From TextBlob to LLM Agents: Sentiment Model Selection for B2B Technical Support with CSAT Ground Truth

Pedro Vidigal
Yugabyte
pvidigal@yugabyte.com

Abstract

We present a five-year case study of sentiment model selection for customer satisfaction (CSAT) prediction in B2B technical support. Our evaluation uses the **complete population** of CSAT-rated tickets from Yugabyte's support operation: 533 tickets (26 bad, 507 good) comprising ∼2,500 customer comments from 100+ organizations over five years. We evaluate 17 approaches across 5 paradigms (lexicon, off-the-shelf transformers, NLI zero-shot, multi-task LLM agent, and 12 dedicated LLM agents from 6 vendor families), plus 11 fine-tuning experiments (all achieving MCC$\leq$0). Key findings: (1) a dedicated single-task LLM agent reduces neutral bias from 69% to 22%, improving MCC from -0.018 to 0.347 ($p<0.001$); (2) our results are consistent with the "Alignment Tax" (Lin et al., 2024; Huang et al., 2025) in sentiment classification: Claude Opus 4.6 exhibits 41% neutral predictions and lower recall than its budget model Haiku 4.5 (p=0.003); (3) 5 of 26 dissatisfied customers are missed by *all* 12 LLMs, and 10 of 26 are detected by fewer than 30% of models, due to administrative requests lacking emotional language; (4) Gemini 3 Flash achieves the best MCC (0.347) at \$0.60/1K, over 100$\times$ cheaper than Claude Opus. We provide practitioner recommendations.

Industry Track Category: Deployed

1 Introduction

B2B enterprise technical support presents a unique challenge for sentiment analysis: customers communicate through formal, technically dense language; class imbalance is extreme (only 4.88% of tickets receive negative satisfaction ratings); and the ground truth, customer satisfaction (CSAT), measures *outcome* satisfaction rather than *in-process* emotion. A customer may express frustration about a production outage yet rate the support experience positively if the issue is resolved well. This fundamental gap between process sentiment and outcome satisfaction is poorly studied in the NLP literature, which primarily evaluates sentiment models on social media and product reviews.

We present a 5-year longitudinal case study spanning the full evolution of sentiment model deployment, from TextBlob (2022) through transformer hybrids (2023) to dedicated LLM agents (2025), evaluated against the complete population of real CSAT outcomes. Our most surprising finding: **the most expensive, most capable model (Claude Opus 4.6, \$67.50/1K tickets) performs worse than its cheapest variant (Haiku 4.5, \$3.60/1K)** for CSAT detection. Our results are consistent with the "Alignment Tax" (Lin et al., 2024; Huang et al., 2025) in a new domain: stronger RLHF safety alignment correlates with more conservative "Neutral" predictions, missing the very signals support teams need.

Our contributions span the full model selection lifecycle:

1. **Comprehensive evaluation**: 17 approaches across 5 paradigms, plus 11 fine-tuning experiments, on the complete population of production CSAT tickets spanning five years, the largest such study in technical support.

2. **Deployed system**: Three-phase production evolution from TextBlob (2022) to dedicated LLM agent (2025), with quantified improvements at each phase.

3. **Cross-vendor LLM comparison**: 12 LLMs from 6 families using identical agent architecture, providing evidence consistent with the Alignment Tax (Lin et al., 2024) in production classification (McNemar p=0.003) and revealing a $>$100$\times$ cost-effectiveness gap.

4. **Structural limitation**: 5 of 26 bad CSAT tickets are missed by every LLM, and 10 of

26 are detected by fewer than 30% of models, motivating non-textual features for text-only CSAT prediction.

2 Related Work

Sentiment in Customer Support Manderscheid and Lee (2023) predicted CSAT with ordinal classification at a contact center. Zhao et al. (2025) deployed an agent-in-the-loop framework at Airbnb. Both use single models; we compare 17 approaches across 5 paradigms plus 11 fine-tuning experiments. Zhang et al. (2024) evaluated LLMs on sentiment benchmarks but not on production CSAT data.

