3": . . . }` with the enum values defined in §3. The *executed* prompts that produced Tables 3–5 are reproduced verbatim in Appendix P. These prompts do not state the §3.3 priority order (imperative > modal/deontic > optative > conditional > interrogative > comparative); it appears only in the P2 definition prompt of Table 6.

Matched-supervision LoRA. Qwen2.5-7B-Instruct, 4-bit QLoRA, trained on 8,285 gold-held-out silver positives (Task B). Details: Appendix S.

Pretraining contamination. The Amazon subset is almost certainly in every instruction-tuned LLM’s pretraining data; this is a structural limit of any public-review benchmark. The central finding (§6) is robust to this contamination because the test signal is the human-applied label, not the text, and per-form gaps are stratified within the same text. All models are scored against the 15K human gold; no model is scored against the silver labels.

6 Results

6.1 Task A: Binary Suggestion Detection

Macro-F1 on the 15,000-row gold (Table 3) places instruction-tuned LLMs at the top, with deepseek-v4-pro at 0.882 and the strongest fine-tuned encoder, RoBERTa-large, at 0.780. The strongest classical baseline (TF-IDF + SVM) reaches 0.740; the gap between it and the strongest encoder is approximately 4 macro-F1 points. Two observations preempt over-reading the table. First, at PRAXIS’s scale, residual binary error appears substantially driven by borderline ambiguity (§3.4), not by feature representation. Second, the LLMs’ binary lead hides a per-form reversal: encoders beat LLMs mainly in one Tier 1 cell (§6.2).

Reconciliation with the form-recall finding. Binary positive recall measures whether the model flags the row at all; per-form recall (Table 4)

Family	Model	macro-F1	95% CI
Instr. LLM	deepseek-v4-pro	0.882	[0.877, 0.887]
Instr. LLM	deepseek-v4-flash	0.870	[0.864, 0.876]
Instr. LLM	gpt-oss:120b	0.865	[0.859, 0.871]
Instr. LLM	gemma4:31b	0.856	[0.850, 0.862]
Instr. LLM	nemotron-3-nano:30b	0.823	[0.817, 0.829]
Encoder	RoBERTa-large (3 seeds)	0.780	[0.773, 0.787]
Encoder	DeBERTa-base (3 seeds)	0.760	[0.752, 0.768]
Encoder	RoBERTa-base (3 seeds)	0.748	[0.740, 0.756]
Linear	TF-IDF + SVM	0.740	[0.732, 0.748]
Encoder	BERT-base-uncased (3 seeds)	0.730	[0.722, 0.738]
Instr. LLM	ministral-3:14b	0.716	[0.708, 0.724]
Instr. LLM	ministral-3:8b	0.702	[0.694, 0.710]
Linear	TF-IDF + LR	0.701	[0.693, 0.709]
Rule	Cue regex	0.533	[0.524, 0.542]
Open LLM	Qwen2.5-7B-Instruct	0.472	[0.463, 0.481]
Open LLM	Qwen2.5-0.5B-Instruct	0.462	[0.453, 0.471]
Open LLM	Qwen2.5-1.5B-Instruct	0.446	[0.437, 0.455]
Open LLM	Qwen2.5-3B-Instruct	0.444	[0.435, 0.453]
Random	Bernoulli($p=0.5$)	0.496	[0.487, 0.505]
Majority	Always-zero	0.333	[0.333, 0.333]

Table 3: Task A binary leaderboard on the 15,000-row gold with 1,000-resample bootstrap 95% confidence intervals (Appendix C). Always-zero produces a constant prediction so the CI degenerates.

measures whether the model assigns the gold Tier 1 form. Binary positive recall on the conditional cell is high for every instruction-tuned LLM (range 0.84–0.95; full table in Appendix E). The conditional gap in Table 4 is therefore a form-classification gap, not a detection gap: the LLM identifies the row as a suggestion but assigns it to a different form bucket, most commonly modal/deontic or optative.

6.2 Task B: Pragmatic-Form Bias

The central finding of the paper appears in Table 4. Encoders peak on optatives (~ 0.89 – 0.90) and conditionals (~ 0.85 – 0.87), both of which carry strong lexical anchors. Instruction-tuned LLMs also peak on optatives (0.91–0.95 for the five strongest) and are strong on modal/deontic, but their recall on conditional suggestions falls sharply. RoBERTa-large recalls 0.852 of conditional suggestions; deepseek-v4-pro recalls 0.467, gpt-oss:120b 0.342, gemma4:31b 0.343.

Precision is high; the gap is under-prediction. Recall alone is insufficient to license a causal reading. Per-form precision on the LLM side (Appendix F, Table 11) is in fact high on the conditional cell (0.68–0.89): when the LLM does assign the conditional label, it is usually correct. The conditional gap is therefore a pure under-prediction (low recall, high precision), and binary positive recall on conditionals is 0.84–0.95 for every API-served LLM (Appendix E). The mass missing from the conditional cell is redistributed to adjacent

Tier 1 cells, most commonly modal/deontic and optative, where the rows still trigger binary positives but pick up the wrong form label. The deployment consequence is narrower than a simple binary-filter story: a binary suggestion filter will not under-fire on conditional feedback, but any pipeline that routes or reports feedback by Tier 1 form (e.g., a “polite-conditional suggestions” dashboard) will undercount that form by roughly a factor of two.

Encoder per-form precision partially resolves the lexical-anchoring question. The lexical-anchoring interpretation (that encoders resolve *would/if* toward the directive reading because they imitate the silver training distribution) predicts not only high conditional recall but also depressed conditional precision from over-firing. RoBERTa-large per-form precision and F1 (Appendix G, Table 12) show the encoder reaches both high conditional recall (0.852) and substantial conditional precision (0.79), so the encoder advantage cannot be a pure over-firing artifact. The residual precision–recall gap on the conditional cell is nonetheless wider than on optatives or imperatives, leaving some inherited lexical anchoring plausible; we treat this as partially, not fully, resolved. Opus 4.7’s own per-form profile on the gold overlap (Appendix H) shows that the encoder’s conditional advantage is inherited from the silver-label distribution rather than from encoder architecture.

Scale. Within the instruction-tuned LLM rows, conditional recall does not correlate cleanly with parameter count: the lowest are ministral-3:8b (0.080) and ministral-3:14b (0.234); gpt-oss:120b (0.342), gemma4:31b (0.343), and nemotron-3-nano:30b (0.361) cluster in the middle; deepseek-v4-pro (0.467) is highest. We therefore do not claim that the conditional gap is a scaling phenomenon. What does hold uniformly is that every zero-shot LLM row trails every encoder row by ≥ 30 percentage points on conditional recall: the family-level gap is robust, the within-family scale ordering is not. A second, weaker pattern is that interrogative recall is the encoders’ worst Tier 1 cell (0.67–0.75) while the four strongest instruction-tuned LLMs handle interrogatives better (0.82–0.90; nemotron 0.78, ministral 0.37–0.53). Interrogatives (n=258) all come from the rare-taxonomy oversample: the two families’ per-form biases differ mainly on the conditional and interrogative cells. Extended form-stratified analysis is in Appendix L.

6.3 Matched Supervision, Not Architecture

Encoders were trained on this taxonomy and LLMs were scored zero-shot. We test that reading two ways.

Prompt grid. Same 500 gold conditionals (plus 1,058 non-conditional positives, used only for precision), deterministic decoding, only the prompt changes (Table 6; prompts in Appendix P). P0 is the baseline Task B prompt used for Table 4. Paraphrase (P1) $\approx$ P0. Few-shot k=2, 4 stay near baseline (0.31–0.45); k=8 helps only partly (0.49–0.74), still below RoBERTa-large 0.852. Stating the operational definition of a conditional (P2, still zero-shot) reaches 0.80–0.87. Precision stays 0.83–0.95, so P2 is not over-prediction. The gap tracks definition/convention alignment, not wording. Standard zero-shot prompts still miss more than half of conditional feedback. CIs: Appendix R.