Safety Alignment Effects The degradation of model capabilities following safety alignment has been termed the "Alignment Tax" by Lin et al. (2024) and the "Safety Tax" by Huang et al. (2025), who showed it degrades reasoning in large models. Aguda et al. (2025) found LLMs exhibit conservative bias, defaulting to "no relation" labels. Our work provides **domain-specific evidence**: we observe patterns consistent with the Alignment Tax in production sentiment classification, where it manifests as excessive neutral predictions that miss dissatisfied customers.

Technical Text Technical support text contains code, logs, and jargon that confound sentiment models (Rosenthal et al., 2017). We quantify this: a 16-point accuracy drop at high technical density for transformers, and a 26-point drop on long tickets (21+ comments) for LLMs, validating Du et al. (2025).

3 Experimental Setup

3.1 Dataset

Our dataset comprises the **complete population** of CSAT-rated tickets from Yugabyte's B2B support operation, collected over a five-year longitudinal period ending in early 2026. Yugabyte is the company behind YugabyteDB, an open-source distributed SQL database for cloud-native applications. Unlike datasets sampled from larger populations, these 533 tickets represent *every* ticket where a customer responded to a post-resolution satisfaction survey (∼8% response rate). The dataset contains:

- **533 tickets** with binary CSAT ratings: 507 good (95.12%) and 26 bad (4.88%), comprising ∼2,500 customer comments
- **100+ enterprise organizations** across Financial Services, FinTech, Technology, Retail, Automotive, and Media/Entertainment
- **3 geographic regions**: predominantly North America (>90%), with Asia-Pacific and Europe/UK

The 4.88% negative rate is not a sampling artifact; it is the **domain-inherent distribution** of B2B enterprise technical support. Unlike consumer support or social media sentiment, enterprise customers engage through contracted support channels with dedicated engineers; most issues are resolved satisfactorily. Bad CSAT occurs predominantly when issues involve unresolved product defects, administrative failures (e.g., license delays), or prolonged resolution times on production-critical systems. This extreme imbalance (19.5:1) makes the task fundamentally different from balanced sentiment benchmarks and demands metrics like MCC that account for class distribution.

For off-the-shelf transformer evaluation, we additionally analyze over 15,000 total comments from the broader support corpus.

3.2 Models

Table 1 shows the 17 approaches evaluated (excluding fine-tuning experiments reported in Section 6). The 12 LLM agents span models released between February 2025 and February 2026, capturing the complete 2025 landscape of LLM evolution, from early Gemini 2.0 variants through Llama 4 and DeepSeek V3.2 to the latest Claude Opus 4.6. Off-the-shelf transformers are evaluated at comment level (n≈2,500); NLI and most dedicated LLM agents are evaluated at ticket level on n=533 (two local/open models have partial coverage: n=531 and n=527).

NLI formulation. The DeBERTa v3 NLI baseline uses the MoritzLaurer DeBERTa-v3-large zero-shot v2.0 model. Each customer comment is the premise; candidate labels are `customer satisfaction` and `customer dissatisfaction` under the hypothesis template "This customer support comment expresses {label}." We average comment-level satisfaction scores to the ticket level and sweep thresholds from 0.3 to 0.9.

Model	Paradigm	Released	$/1K	N
TextBlob	Lexicon	2018	$0	∼2.5K
Siebert RoBERTa	Transformer	2022	$0	∼2.5K
NLPTown BERT	Transformer	2019	$0	∼2.5K
DeBERTa v3 NLI	Zero-shot	2024	$0	533
Gemini 2.0 Flash (V1)	Multi-task LLM	Feb 2025	$0.40	∼200
Gemini 2.0 Flash Lite	Dedicated LLM	Feb 2025	$0.15	533
Gemini 2.0 Flash	Dedicated LLM	Feb 2025	$0.40	533
Gemini 3 Flash	**Dedicated LLM**	**Dec 2025**	**$0.60**	**533**
Gemini 3 Pro	Dedicated LLM	Jan 2026	$5.00	533
Claude Haiku 4.5	Dedicated LLM	Oct 2025	$3.60	533
Claude Sonnet 4.5	Dedicated LLM	Sep 2025	$13.50	533
Claude Opus 4.6	Dedicated LLM	Feb 2026	$67.50	533
DeepSeek V3.2	Dedicated LLM	Sep 2025	$0.77	533
Llama 4 Maverick	Dedicated LLM	Apr 2025	$0.60	531
Llama 4 Scout	Dedicated LLM	Apr 2025	$0.31	533
Mistral Medium 3	Dedicated LLM	Feb 2025	$1.80	533
Qwen3 8B	Dedicated LLM	May 2025	$0.00	527

Table 1: All 17 approaches evaluated (fine-tuning experiments in Section 6). LLM release dates span Feb 2025 to Feb 2026. All dedicated LLMs use identical agent architecture. $/1K = estimated cost per 1,000 tickets.

3.3 Agent Architecture

All LLM agents use identical infrastructure: a dedicated single-task sentiment agent (Pydantic AI) with tool-calling access to customer-only comments (internal staff filtered out). The system prompt instructs: ignore technical jargon, focus on emotional language, classify as Positive/Neutral/Negative/Frustrated with confidence and reasoning (full prompt in Appendix F). Structured output constrains the response to the valid schema used by the production pipeline. The prompt was developed iteratively using qualitative error analysis on early production tickets before the formal evaluation period; the model set was then evaluated on the CSAT population using this fixed prompt, with available outputs per model reported in Table 1. This controlled design isolates the model effect rather than optimizing prompts per vendor.

3.4 Metrics

For CSAT prediction, we map agent outputs to binary: Negative or Frustrated → predicted Bad CSAT; Positive or Neutral → predicted Good CSAT. This binarization reflects the operational goal of detecting at-risk customers before formal complaints. We use MCC as primary metric (Chicco and Jurman, 2020), reporting bootstrap 95% CIs (1,000 iterations), McNemar's test, Bad CSAT Recall/Precision/F1, ECE, cost/1K, and latency. We did not separately estimate repeated-run decoding variance for LLMs; the reported uncertainty is over the fixed prediction set and paired ticket population.

Model	MCC	Rec	Pre	ECE	$/1K
Dedicated LLM Agents (ticket-level)					
Gemini 3 Flash	**.347**	**65.4**	23.9	**.053**	$0.60
Gemini 3 Pro	.341	53.8	**27.5**	.056	$5.00
Claude Sonnet 4.5	.224	57.7	14.9	.061	$13.50
Claude Haiku 4.5	.176	50.0	12.6	.058	$3.60
Claude Opus 4.6	.165	38.5	13.9	.069	$67.50
DeepSeek V3.2	.129	61.5	8.7	.188	$0.77
Gemini 2.0 Flash	.128	46.2	10.0	.067	$0.40
G2.0 Flash Lite	.069	65.4	6.4	.344	$0.15
Llama 4 Maverick	.032	65.4	5.5	.420	$0.60
Qwen3 8B	.008	38.5	5.2	.284	$0.00
L. Scout / Mistral‡	≤.006	35–46	4–5	>.3	$0.31–1.80
NLI Zero-Shot (ticket-level)					
DeBERTa v3 NLI	.174	45.8	13.1	–	$0.00
Off-the-shelf Models (comment-level)					
Siebert RoBERTa	.086	61.5	3.4	–	$0.00
NLPTown BERT	.056	78.8	2.7	–	$0.00
TextBlob	−.012	51.9	1.9	–	$0.00
Multi-task LLM Agent (ticket-level)					
Gemini 2.0 Flash V1	−.018	0.0	0.0	.464	$0.40

Table 2: All approaches ranked by MCC. Rec/Pre = Bad CSAT Recall/Precision (%). Dedicated LLMs are evaluated at ticket level on the complete CSAT population where model outputs are available (full results in Appendix A). The multi-task agent achieves 96.4% accuracy but MCC<0. ‡Mistral MCC=−.022. *Note*: off-the-shelf transformers are evaluated at comment level; all others at ticket level; these are not directly comparable.

4 Results

4.1 The Full Paradigm Landscape

Table 2 ranks all approaches by MCC, but it should be read as a paradigm landscape rather than a single apples-to-apples leaderboard: off-the-shelf transformers are comment-level diagnostic baselines, while NLI and LLM systems are ticket-level. The paradigm hierarchy is clear: dedicated LLM agents dominate, followed by NLI zero-shot, then off-the-shelf transformers. The multi-task LLM agent (MCC=−0.018) achieves 96.4% accuracy but detects zero dissatisfied customers, a cautionary example of why accuracy misleads on imbalanced data.