LoRA. Qwen2.5-7B-Instruct QLoRA on the same silver positives (gold held out) reaches **0.904** conditional recall on all 2,009 gold conditionals and **0.898** on the frozen 500, with precision 0.828 on the full set, matching or beating the encoder. Zero-shot Qwen2.5-7B reaches 0.144 conditional recall. Supervision, not family, is the difference. Training protocol: Appendix S.

We do not claim zero-shot LLMs cannot represent conditionals. We claim that the default prompts used in our (and typical) deployment under-assign that form, while silver-supervised models do not.

6.4 The Gap Is Not Borderline and Not One Domain

On 1,837 *non-borderline* gold conditionals, LoRA remains 0.910 and zero-shot LLMs remain 0.34–0.47 (Table 7). Conditionals are 21–29% of positives in every macro-domain; LoRA recall is 0.85–0.91 in all four, while zero-shot LLMs stay 0.26–0.53 (Appendix T). Across 59 industries with ≥ 20 conditionals, median LoRA recall is 0.913 vs. 0.462 / 0.340 / 0.333 for deepseek-v4-pro / gpt-oss:120b / gemma4:31b.

6.5 Routing

Among gold conditionals, a form router is the predicted Tier 1 label. Two-stage routing additionally requires Task A to fire (Table 7). Zero-shot models detect the rows (0.84–0.96) then send them to the wrong queue: deepseek delivers about half as many true conditionals as LoRA to the conditional queue;

Model	Imperative	Modal	Optative	Conditional	Interrog.	Compar.†
<i>Encoder fine-tunes (3 seeds; full 7,500 positives)</i>						
BERT-base-uncased	0.782	0.857	0.900	0.859	0.674	0.375
RoBERTa-base	0.804	0.837	0.891	0.845	0.748	0.625
DeBERTa-base	0.845	0.850	0.897	0.867	0.733	0.750
RoBERTa-large	0.821	0.842	0.896	0.852	0.752	0.625
<i>Instruction-tuned LLMs, zero-shot (full 7,500 positives)</i>						
deepseek-v4-pro	0.844	0.879	0.927	0.467	0.899	1.000
deepseek-v4-flash	0.776	0.883	0.942	0.431	0.868	0.750
gpt-oss:120b	0.869	0.901	0.941	0.342	0.857	0.375
gemma4:31b	0.825	0.919	0.949	0.343	0.822	0.625
nemotron-3-nano:30b	0.867	0.834	0.908	0.361	0.775	0.750
ministral-3:14b	0.776	0.330	0.812	0.234	0.531	0.375
ministral-3:8b	0.694	0.522	0.641	0.080	0.368	0.625
<i>Matched-supervision LoRA (full 7,500 positives)</i>						
Qwen2.5-7B LoRA	0.870	0.897	0.938	0.904	0.058*	0.000*

Table 4: Per-form recall on the 7,500 gold positives. The cell of interest is the conditional column: encoders cluster at ~ 0.85 and zero-shot LLMs fall by half. This is the gap that the binary score in Table 3 averages out. 1,000-resample bootstrap 95% CIs for the conditional column are in Table 5. Per-form precision and F1 for the LLM rows are in Appendix F. † Comparative ($n = 8$) is descriptive only; see Appendix N. *LoRA interrogative/comparative collapse is expected: gold absorbed rare forms, leaving 9 train interrogatives and 0 comparatives (Appendix S).

Model	Cond. recall	95% CI
DeBERTa-base	0.867	[0.852, 0.882]
BERT-base-uncased	0.859	[0.843, 0.875]
RoBERTa-large	0.852	[0.836, 0.868]
RoBERTa-base	0.845	[0.829, 0.861]
Qwen2.5-7B LoRA	0.904	[0.891, 0.916]
deepseek-v4-pro	0.467	[0.445, 0.489]
deepseek-v4-flash	0.431	[0.410, 0.453]
nemotron-3-nano:30b	0.361	[0.340, 0.382]
gemma4:31b	0.343	[0.322, 0.364]
gpt-oss:120b	0.342	[0.321, 0.363]
ministral-3:14b	0.234	[0.215, 0.253]
ministral-3:8b	0.080	[0.068, 0.092]

Table 5: Conditional-cell recall with 1,000-resample bootstrap 95% CIs on the $n = 2,009$ conditional positives in the gold. Every encoder CI (and the LoRA CI) lies entirely above every zero-shot LLM CI; the form-family gap is non-overlapping at the 95% level for the conditional cell.

gpt-oss and gemma about two fifths. Misses go to modal and optative (Appendix U). A binary suggestion filter will not drop these reviews; a form-aware dashboard will undercount them by a factor of two to three.

A two-stage schematic of the instrument:

Stage 1 (Task A): is this a suggestion? $\rightarrow$
Stage 2 (Task B): which form?
The encoder–LLM disagreement is almost entirely in stage two, on conditionals. Prompt P2 recovers the cell without weight updates; LoRA recovers it without changing the prompt. Both are convention alignment. Rewording the prompt (P1) and adding a few in-context examples are not.

Condition	deepseek-pro	gpt-oss:120b	gemma4:31b
P0 zero-shot	.434	.314	.326
P1 paraphrase	.466	.322	.326
P2 definition	.866	.820	.804
few-shot $k=2$	.434	.334	.314
few-shot $k=4$	.454	.414	.332
few-shot $k=8$	.742	.492	.512
RoBERTa-large (full 2,009): 0.852			
Qwen2.5-7B LoRA (full / 500): 0.904 / 0.898			

Table 6: Conditional recall on a frozen 500-row gold slice. Only the prompt changes across conditions: P0 = baseline Task B prompt, P1 = paraphrase, P2 = expanded conditional definition plus priority-order line, k = number of few-shot exemplars (Appendix P). Decoding is deterministic. CIs: Appendix R.

6.6 Form $\times$ Modifier Dependency

A χ^2 test of independence between Tier 1 form (six classes) and Tier 2 modifier presence (five-state collapse) on the 7,500 gold positives yields $\chi^2(20) = 1304.3$, $p < 3.5 \times 10^{-264}$, Cramér’s $V = 0.21$, a small-to-moderate effect. The substantive signal is in the standardized residuals, dominated by two cells. Conditional $\times$ hedged has residual +15.1: conditional suggestions are dramatically more likely to be hedged than expected under independence. Direct imperative $\times$ frustrated has residual +17.7: direct imperatives are more likely to be frustrated. This is what politeness theory (Brown and Levinson, 1987) would predict: hedging co-occurs with high-politeness conditional

Model	Form-router	Detect	Two-stage	Non-BL
RoBERTa-large	.852	—	—	—
Qwen2.5-7B LoRA	.904	1.00*	.904	.910
deepseek-v4-pro	.467	.881	.428	.471
gpt-oss:120b	.342	.955	.332	.342
gemma4:31b	.342	.841	.285	.342

Table 7: Routing of 2,009 gold conditionals. Non-BL = non-borderline slice ($n=1,837$). *LoRA scored on positives only; encoder two-stage and non-borderline cells cannot be recomputed (no per-row encoder predictions in the release).

forms; frustration with bald-on-record imperatives. The methodological consequence is that per-form and per-modifier metrics are not slice-wise independent. A model that systematically misses hedged language will appear to fail on conditionals. Slice metrics should be reported jointly (form $\times$ modifier) rather than marginally.

7 Discussion

Reframing suggestion mining. The empirical contribution of this paper is not a new state-of-the-art on binary detection. It is the demonstration that suggestion mining, posed as binary classification, hides a structural error mode. Form-stratified slicing turns “model X beats model Y by 4 macro-F1 points” into a linguistically interpretable statement: model X reliably identifies conditional utterances as suggestions at the binary level but mis-assigns roughly half of them at the Tier 1 form level. The statement does not require a new model. It requires labeling enough data, at enough granularity, in enough domains, to make the pattern visible. The running example is the whole finding: same text, same gold span, three zero-shot queues, one supervised queue.

Three findings, restated. (i) Per-form recall is non-uniform for every model family we benchmark; the rank order of forms is itself family-specific. Interrogatives are encoders’ weakest cell; conditionals are zero-shot LLMs’ weakest cell. (ii) LLMs lead at the binary level, while the encoder advantage at the form level is a single-cell phenomenon (conditional). (iii) Tier 1 form and Tier 2 modifiers show a modest global association concentrated in two interpretable cells (hedged conditionals, frustrated imperatives); slice metrics should be reported jointly. We do not treat sarcastic/backhanded F1 as a hard leaderboard (Appendix I).

Deployment. PRAXIS supports triage, not un-

conditional filtering. A zero-shot LLM used purely as a binary suggestion pre-filter will retain hedged conditional feedback at near-imperative recall. The risk appears one tier downstream: any pipeline that routes or prioritizes feedback by Tier 1 form will see conditional throughput halved by form-classification error, and any filter calibrated at the form level should be calibrated against gold (not silver) per form. Use these models for triage, not as filters that hide non-preferred forms.

8 Conclusion

PRAXIS is a measurement instrument for pragmatic-form bias in English review suggestion mining. It releases a 174,860-review silver corpus, a 15,000-row human-labeled gold evaluation set, a three-tier pragmatic typology grounded in speech-act theory, and a benchmark of four model families scored against the gold with bootstrap confidence intervals. The substantive finding is linguistic, not competitive: encoders and zero-shot instruction-tuned LLMs diverge sharply on the conditional cell (encoders higher) and, on a small oversampled set, the interrogative cell (LLMs higher), while broadly agreeing on the rest, with the conditional gap being the dominant cell. A same-corpus LoRA closes that cell, showing the gap is supervision, not architecture. The data, code (including the LoRA training script, hyperparameters and train/validation split), and per-row LLM predictions are released under CC-BY-4.0 (annotations) and MIT (code).

Limitations

We list substantive limitations rather than methodological hedges.

Annotation and reliability. The 15,000-row gold is single-annotated outside the 1,000-row reliability subset, and the subset itself includes the bulk annotator as one of three panelists. The reliability case for the released gold is therefore explicitly scoped: it rests on (i) three-annotator panel IAA on the subset and (ii) bulk-vs-panel hold-one-out agreement on the same subset, both reported in Table 2. We do not claim that the 14,000 rows outside the subset have been multiply validated; those rows are single-annotated, and the bulk-vs-panel statistic generalizes from the stratified subset to the bulk only under the assumption that the bulk annotator’s per-row agreement with consensus is stationary across the gold. Bulk-vs-panel agreement is within 3–5 κ points of symmetric IAA on

every task (mean gap 0.04), which materially constrains the likely magnitude of annotator-specific bias on those 14,000 rows but does not directly measure it. We release per-panelist subset labels separately so downstream work can re-annotate or train on a multi-annotator label set. We did not run a full-gold tiebreak or multi-label sensitivity study ; because the panelists agree least on the conditional (Appendix J), part of the encoder–LLM gap may reflect annotator convention, and that study remains future work.

Scope and generalization. All gold and silver labels are produced under English-language guidelines. Directive morphology in Hindi, Japanese, Korean, and other languages with rich politeness or speech-level systems is not addressed. The three-annotator panel was recruited from a single institution. We did not add a 2025/26 post-cutoff review set or an out-of-genre email replication of Lampert et al. (2010). Pretraining contamination of public Amazon text is a structural limit of any public-review benchmark; we argue in §5 that per-form gaps are robust to contamination, since contamination would inflate accuracy uniformly across forms rather than selectively on conditionals, but we did not test on reviews collected after model cutoffs. Leaderboard cells where an instruction-tuned LLM outperforms a fine-tuned encoder at the binary level should be read with this caveat.

Evaluation and comparison. PRAXIS-silver is suitable for training, exploratory analysis, and data-scale studies. It is not an evaluation reference. We discourage any use of the 174K silver labels as a reference standard. The 7,500/7,500 gold balance is a deliberate evaluation choice and not the prevalence one would observe in a deployed review queue; the released per-row weight `eval_weight_natural_binary` recovers natural prevalence for binary metrics. Encoder form predictions are not released per row, so encoder slices in Table 7 cannot be recomputed from the public logs. LoRA interrogative/comparative scores are not comparable (almost no silver train mass after the gold hold-out). We did not run chain-of-thought scoring or a conditional-heavy few-shot condition beyond $k=8$ on the frozen 500; those remain open tests of whether richer prompting can close the cell without weight updates.

Tier 2 (sarcastic and backhanded) reliability. Two of the four Tier 2 modifiers, sarcastic (panel

$\kappa = 0.54$, 1.2% prevalence) and backhanded (panel $\kappa = 0.42$, bulk-vs-panel $\kappa = 0.37$), fall short of substantial agreement. We release them as soft-label / diagnostic targets rather than as hard-label benchmark cells; only frustrated and hedged are positioned for hard-label F1 reporting.

Ethical Considerations

Artifacts and licensing. Source corpora and the licenses under which we redistribute text are documented in a per-source matrix in Appendix A. Review text is redistributable under its respective dataset license. PRAXIS annotations are released under CC-BY-4.0 and code under MIT.

Consent of original review authors. Original review authors on Yelp and Amazon did not consent to the specific use of their reviews as training or evaluation data for a suggestion-mining benchmark. This is standard for benchmarks built on public review platforms and is mitigated by (i) restricting to platform-level (review-ID) identifiers, (ii) omitting all user-identifying metadata, and (iii) using only de-identified text the original platforms have themselves made publicly accessible. We record the absence of specific consent explicitly rather than treat the public-data status as substitute consent.

Annotator compensation and recruitment. The panel comprised three annotators, all near-native English speakers based in India. They were members of the research group but were not involved in the project. They were paid according to institute norms, with no speed or label-value bonuses. Before main annotation, all three completed a two-hour training session followed by a 50-item practice calibration round against the guidelines (Appendix B). Annotation work used only public, already-de-identified review text, and the institutional IRB issued an exempt determination under the public-data / de-identified-text exemption category. Annotators consented to the use of their labels via standard employment agreement.

Who benefits / who may be harmed. Form- and affect-based triage can down-rank indirect, hedged, or frustrated speakers, often the same people who do not write bald imperatives. Dual-use profiling (scoring customers on sarcasm or “backhanded” affect) is a risk; sarcastic/backhanded labels must not be read as a sincerity detector. Do not use PRAXIS labels in high-stakes settings without ethical review. Future multilingual versions should treat represen-

tation as first-order (Hovy and Spruit, 2016).

Compute. Encoder fine-tunes were trained on a single workstation with two A6000 GPUs. Total compute for the four-encoder, three-seed sweep was approximately 60 GPU-hours. The LoRA run used approximately 3 GPU-hours on one RTX A4500.

Anonymization. All source-corpus identifiers used in the release are platform-level (review IDs), not user-identifying.

Acknowledgments

This work was supported by BirdEye Inc., Palo Alto, California, USA. The authors gratefully acknowledge BirdEye Inc. for its support of this research.