McNemar's test shows that Gemini 3 Flash significantly outperforms Claude Sonnet (χ^2=17.56, p<0.001), Claude Opus (p=0.033), and Gemini 2.0 Flash (p<0.001). We note that bootstrap 95% CIs for adjacent models overlap substantially (e.g., G3 Flash [0.203, 0.468] vs. G3 Pro [0.192, 0.479]), reflecting the small minority class (26 bad tickets); we therefore rely on McNemar's pairwise error-pattern tests rather than MCC point estimates for model comparison.

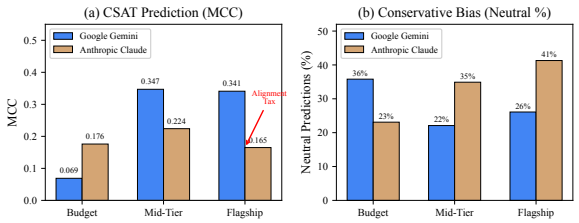

Figure 1: The Alignment Tax. (a) MCC by tier: Google improves; Anthropic inverts. (b) Neutral rate: Opus defaults to Neutral 41.3%. McNemar Haiku vs. Opus: p=0.003.

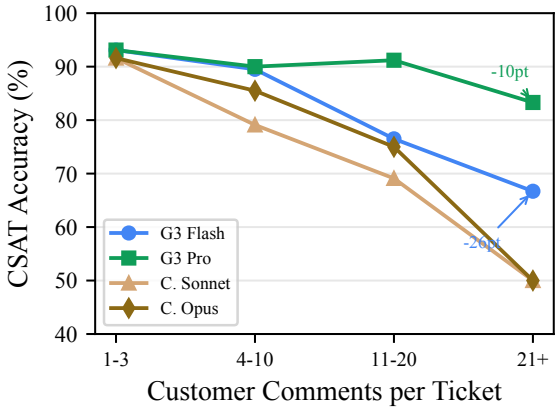

Figure 2: CSAT prediction accuracy by ticket length. Gemini 3 Pro is most length-resilient (−10pt). Claude Sonnet and Opus degrade to 50% at 21+ comments. Accuracy is reported because per-bin minority counts (1–10 bad tickets) are too small for stable MCC. Bin sizes: n=202, 220, 68, 18.

4.2 The Alignment Tax

Figure 1 reveals a tier inversion within Anthropic's Claude family. Google's Gemini improves monotonically across generations and tiers: Flash Lite (.069) → G3 Flash (.347) → G3 Pro (.341). Anthropic *inverts*: Haiku 4.5 (.176) → Sonnet 4.5 (.224) → **Opus 4.6 (.165)**—the flagship is worst.

The mechanism: Opus classifies 41% of tickets as Neutral (highest of any model), yielding only 38.5% bad recall (lowest in its family). This conservative bias is **consistent with** the Alignment Tax (Lin et al., 2024; Huang et al., 2025): the safety-induced degradation previously observed in reasoning tasks may extend to classification calibration, manifesting as a "Neutral Bias" that reduces minority-class recall. We note that alternative explanations exist: tier differences in architecture, training data, and model size could contribute to the observed pattern independently of alignment procedures. With only three models per vendor family, the correlation between tier and neutral rate is suggestive but not conclusive.

4.3 The Invisibility Floor

Five bad CSAT tickets are detected by **zero** of 12 models. All are administrative requests with no emotional language, e.g., *"Please provide a license for YugabyteDB"* (customer waited two weeks; rated Bad), *"YugabyteDB image pull secret"* (three-word subject; too terse for any model), and *"We need to link our YugabyteDB account with [tool]"* (routine request; dissatisfaction arose from a one-week response delay invisible in the text). The common pattern: dissatisfaction is encoded in **metadata** (time-to-resolve, SLA breaches) rather than language. Broadly, 10 of 26 bad tickets (38.5%) are detected by fewer than 30% of models, revealing a structural weakness in text-only CSAT prediction. Addressing this class requires non-textual features.

This is why we do not frame retrieval-augmented generation (RAG) as a direct missing baseline: the model already receives the customer conversation. The hard cases require metadata such as response delay, SLA breach, or entitlement state, making metadata-augmented CSAT prediction a distinct future system design.

4.4 Reasoning Fingerprints

LLM reasoning traces systematically predict errors (Appendix D). Pooling across all models, FP reasoning mentions "production" 26% more than TN, while "thank" appears 39% less. FP predictions are 20–30% longer than correct ones across all models, extending Su et al. (2025). This suggests reasoning content as a zero-cost post-hoc error filter.

4.5 Technical Content and Context Length

Off-the-shelf transformers degrade 16 points at high technical density (Q4 vs. Q1). Figure 2 shows that LLM agents degrade substantially on long tickets, validating Du et al. (2025) in production: Gemini 3 Flash drops 26 points (93.1% at 1–3 comments to 66.7% at 21+), while Claude Sonnet and Opus both fall to 50%. Gemini 3 Pro is most length-resilient (10-point drop vs. 26 for Flash).

4.6 Cost-Effectiveness

Figure 3 shows the Pareto frontier of MCC vs. cost. Gemini 3 Flash dominates: it achieves the

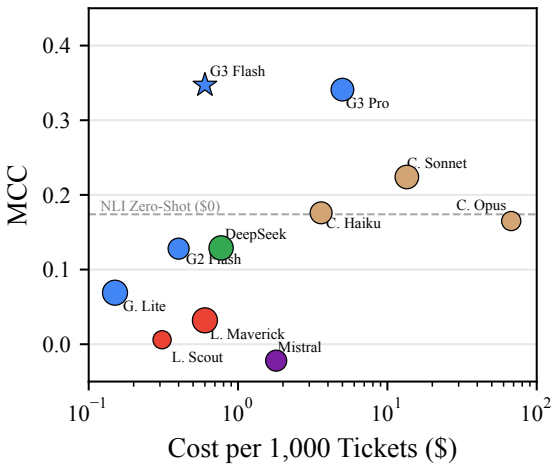

Figure 3: Cost-effectiveness Pareto frontier. Gemini 3 Flash (star) dominates: highest MCC at low cost. Claude Opus (rightmost) pays 112× more for worse results. Dashed line = NLI zero-shot (free baseline). Marker size proportional to Bad CSAT Recall.

highest MCC (0.347) at $0.60/1K, while Claude Opus 4.6 pays 112× more for lower MCC (0.165). DeepSeek V3.2 ($0.77/1K) offers the best recall (61.5%) for budget-constrained high-recall deployments. The free NLI zero-shot baseline (MCC=0.174) outperforms several paid LLMs.

5 Deployed System

The production system evolved through three phases, each informed by the findings above:

Phase 0: TextBlob (2022) Lexicon-based scoring achieving 44.1% CSAT accuracy with *negative* separation (-0.022)—worse than random. This motivated the model search.

Phase 1: Hybrid Transformer + Multi-task LLM (2023–2024) Siebert + NLPTown with technical-density confidence discount (38.5% bad recall, F1=0.065). We also deployed a multi-task LLM agent that generated sentiment alongside 20+ other ticket fields (priority, next steps, closing notes). Despite 96.4% accuracy, this agent achieved 0% bad CSAT recall (MCC=−0.018); it optimized for the aggregate task and defaulted to "Neutral" 69% of the time. This *multi-task dilution* motivated the Phase 2 redesign.

Phase 2: Dedicated LLM Agent (2024) Gemini 3 Flash (current LLM in production) with customer-comment filtering and single-task prompt. Achieves 65.4% bad recall, 23.9% precision (MCC=0.347). Architecture: Pydantic AI agent on Google Cloud Functions invoking Gemini 3 Flash via Vertex AI with structured output; PII-filtering pipeline strips emails/IPs before LLM ingestion; tool-calling retrieves only customer-facing comments. Latency: 14.5s/ticket. Cost: $0.60/1K (∼$15/month).

Phase	Bad Rec.	MCC	Cost
0: TextBlob (2022)	∼0%	−.012	$0
1: Hybrid (2023)	38.5%	.065†	$0
2: LLM Agent (2024)	**65.4%**	**.347**	$0.60/1K

Table 3: Deployment evolution. †F1 reported as MCC proxy (comment-level evaluation). Each phase was motivated by the empirical findings in this paper.