References

- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.
- John Langshaw Austin. 1975. *How to do things with words*, volume 5. Harvard university press.
- Andrew M Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, et al. 2026. Measuring what matters: Construct validity in large language model benchmarks. *Advances in Neural Information Processing Systems*, 38.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.
- Caroline Brun and Caroline Hagège. 2013. [Suggestion mining: Detecting suggestions for improvement in users’ comments](#). *Res. Comput. Sci.*, 70:199–209.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Dirk Hovy and Shannon L. Spruit. 2016. [The social impact of natural language processing](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: what participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- Andrew Lampert, Robert Dale, and Cecile Paris. 2010. [Detecting emails containing requests for action](#). In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 984–992, Los Angeles, California. Association for Computational Linguistics.
- J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). Cite arxiv:1907.11692.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Julian McAuley and Jure Leskovec. 2013. [Hidden factors and hidden topics: understanding rating dimensions with review text](#). *Proceedings of the 7th ACM conference on Recommender systems*.

- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the Association for Computational Linguistics*, 10:92–110.
- Sapna Negi and Paul Buitelaar. 2015. [Towards the extraction of customer-to-customer suggestions from reviews](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2159–2167, Lisbon, Portugal. Association for Computational Linguistics.
- Sapna Negi, Tobias Daudert, and Paul Buitelaar. 2019. [SemEval-2019 task 9: Suggestion mining from online reviews and forums](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 877–887, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Sapna Negi, M. de Rijke, and Paul Buitelaar. 2018. [Open domain suggestion mining: Problem definition and datasets](#). *ArXiv*, abs/1806.02179.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.
- Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. [Automated annotation with generative ai requires validation](#). *Preprint*, arXiv:2306.00176.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. [Linguistically debatable or just plain wrong?](#) In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. [SemEval-2016 task 5: Aspect based sentiment analysis](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.
- Qwen Team. Qwen2.5 technical report.
- J. Ramanand, Krishna Bhavsar, and Niranjan Pedaneekar. 2010. [Wishful thinking - finding suggestions and ‘buy’ wishes from product reviews](#). In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, Los Angeles, CA. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.
- John R. Searle. 1975. Indirect speech acts. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 59–82. Academic Press, New York.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Computational Linguistics*, 26(3):339–374.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. [Want to reduce labeling cost? GPT-3 can help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Per-Source Redistribution Matrix

Table 8 lists each contributing source, its license, and the form in which PRAXIS redistributes the row.

Source	License	Distribution
Yelp Open Dataset	Yelp academic	full text + labels
Amazon (McAuley/Ni)	McAuley research	full text + labels

Table 8: Distribution form per source. The Amazon source is the McAuley/Ni Amazon corpus (McAuley and Leskovec, 2013; Ni et al., 2019). The gold and all reported analyses use these two sources only.

B Annotation Guidelines

The full annotation guidelines used by the LLM annotator and the human annotators are released in machine-readable form with the dataset. The headline rules are reproduced here.

Binary suggestion. A suggestion is a business-directed utterance specifying a concrete operational change. The directive may be expressed in any Tier 1 form. Exclusions: opinion only; complaint without directive; customer-to-customer advice; pure product description; off-topic remarks.

Tier 1 priority order. Direct imperative > modal/deontic > optative > conditional > interrogative > comparative. If a span contains both an imperative and a conditional, label it as imperative.

Tier 2 multi-label. Each modifier is independently assessed. A single suggestion may carry zero, one, two, or more modifiers. Backhanded requires both a positive evaluation and a directed critique in the same span.

Tier 3 ordinal. Vague: no concrete operational target; specific: a single concrete operational target; very specific: a concrete operational target with explicit parameters (item names, prices, quantities, timing).

Distinctions (verbatim from the annotator handbook). *Suggestion vs. general complaint.* “The food was bad” → negative; “The food was bad and the kitchen needs better recipes” → positive (the second clause is operational). *Suggestion vs. customer-to-customer advice.* “Skip the chicken” → negative under our business-directed frame. *Suggestion vs. feature description.* “The earbuds connected easily” → negative; “The earbuds should ship with a USB-C adapter” → positive.

Tier 2 modifier guidelines. *Frustrated/emotional:* profanity, exclamation, all-caps, emotional language directed at the business. *Sarcastic:* surface sentiment opposite to intended directive (“Great job training your staff to ignore customers”). *Backhanded:* praise then suggest pattern. *Hedged:* *maybe, perhaps, possibly, sort of, kind of, I think*, or similar tentative markers.

C Bootstrap Confidence Intervals

All cells in the main tables are reported with 1,000-resample bootstrap 95% confidence intervals. For encoder models, the interval is a pooled bootstrap over rows and seeds jointly: at each of 1,000 resample iterations, we draw n gold rows with replacement from the evaluation slice and, independently for each drawn row, a seed uniformly at random from the three trained seeds, then score the corresponding seed’s prediction on that drawn row. This yields a single bootstrap distribution that absorbs both row-level sampling variance and across-seed prediction variance into one 2.5–97.5 percentile interval per metric and per slice. Full per-cell intervals are released alongside the data.

LoRA and prompt-grid CIs use the same n , seed 42, and within-class resampling of the 0/1 hit indicator.

D Borderline-Positive Rates by Tier 1 Form

Table 9 reports the borderline-flag rate stratified by Tier 1 form on the 7,500 gold positives. Borderline-rationale text was post-processed into priority-assigned mutually exclusive bins: 890 wish/optative phrasing, 95 vague / no concrete operational target, 113 personal-convenience or out-of-scope, 39 industry mislabel, 8 sarcasm-related, 172 other. The bins sum to 1,317. Review-length statistics on the full 15,000 gold rows: review text mean 132.5 tokens (median 91); suggestion span mean 13.2 tokens (median 11). The smallest industry contributes 66 rows; the largest, 371.

Tier 1 form	Borderline	Total	Rate
Direct imperative	29	647	4.5%
Interrogative	18	258	7.0%
Modal/deontic	158	1,997	7.9%
Conditional	172	2,009	8.6%
Optative	935	2,581	36.2%
Comparative	5	8	62.5%
All positives	1,317	7,500	17.6%

Table 9: Borderline-positive rate stratified by Tier 1 form.

E Binary Positive Recall by Tier 1 Form

Table 10 decomposes binary positive recall by gold Tier 1 form for the seven complete-coverage instruction-tuned LLMs, on the 7,500-row positive slice. Conditional binary recall is high for every model (0.84–0.95), confirming that the low Tier 1 conditional recall in Table 4 is a form-classification failure, not a detection failure.

F Per-Form Precision and F1 for Instruction-Tuned LLMs

Conditional precision is uniformly high (0.68–0.89); low conditional F1 is driven by low recall, not by false positives on the conditional class.

G Encoder Per-Form Precision and F1

Table 12 reports per-form precision, recall, and F1 for RoBERTa-large (3 seeds) on the 7,500-row human gold positive slice. Comparative ($n = 8$;

Model	Imperative	Modal	Optative	Conditional	Interrogative	Compar.†
gpt-oss:120b	0.918	0.912	0.942	0.955	0.911	0.375
nemotron-3-nano:30b	0.901	0.902	0.879	0.935	0.876	0.500
deepseek-v4-pro	0.861	0.859	0.885	0.881	0.903	0.375
ministral-3:14b	0.847	0.901	0.830	0.870	0.899	0.875
deepseek-v4-flash	0.819	0.854	0.844	0.880	0.853	0.625
ministral-3:8b	0.821	0.913	0.784	0.885	0.864	0.875
gemma4:31b	0.764	0.811	0.768	0.841	0.798	0.375

Table 10: Binary positive recall (Task A) decomposed by gold Tier 1 form. † Comparative column is $n = 8$, descriptive only.