Operational Integration Sentiment predictions feed a live monitoring dashboard where all active tickets are tracked with their predicted sentiment, confidence scores, and model reasoning. Tickets predicted as Negative or Frustrated trigger alerts to dedicated Slack channels, notifying the support team of potentially at-risk customers *before* a formal complaint is filed. Each alert includes the model's reasoning and confidence score, enabling rapid triage. Engineers can dismiss clear false positives in seconds. In B2B enterprise support, the asymmetric cost structure justifies the low precision (23.9%): a single missed escalation on a production-critical system can cost orders of magnitude more in customer trust and contract renewals than the time spent reviewing false alerts. Since the Phase 2 deployment in late 2024, the system has analyzed over 3,000 tickets, covering >95% of monthly ticket volume (∼250–300 tickets/month) across the full support operation.

The cross-vendor evaluation (Section 4.2) supported this model choice: Gemini 3 Flash outperforms all 11 alternatives tested, including models costing 112× more. Table 3 summarizes the improvement trajectory.

6 Discussion

Practitioner Recommendations (1) **Default**: Gemini 3 Flash ($0.60/1K, MCC=.347, best calibration). (2) **Long tickets**: Gemini 3 Pro (most length-resilient). (3) **Maximum recall**: DeepSeek V3.2 (61.5% at $0.77/1K). (4) **Budget**: NLI zero-shot (MCC=.174, $0/1K). (5) **Avoid**: Claude Opus for classification ($67.50/1K, MCC=.165).

Why Accuracy Misleads The V1 multi-task agent (96.4% accuracy, MCC=−0.018) illustrates why MCC must replace accuracy for imbalanced CSAT (Chicco and Jurman, 2020).

Fine-Tuning Cannot Compete Eleven fine-tuning experiments on Siebert RoBERTa-large **all achieved ticket-level MCC≤0** with zero bad recall (Table 7 in Appendix G). Techniques spanned standard cross-entropy, focal loss (γ=2,5), EDA augmentation, LoRA (r=8,16), K-fold ensemble, self-training, and domain-adaptive pre-training (DAPT). With only 23 bad CSAT tickets in the filtered fine-tuning dataset (4.63%; 36.9:1 imbalance), no technique overcame data scarcity and the supervision mismatch of propagating ticket-level CSAT outcomes to individual comments. Total cost: $0.19 (spot GPU, 0.53 hours). LLM agents are the pragmatic choice.

Critical Deployment Lessons Two architectural decisions produced the largest improvements (detailed in Appendix E). First, **context control**: filtering input to customer-facing comments only, removing internal staff discussions, Slack threads, and JIRA references, was the single largest contributor to Phase 2's improvement. Second, **single-task focus**: separating sentiment from a 20-field multi-task agent into a dedicated single-task agent reduced neutral bias from 69% to 22% and was the difference between a useless system (MCC=−0.018) and a useful one (MCC=0.347).

Customer Segments Enterprise customers: 2.2% bad CSAT, 86% detection. OSS community: 42.9% bad CSAT, 33% detection. The 53-point recall gap is driven by administrative requests with no emotional signal.

Precision and Alert Management The best model's 23.9% precision means approximately three false alerts per true detection. In practice, this is operationally sustainable because: (1) each alert surfaces the model's reasoning and confidence, enabling sub-minute triage; (2) the base rate of flagged tickets is low (∼13% of all tickets), producing a manageable alert volume; and (3) the asymmetric cost of a missed at-risk enterprise customer far exceeds the cost of reviewing false positives. The reasoning fingerprints (Section 4.4) offer a path to further precision improvement as a zero-cost post-hoc filter.

7 Conclusion

We present the most comprehensive evaluation of sentiment models for technical support CSAT prediction: 17 approaches across 5 paradigms plus 11 fine-tuning experiments, including 12 dedicated LLMs from 6 vendor families, evaluated on production CSAT tickets spanning five years. Dedicated LLM agents dominate all other paradigms (MCC=0.347 vs. 0.086 for the best transformer), but model tier and cost are not reliable proxies for CSAT performance. For practitioners: deploy Gemini 3 Flash at $0.60/1K, and accept that some dissatisfied customers are invisible to text-only sentiment models; addressing them requires non-textual signals.