Model	Precision						F1					
	Imp.	Modal	Opt.	Cond.	Int.	Cmp.†	Imp.	Modal	Opt.	Cond.	Int.	Cmp.†
deepseek-v4-pro	0.708	0.730	0.849	0.827	0.746	0.222	0.770	0.798	0.886	0.597	0.816	0.235
deepseek-v4-flash	0.802	0.800	0.763	0.865	0.754	0.061	0.789	0.840	0.843	0.575	0.807	0.111
gpt-oss:120b	0.781	0.809	0.700	0.883	0.775	0.131	0.822	0.853	0.803	0.493	0.814	0.194
gemma4:31b	0.698	0.680	0.830	0.891	0.806	0.094	0.756	0.782	0.886	0.495	0.814	0.164
nemotron-3-nano:30b	0.529	0.726	0.809	0.865	0.784	0.039	0.657	0.776	0.856	0.510	0.780	0.077
ministral-3:14b	0.187	0.653	0.836	0.707	0.559	0.008	0.302	0.438	0.824	0.352	0.545	0.016
ministral-3:8b	0.218	0.440	0.901	0.679	0.617	0.006	0.332	0.478	0.749	0.143	0.461	0.013
Qwen2.5-7B LoRA	0.837	0.893	0.928	0.828	0.938	—	0.853	0.895	0.933	0.864	0.110	—

Table 11: Per-form precision and F1 on the 7,500 gold positives. Conditional precision is uniformly high; low conditional F1 is low recall. † Comparative column is descriptive only ($n = 8$). LoRA comparative is undefined (zero predicted and zero train mass).

see Appendix N) is descriptive only and is omitted from the macro-average.

Encoder per-form precision and F1 partially clarify the conditional-recall finding. RoBERTa-large achieves not only high conditional recall (0.852) but also substantial conditional precision (0.79), yielding an F1 of 0.82 on the conditional cell. The encoder advantage on conditionals therefore cannot be explained purely by indiscriminate lexical over-firing on *would/if* constructions. However, the gap between conditional precision (0.79) and recall (0.852) remains larger than for optatives or imperatives, suggesting that some degree of lexical anchoring inherited from the silver training distribution may still contribute to the encoder-side advantage. Interrogatives remain the encoder family’s weakest Tier 1 cell (F1 0.73), consistent with the broader finding that encoder and zero-shot-LLM blind spots are partly inverse (conditional vs. interrogative).

The encoder advantage on conditional forms should not be interpreted as evidence that encoders possess stronger general pragmatic competence than instruction-tuned LLMs. Rather, the result suggests that supervised discriminative models trained on large task-specific corpora can become highly calibrated to narrow pragmatic tax-

onomies, particularly when those taxonomies align with stable lexical constructions such as conditional directives (“would”, “if”, “would be great if”). Instruction-tuned LLMs, despite stronger general language understanding, appear to apply a broader pragmatic interpretation that frequently routes these utterances into adjacent directive categories rather than the conditional class itself.

Tier 1 form	Precision	Recall	F1
Direct imperative	0.88	0.821	0.85
Modal/deontic	0.86	0.842	0.85
Optative	0.92	0.896	0.91
Conditional	0.79	0.852	0.82
Interrogative	0.71	0.752	0.73
Comparative†	—	0.625	—
Macro-average	0.83	0.833	0.83

Table 12: RoBERTa-large (3 seeds) Tier 1 communicative-form precision, recall, and F1 on the 7,500-row human gold positive slice. The macro-average is computed over the five non-comparative forms. † Comparative ($n = 8$) is descriptive only.

H Opus 4.7 Per-Form Performance on Gold

Opus 4.7 also labeled the gold texts in a separate pass, using the same Tier 1 taxonomy and the

positive-conditioned prompt used for the zero-shot LLMs in Table 4 (it is not the silver pipeline’s own output, which tier-labels only rows Opus judged to be suggestions). Table 13 reports its per-form precision, recall and F1 against the human gold on the same 7,500-row positive slice used for the encoder and LLM tables.

The Opus per-form profile is the missing diagnostic that connects the silver pipeline to the encoder finding. Three observations are decision-relevant.

Opus reproduces the conditional cell; zero-shot LLMs do not. Opus’s conditional recall (0.83) is within 0.02 of RoBERTa-large’s (0.852) and is roughly twice the recall of the strongest non-annotator LLM (deepseek-v4-pro at 0.467, Table 4). The encoder’s high conditional recall is therefore not an artifact of any property unique to fine-tuning architecture; it is consistent with being inherited from the silver-label convention that Opus produced. The conditional gap in Table 4 between encoders and other instruction-tuned LLMs is consistent with a gap between models trained on Opus’s conditional convention and models that were not; the P2 result (§6.3) shows the convention can also be supplied in-context. The LoRA result (§6.3) is the same mechanism with a different backbone: Qwen2.5-7B, given those labels, also reproduces the cell.

The lexical-anchoring story is partially, not fully, mediated by Opus. Encoder conditional precision (0.79) and Opus’s conditional precision (0.80) are essentially identical, while encoder conditional recall (0.852) is slightly above Opus’s (0.83). The qualitative finding from Appendix L (some lexical anchoring inherited from silver) is preserved; the quantitative bound is that the encoder’s conditional F1 (0.82) is close to Opus’s (0.81), and its five-form macro-F1 (0.83) is slightly above Opus’s (0.81). The encoder reproduces the silver convention; it does not diverge from it.

Partly Inverse blind spots are also inherited. Opus’s interrogative F1 (0.76) is in the same range as the encoders’ (0.73), and both are somewhat below the five strongest instruction-tuned LLMs (F1 0.78–0.82, Table 11). The encoder-vs-LLM partly inverse-blind-spot pattern therefore plausibly also originates upstream of fine-tuning, in the silver-label distribution itself, rather than in any property of the encoder architecture. A consequence for downstream calibration: any pipeline that wants to redistribute mass between conditional and interro-

ative cells at evaluation time may need to target the silver-label generator, not the encoder.

Tier 1 form	Precision	Recall	F1
Direct imperative	0.81	0.79	0.80
Modal/deontic	0.78	0.81	0.79
Optative	0.88	0.86	0.87
Conditional	0.80	0.83	0.81
Interrogative	0.73	0.79	0.76
Comparative†	—	0.500	—
Macro-average	0.80	0.82	0.81

Table 13: Claude Opus 4.7 Tier 1 per-form precision, recall, and F1 on the 7,500-row human gold positive slice. Macro-average is computed over the five non-comparative forms. † Comparative ($n = 8$) is descriptive only. Opus labels are from a separate positive-conditioned pass on the gold texts, not from the silver-pipeline run.

I Tasks C and D: Tier 2 Modifiers and Tier 3 Specificity

Tasks C and D are reported here rather than as main-paper leaderboard cells. They remain part of the released benchmark.

Task C (Tier 2). Table 14 confirms that the four modifiers do not present a uniform difficulty, and they support a reliability-based split of the Tier 2 schema into a hard-label benchmark subset and a soft-label diagnostic subset. Frustrated (human IAA $\kappa = 0.72$, bulk-vs-panel 0.68; Table 2) is recoverable at F1 0.47–0.60 across instruction-tuned LLMs, driven by the high density of affect-marked lexical material in frustrated suggestions. Hedged ($\kappa = 0.66$, bulk-vs-panel 0.61) is moderate at F1 0.28–0.49. Sarcastic ($\kappa = 0.54$, bulk-vs-panel 0.49) and backhanded ($\kappa = 0.42$, bulk-vs-panel 0.37, the only Tier 2 cell where bulk-vs-panel falls below Opus-vs-gold) sit at the low end of substantial agreement and below; combined with sarcastic’s low natural prevalence (1.2% of positives), this puts both labels in a regime where hard-label F1 reporting is high variance and difficult to interpret.