Limitations

Single domain. YugabyteDB enterprise software technical support; findings may not generalize to consumer support, high-volume contact centers, or non-technical domains where customers use more explicit affective language. **Complete but small minority class.** Our bad CSAT tickets (26 of 533; 4.88%) represent the *complete population* over five years, not a sample that could be enlarged. This is the genuine prevalence of B2B dissatisfaction, analogous to rare event classification in medical diagnostics. Bootstrap CIs are correspondingly wide (MCC ±0.13), and we rely on McNemar's test for pairwise comparisons precisely because it tests error patterns rather than point estimates. We rejected synthetic data augmentation (SMOTE, GPT-generated examples) because our qualitative analysis showed that synthetic examples failed to capture the subtle "neutral-negative" duality of administrative failures, the very category that defines our Invisibility Floor. **LLM stochasticity.** We report bootstrap CIs and paired McNemar tests over fixed model outputs, but do not separately measure repeated-run decoding variance. **US-dominated.** Over 90% of tickets from North American timezones. **Survey response bias.** Only ∼8% of customers responded to CSAT surveys; non-respondents may differ systematically. **Infrastructure.** All LLMs accessed via Google Vertex AI; Google models may benefit from infrastructure advantages (lower latency, native tool-calling). Llama models required a tool_choice workaround; results on Anthropic's native API may differ. **Cost estimates.** Based on published pricing and estimated token counts, not measured

actual usage.

Ethical Considerations

Customer and organization identifiers are anonymized in the analysis. The system is advisory (no auto-escalation). Models perform differently across customer segments (Section 6), which could cause unequal service if deployed without segment-aware calibration.

Acknowledgments

I thank Karthik Ranganathan and Kannan Muthukkaruppan for their visionary leadership and commitment to advancing AI-driven solutions at Yugabyte. Their emphasis on data-driven decision making and customer-centric engineering culture made this research possible. I thank the Yugabyte Support Engineering team for their domain expertise, help building this solution, and invaluable feedback throughout this study. I also thank Tyler Ramer for early encouragement to build Hagen and pursue AI-assisted workflows across Yugabyte, Eugene Kim for sustained support throughout the project, and my sister Joana Vidigal for inspiring me to write and complete this paper.

References

Toyin Aguda, Erik Wilson, Allan Anzagira, Simerjot Kaur, and Charese Smiley. 2025. Conservative bias in large language models: Measuring relation predictions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18989–18998, Vienna, Austria. Association for Computational Linguistics.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:6.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298, Suzhou, China. Association for Computational Linguistics.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *arXiv preprint arXiv:2503.00555*.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Etienne Manderscheid and Matthias Lee. 2023. Predicting customer satisfaction with soft labels for ordinal classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 652–659, Toronto, Canada. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518. Association for Computational Linguistics.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025. Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs. *arXiv preprint arXiv:2505.00127*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Cen Zhao, Tiantian Zhang, Hanchen Su, Yufeng Zhang, Shaowei Su, Mingzhi Xu, Yu Liu, Wei Han, Jeremy Werner, Claire Na Cheng, and Yashar Mehdad. 2025. Agent-in-the-loop: A data flywheel for continuous improvement in LLM-based customer support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1919–1930, Suzhou, China. Association for Computational Linguistics.

A Full Confusion Matrices

B Statistical Tests

C Additional Analysis

Reasoning Fingerprints. LLM reasoning traces predict errors: "production" appears 26% more in FP than TN; "frustrated" +24%; "thank" −39%. FP reasoning is 20–30% longer than TP across all models, extending Su et al. (2025). A post-hoc filter flagging predictions with 3+ FP keywords identifies FPs 87% of the time in G3 Flash (n=9, preliminary).

Model	TP	FP	FN	TN
G3 Flash	17	54	9	453
G3 Pro	14	37	12	470
C. Sonnet	15	86	11	421
C. Haiku	13	90	13	417
C. Opus	10	62	16	445
DeepSeek	16	168	10	339
G2.0 Flash	12	108	14	399
Flash Lite	17	250	9	257
L. Maverick	17	293	9	212
Qwen3 8B	10	184	16	317
L. Scout	9	169	17	338
Mistral	12	260	14	247

Table 4: Confusion matrices for all 12 dedicated LLM agents. TP = correctly detected bad CSAT. FP = flagged good CSAT as bad. Most models have n=533; Llama Maverick has n=531 and Qwen3 has n=527.