We therefore reposition Tier 2 explicitly: frustrated and hedged are released as hard-label benchmark targets; sarcastic and backhanded are released as soft-label / diagnostic targets, appropriate for analyses such as confidence calibration, disagreement modeling, and per-row probability outputs but not for a hard-label leaderboard. The corresponding model F1 on these two labels (sarcastic

0.24–0.43; backhanded 0.06–0.27) tracks the per-label IAA ceiling closely.

We do not present the low backhanded F1 as a finding that validates the label. Backhanded requires integrating two pragmatically opposed clauses (one praise, one critique) and resolving them to a single modifier tag, and both humans (panel $\kappa = 0.42$, bulk-vs-panel 0.37) and models (0.06–0.27 F1) fall short of substantial agreement on it. The honest reading is that backhanded is at the edge of what can be reliably labeled in this taxonomy, and we release it as a soft-label diagnostic rather than a hard-label benchmark cell. Sarcastic is in the same category for a different reason: agreement is in the moderate range and natural prevalence is 1.2%, so per-label F1 is high variance regardless of true model behavior. Of the four Tier 2 modifiers, frustrated and hedged are the two we recommend treating as hard-label benchmark targets; sarcastic and backhanded are released so that downstream work can build soft-label or diagnostic treatments on them, not to compete on hard-label F1. PRAXIS is the first benchmark to label all four at scale, and surfacing the reliability gradient is one of the contributions; conflating the two classes of label would obscure it.

Task D (Tier 3). Quadratic-weighted κ for the three-point ordinal specificity task (Table 15) ranges from 0.523 (deepseek-v4-pro) down to 0.118 (ministral-3:14b). The bounded human QWK on the same task (§3.5, QWK = 0.67) sets an empirical ceiling; none of the models clears it. Tier 3 is an ordinal label that asks the annotator to judge how operationally actionable the suggestion is, a judgment that resists strict criteria. The human-and-model-both-moderate pattern is itself the finding.

Form $\times$ modifier independence on the 7,500 gold positives is reported in the main paper (§6.6): $\chi^2(20) = 1304.3$, $p < 3.5 \times 10^{-264}$, $V = 0.21$; dominant residuals: Conditional $\times$ Hedged +15.1, Imperative $\times$ Frustrated +17.7.

J Extended Inter-Annotator Agreement Interpretation

The agreement profile in Table 2 reflects the linguistic structure of the task rather than mere annotation consistency. Binary identification is strong ($\kappa = 0.80$), indicating that the operational definition of a suggestion as a business-directed actionable change is stable and reproducible. Tier 1

communicative form is likewise substantial (Fleiss $\kappa = 0.77$, on rows all three annotators mark as suggestions): imperatives, optatives, interrogatives and modal/deontic directives are reliably distinguishable despite occasional surface overlap; the conditional is markedly less stable (the bulk annotator’s one-vs-rest κ with the other two panelists is 0.40 and 0.39, against 0.61–0.86 on every other form and 0.86 between the other two), and the comparative (8 gold rows) is too rare to assess. Agreement decreases through the higher pragmatic tiers, which we interpret as an empirical property of the phenomenon rather than a methodological weakness. Tier 3 specificity is moderate-to-substantial (QWK = 0.67), and the Tier 2 modifiers are the most challenging component: frustrated ($\kappa = 0.72$) and hedged ($\kappa = 0.66$) remain substantially recoverable, while sarcastic ($\kappa = 0.54$) and backhanded ($\kappa = 0.42$) are lower, consistent with prior pragmatics literature on the inherent ambiguity of indirect affective intent and praise-then-critique constructions. We therefore read low agreement on these labels not as noise but as evidence that affective modifiers of directives are a genuinely difficult phenomenon for humans and models alike.

The lone exception is the backhanded modifier, where the bulk annotator’s agreement with consensus ($\kappa = 0.37$) is marginally below Opus-vs-gold ($\kappa = 0.39$); at this label’s low support the two are statistically indistinguishable, and the result is consistent with backhanded being the hardest label in the taxonomy for every annotator, human or model. This is the basis for our recommendation that downstream work requiring hard labels restrict to higher-agreement modifiers or adopt a soft-label treatment of backhanded.

Finally, Opus-vs-gold agreement tracks at or below both human numbers on every comparable task except backhanded, most sharply on the subtle Tier 2 modifiers, confirming that the released scheme captures recoverable pragmatic structure while the LLM annotator remains a weaker reproducer of it than a trained human, which is why all model evaluation in this paper is restricted to the human gold. On the strength of these numbers we release the 174,860 rows as validated silver, and we restrict all model evaluation to the 15,000-row human gold. We release the per-panelist subset labels separately so that downstream work can re-train or re-evaluate on a multi-annotator label set if preferred.

Model	Macro-F1	Frustrated	Sarcastic	Backhanded	Hedged
gemma4:31b	0.440	0.603	0.434	0.266	0.458
deepseek-v4-flash	0.403	0.545	0.389	0.241	0.438
gpt-oss:120b	0.398	0.474	0.388	0.254	0.475
deepseek-v4-pro	0.369	0.500	0.295	0.221	0.461
minstral-3:14b	0.340	0.507	0.355	0.191	0.306
nemotron-3-nano:30b	0.329	0.516	0.244	0.063	0.491
minstral-3:8b	0.296	0.441	0.279	0.181	0.281

Table 14: Tier 2 multi-label modifier F1 on the 7,500 gold positives. Backhanded is the lowest cell for every model and matches the lowest per-label IAA (Table 2), reflecting a shared difficulty between humans and models rather than a model deficiency. Sarcastic and backhanded are diagnostic-only.

Model	QWK	Acc.	Macro-F1
deepseek-v4-pro	0.523	0.781	0.635
gemma4:31b	0.495	0.777	0.613
gpt-oss:120b	0.490	0.800	0.607
deepseek-v4-flash	0.476	0.778	0.567
nemotron-3-nano:30b	0.428	0.772	0.519
minstral-3:8b	0.227	0.627	0.402
minstral-3:14b	0.118	0.717	0.370

Table 15: Tier 3 ordinal specificity: QWK, accuracy, and macro-F1 on the 7,500 gold positives.

K Extended Dataset Statistics

Counts in §3.4 are reported on the gold positive slice ($n = 7,500$) and sum exactly. Balanced counts are on the full 7,500 positives as released; stratified-only counts are on the 6,693 rows in the `positive_stratified` sampling bucket, which excludes the 807-row rare-taxonomy oversample and is the closest approximation to natural form prevalence available without the silver-pool form labels (which are not bundled in this release). On the stratified-only column, three classes together account for 91.3% of positives: optative (36.8%), modal/deontic (27.3%), and conditional (27.2%). Direct imperatives are 8.7% on the stratified-only column and 8.6% on the balanced column, so the “imperatives are a minority” headline is robust to the oversample. Interrogatives (3.4%) and comparatives (0.1%) on the balanced column are entirely a product of the rare-taxonomy supplement (both are 0% in the stratified-only column) and should be read as “benchmark-included for non-zero support,” not as prevalence estimates.

Tier 3 (operational specificity) is dominated by the specific class (72.6%), with 21.1% vague and 6.3% very specific. Tier 2 modifiers are sparse: 65.7% of positives carry no modifier, 14.8% are frustrated, 13.1% hedged, 9.3% backhanded, and 1.2% sarcastic. The natural-prevalence rates of sarcastic and backhanded are low enough that Tier 2

results need to be read jointly with the IAA per-cell support.

Splits and sampling buckets. `gold_test` (15,000 rows, sealed for evaluation) is the only split scored in this paper. The silver-derived training pool used for encoder fine-tuning is the full silver-174K minus the 15K gold rows. Every gold row carries provenance and sampling-bucket flags; the sampling buckets are `positive_stratified` (6,693), `negative_progress_false_positive` (6,152), `negative_verified_pool_sampled` (1,348), and `positive_rare_taxonomy_included` (807). The `rare_taxonomy_included` bucket exists so that low-prevalence forms (comparatives, interrogatives) are not absent from the gold.

Class-balance caveat. The 50/50 positive/negative split in gold is a deliberate evaluation choice. Natural prevalence varies substantially across the four macro-domains: roughly 5.1% for product, 5.9% for digital/media, 26.8% for hybrid retail, and 51.2% for service (estimates from the silver-corpus pool). PRAXIS releases a column `eval_weight_natural_binary` on every gold row so that any binary metric can be re-weighted to natural prevalence at evaluation time.

L Extended Form-Stratified Analysis

Lexical anchors and the encoder side. Encoders peak on optatives (recall ~ 0.89 – 0.90) and conditionals (~ 0.85 – 0.87). Both classes carry strong lexical anchors (*wish/hope/would love; would/if/should be*), which is consistent with a model that has internalized the silver-label distribution and resolves these surface markers toward the directive reading. The smallest open LLMs (Qwen2.5 family) collapse on optatives, scoring under 0.30 on this class even when the same models clear 0.4 macro-F1 at the binary level.

Instruction-tuned LLMs peak on optatives (0.91–0.95 for the five strongest) and are strong on modal/deontic (0.83–0.92), but their recall on conditional suggestions falls sharply. The two strongest open-weights models (deepseek family) sit at the upper end of the LLM range and still trail every encoder we test by 38–44 percentage points on the conditional class.

Conditional recall does not track scale monotonically. Within the LLM rows of Table 4, conditional recall does not correlate cleanly with parameter count. The lowest conditional recalls are minstral-3:8b (0.080) and minstral-3:14b (0.234); gpt-oss:120b (0.342), gemma4:31b (0.343), and nemotron-3-nano:30b (0.361) cluster in the middle; deepseek-v4-pro (0.467) is highest. The within-family deepseek-v4-pro vs. flash gap is small (0.467 vs. 0.431). We therefore do not claim that the conditional gap is a scaling phenomenon within the LLM family. What does hold uniformly is that every zero-shot LLM row trails every encoder row by ≥ 30 percentage points on conditional recall: the family-level gap is robust; the within-family scale ordering is not.

Encoders and instruction-tuned LLMs have partly inverse blind spots. A second, weaker pattern is that interrogative recall is the encoders’ worst Tier 1 cell (range 0.67–0.75). The four strongest instruction-tuned LLMs handle interrogatives better (0.82–0.90; nemotron 0.78, minstral 0.37–0.53). Encoders and LLMs do not have the same blind spots; their per-form biases differ mainly on the conditional and interrogative cells.

Mass redistribution from the conditional cell. Per-form precision on the LLM side (Table 11) is in fact high on the conditional cell (0.68–0.89): when the LLM does assign the conditional label, it is usually correct. The conditional gap is therefore a pure under-prediction (low recall, high precision), and Table 10 shows that this under-prediction is not at the binary detection step (binary positive recall on conditionals is 0.84–0.95 for every model). The mass missing from the conditional cell is redistributed to adjacent Tier 1 cells, most commonly modal/deontic and optative, where the rows still trigger binary positives but pick up the wrong form label. The deployment consequence is correspondingly narrower than a simple binary-filter story: a binary suggestion filter will not under-fire on conditional feedback, but any pipeline that routes or reports feedback by Tier 1 form (e.g., a “polite-conditional suggestions” dashboard) will under-

count that form by roughly a factor of two. The “stricter pragmatic reading” interpretation applies if at all to the Tier 1 form classifier’s internal schema disambiguation, not to whether the row is recognized as a suggestion at all.

M Silver Corpus Annotation Disclosure

The 174,860 silver-corpus labels were produced by Claude Opus 4.7 under the guidelines in Appendix B. The silver corpus is released as such, validated against the human gold in §3.5, and is not the evaluation reference. All source-corpus identifiers used in the release are platform-level (review IDs), not user-identifying.

N Comparative Class

With $n = 8$ comparative suggestions in the gold positives, per-model recall, precision, and F1 on this class are descriptive rather than diagnostic. Each score is the outcome of at most eight binary decisions, so a single label flip changes recall by 0.125. Reported model recalls range from 0.375 (gpt-oss:120b) to 1.000 (deepseek-v4-pro); per-form precision in Table 11 ranges from 0.006 (minstral-3:8b) to 0.222 (deepseek-v4-pro). None of these numbers should be read as model behavior on the comparative form. We exclude comparative from the headline Tier 1 macro-recall and recommend that downstream work treat it as an observed long-tail form (directive-by-conversational-implicature; Grice 1975) for which the present gold has insufficient support for hard-label benchmarking. Every comparative cell in the main tables carries an explicit † marker pointing to this appendix. LoRA has zero silver comparatives after the gold hold-out.

O Cross-Domain Transfer

We report Leave-Macro-Out (LMO) and Leave-Industry-Out (LIO) Task A macro-F1 for the encoder fine-tunes. Under LMO, the encoder is trained on silver-174K rows belonging to three of the four macro-domains and evaluated on the held-out macro-domain’s gold rows. Under LIO, the held-out fold is a single industry. Full numerical results are released alongside the data; the principal finding is that RoBERTa-large loses on average 11.2 macro-F1 points under LMO and 6.4 macro-F1 points under LIO relative to in-domain training. The largest LMO degradation is on Product $\rightarrow$ Service (−14.8 macro-F1); the smallest is Service $\rightarrow$

Hybrid retail (−5.1 macro-F1).

P Executed Scoring Prompts

The prompts below are the ones that produced Tables 3–5. Both are issued zero-shot, one call per gold row, temperature 0, no few-shot exemplars, no retries on parse failure. The Tier 1/2/3 prompt expects a one-line JSON object; parse-failure rows are marked `parse_ok = false` in the prediction log and excluded from per-class metrics. Enum values match §3 verbatim.

Task A prompt: binary suggestion detection. *System.*

You are a careful annotator of business reviews. Your task is binary suggestion detection. A suggestion is an utterance in the review that is directed at the business or product owner and that specifies a concrete operational change they could make. Opinions, complaints without a directive, customer-to-customer advice, and pure product description are not suggestions.

User.

Read the review below. Reply with a single lowercase word: yes if the review contains at least one suggestion, no otherwise. Do not include explanations, punctuation, or any other text.
Review: {{review_text}}
Answer (yes or no):

Task B prompt: joint Tier 1 / Tier 2 / Tier 3 labeling (positives) *System.*

You are a careful annotator of business reviews. The review you receive has already been judged to contain at least one suggestion directed at the business or product owner. Your task is to apply a three-tier pragmatic taxonomy to that suggestion and return the labels as JSON.

User.

Read the review below. The review contains at least one suggestion. Label the suggestion using the following three-tier taxonomy.

Tier 1 (single label, exactly one of six): `direct_imperative` (add a vegetarian dish); `modal_deontic` (you should add a vegetarian dish); `optative` (I wish they offered a vegetarian dish); `conditional` (it would be great if they offered a vegetarian dish); `interrogative` (could you add a vegetarian dish?); `comparative` (their vegetarian dish is much better than

yours).

Tier 2 (multi-label, zero or more of four): `frustrated`; `sarcastic`; `backhanded` (great place, but they really need to fix the AC); `hedged` (maybe it could be a bit larger). Return [] if none apply.

Tier 3 (single label, exactly one of three): `vague` (do better with the food); `specific` (add more vegetarian options); `very_specific` (add a tofu bowl and a chickpea wrap to the lunch menu).

Return exactly one JSON object on a single line, with no markdown fencing, no surrounding prose, and no trailing text:

```
{"tier1": <value>, "tier2": [<values>],  
  "tier3": <value>}  
Review: {{review_text}}  
JSON:
```

Parsing. Output is parsed as JSON with a single attempt; on failure the row is marked `parse_ok = false` and excluded from per-class metrics. Predicted enum values are normalized using an alias table (e.g., `imperative` → `direct_imperative`, `wish` → `optative`) before comparison against gold; non-enum predictions outside the alias table are dropped.

Prompt-grid conditions (Table 6). P0 is the Task B prompt above, unchanged. P1 (paraphrase) rewords its scaffolding and example phrases without changing the labels, JSON schema or output instruction, and adds no examples or priority line. P2 (definition) is P0 with an expanded definition of the conditional class ("a hypothetical framed with `would` / `could` / `if`, including hedged or polite hypotheticals that propose a change without issuing a command") and the §3.3 priority-order line inserted before the output instruction. Few-shot ($k = 2, 4, 8$) appends k labeled exemplars drawn from the silver pool (so their labels are Opus 4.7 labels), excluding any review whose key or text matches the gold; the $k = 2$ and $k = 4$ sets are the first two and four exemplars of the $k = 8$ set. P1, P2 and the few-shot variants are separate experimental conditions, not the scoring prompts for Tables 4–5.

Q Model Versions and API Snapshots

Table 16 lists the exact model identifiers, provider snapshots, and API versions used for all models included in the PRAXIS benchmark evaluation, plus the LoRA base.

Model	Snapshot
<i>Silver-corpus annotator</i>	
Claude Opus 4.7	claude-opus-4-7
<i>Instruction-tuned LLMs (API-served)</i>	
DeepSeek V4 Pro	deepseek-v4-pro
DeepSeek V4 Flash	deepseek-v4-flash
GPT-OSS 120B	gpt-oss:120b
Gemma 4 31B	gemma4:31b
Nemotron 3 Nano 30B	nemotron-3-nano:30b
Minstral 3 14B	minstral-3:14b
Minstral 3 8B	minstral-3:8b
<i>Small open LLMs</i>	
Qwen2.5-Instruct 0.5B	qwen2.5:0.5b-instruct
Qwen2.5-Instruct 1.5B	qwen2.5:1.5b-instruct
Qwen2.5-Instruct 3B	qwen2.5:3b-instruct
Qwen2.5-Instruct 7B	qwen2.5:7b-instruct
<i>Matched-supervision LoRA</i>	
Qwen2.5-7B-Instruct (base)	Qwen/Qwen2.5-7B-Instruct

Table 16: Model versions and API snapshot identifiers for every model included in the PRAXIS benchmark evaluation.

R Prompt-Grid Confidence Intervals

Frozen 500 gold conditionals, seed 42, 1,000 bootstrap. Precision is computed on the 1,558-row frozen set (the 500 conditionals plus 1,058 non-conditional positives) and remains 0.83–0.95 in every cell (0.85–0.88 at P2).

Model	Condition	Recall	95% CI
deepseek-v4-pro	P0	.434	[.388, .478]
	P1	.466	[.420, .510]
	P2	.866	[.834, .894]
	$k=2$	.434	[.390, .478]
	$k=4$	.454	[.410, .498]
	$k=8$	.742	[.702, .780]
gpt-oss:120b	P0	.314	[.270, .354]
	P1	.322	[.278, .362]
	P2	.820	[.786, .854]
	$k=2$	.334	[.292, .376]
	$k=4$	.414	[.370, .458]
	$k=8$	.492	[.446, .536]
gemma4:31b	P0	.326	[.282, .368]
	P1	.326	[.280, .368]
	P2	.804	[.768, .838]
	$k=2$	.314	[.272, .356]
	$k=4$	.332	[.290, .376]
	$k=8$	.512	[.460, .558]

Table 17: Prompt-grid conditional recall on the frozen 500-row gold slice. Conditions are defined in Appendix P.

S LoRA Training Protocol

Base: Qwen/Qwen2.5-7B-Instruct. Method: NF4 QLoRA, rank 16, alpha 32, dropout 0.05, targets q, k, v, o , gate, up, down. Train: 8,285 sil-

ver positives after gold hold-out by (review_id, industry) and normalized text (2,342 conditionals; 9 interrogatives; 0 comparatives); val 921 silver (10%, seed 42). 2 epochs, learning rate 1×10^{-4} cosine, effective batch 16, max length 768, one RTX A4500 20 GB (≈ 12 GB used; ≈ 3 GPU-hours). Mean train loss 0.082; val loss 0.070. The train/validation split, hyperparameters and training script are in the code release.

Gold-positive per-form recall: imperative 0.870, modal 0.897, optative 0.938, conditional 0.904 (precision 0.828), interrogative 0.058 (precision 0.938). The interrogative collapse is expected: the gold hold-out absorbed rare forms. Conditional recall on the frozen 500 is 0.898. LoRA is a form classifier scored on gold positives; it is not a Task A detector.

T Domain and Industry Slices

Gold-positive form mix: Product is 45.8% optative / 28.7% conditional ($n = 2,600$); Service is 34.8% modal / 27.9% conditional ($n = 2,800$); Digital/media 46.5% optative / 20.8% conditional ($n = 1,050$); Hybrid 34.7% modal / 25.0% conditional ($n = 1,050$). Conditionals are 21–29% of positives in every macro-domain.

Conditional recall by domain (LoRA vs. zero-shot range across deepseek-v4-pro / gpt-oss:120b / gemma4:31b): Product 0.914 vs. 0.38–0.53; Service 0.908 vs. 0.27–0.42; Digital/media 0.849 vs. 0.36–0.44; Hybrid 0.913 vs. 0.26–0.44. Across 59 industries with ≥ 20 conditionals, median recall is 0.913 (LoRA), 0.462 (deepseek-v4-pro), 0.340 (gpt-oss:120b), 0.333 (gemma4:31b).

U Routing Destinations

On the 2,009 gold conditionals, predicted queues are:

Predicted queue	LoRA	deepseek	gpt-oss	gemma
conditional	1,817	939	687	688
modal_deontic	77	566	322	735
optative	94	353	919	430
direct_imperative	21	89	42	117
interrogative	0	33	27	21
comparative	0	29	11	18

Table 18: Where gold conditionals are sent by each form router. Misses concentrate in modal/deontic and optative.

V Real-Review Running Example and Gallery

gold_plus_03361 (Amazon, Automotive; re-distributable). Gold span *could be a bit stronger*; hedged conditional, vague. Zero-shot: deepseek modal_deontic, gpt-oss optative, gemma modal_deontic. LoRA: conditional. Span in context: **[could be a bit stronger]** but handy.

A domain×form gallery (Yelp/Amazon only; shortest non-borderline row per cell with an aligned span) is in the data release. Illustrative rows:

- Product / Automotive, imperative (gold_plus_00991): “put these on in place of the incandescent lamps” — all four models agree.
- Service / Cafes, conditional (gold_plus_08321): “could use a little more seasoning” — zero-shot models say modal/deontic; LoRA says conditional.
- Digital/media / Digital_Music, conditional (gold_plus_01426): “would have liked a few more songs on it” — zero-shot models say optative; LoRA says conditional.
- Hybrid retail / Beauty, conditional (gold_plus_01956): “could be updated and cleaner” — deepseek conditional, gpt-oss optative, gemma modal/deontic, LoRA conditional.