Model	MCC	Bootstrap 95% CI
G3 Flash	.347	[.203, .468]
G3 Pro	.341	[.192, .479]
C. Sonnet	.224	[.108, .334]
C. Haiku	.176	[.070, .290]
C. Opus	.165	[.044, .286]
DeepSeek	.129	[.036, .217]
G2 Flash	.128	[.030, .231]
G2 Lite	.069	[−.017, .144]
L. Maverick	.032	[−.050, .105]
Qwen3	.008	[−.081, .097]
L. Scout	.006	[−.074, .098]
Mistral	−.022	[−.106, .065]

Table 5: Bootstrap 95% CIs (1,000 iterations) for MCC. Models below G2 Lite have CIs crossing zero, indicating performance not distinguishable from random.

Inter-Model Agreement. Within-Google κ=0.385 (moderate; high internal diversity); within-Anthropic κ=0.606 (substantial; internally consistent); cross-family Google–Anthropic κ=0.476 (moderate). Notably, Mistral Medium 3 and Gemini Flash Lite exhibit κ=0.792 despite being from different vendors, suggesting similar training biases toward aggressive negative prediction.

Customer Segments. Bad CSAT by customer type: Enterprise (Fortune 500) ∼2%, Mid-Market ∼1.5%, Open-Source Community ∼43%. Model detection (G3 Flash recall): Enterprise 86%, OSS Community 33%, a 53-point gap driven by administrative requests with no emotional signal.

Temporal Trend. Bad CSAT dropped from ∼25% (late 2021) to ∼1.5% (mid-2025), a >16× improvement over four years, reflecting maturation of the support operation. Saturday tickets show 2× the weekday bad CSAT rate.

Model A	Model B	χ^2	p	Sig.
G3 Flash	C. Sonnet	17.56	<.001	***
G3 Flash	G2 Flash	36.17	<.001	***
C. Haiku	C. Opus	9.14	.003	**
G3 Flash	C. Opus	4.56	.033	*

Table 6: McNemar's pairwise tests on error patterns. Tests whether two models make significantly different errors on the same tickets. *p<.05, **p<.01, ***p<.001.

D Lessons Learned: Critical Success Factors

Our five-year deployment journey identified four critical success factors for LLM-based sentiment agents in B2B support:

Context Control. Our largest improvement came from filtering input to customer-facing comments only, removing internal staff discussions, Slack threads, and JIRA references that diluted the sentiment signal. A PII-filtering pipeline strips emails and IPs before LLM ingestion. This single change, feeding the model *only what the customer wrote*, was the largest contributor to Phase 2's improvement. We recommend that practitioners implement similar input filtering before LLM ingestion, particularly in B2B environments where internal and external communications are interleaved.

Single-Task Focus. Our V1 multi-task agent produced sentiment alongside 20+ other ticket fields, resulting in 69% neutral bias and 0% bad recall. Separating sentiment into a dedicated single-task agent reduced neutral bias to 22% and achieved MCC=0.347, the difference between a useless and a useful system. For any classification task where minority-class detection is critical, we recommend dedicated agents over multi-task generation, even when the underlying model is the same.

Explicit Prompt Engineering. Generic sentiment prompts failed in our B2B domain. Performance improved only after the prompt explicitly addressed: (1) ignoring technical jargon (error messages are diagnostic, not emotional); (2) detecting polite masking (formal language does not mean satisfied); (3) distinguishing urgency from dissatisfaction (production outages cause stress regardless of politeness). Practitioners deploying LLM sentiment in specialized domains should invest in domain-specific prompt instructions rather

than relying on the model's general-purpose capabilities.

Structured Output. Enforcing a strict output schema (sentiment label, confidence score, reasoning, escalation flag) via Pydantic AI prevented format errors that broke our downstream pipelines during early testing. The reasoning field proved additionally valuable: its content predicts model errors (Section 4.4), enabling zero-cost post-hoc quality filtering. We recommend requiring structured output with a reasoning field for any production sentiment system; it serves both operational reliability and introspective error detection.

E System Prompt

The following prompt is used identically across all 12 dedicated LLM agents (only the underlying model differs):

```
You are a specialized Customer Sentiment
Analyst for technical support.

YOUR TASK: Analyze ONLY the customer's
emotional state from their support ticket
comments. You will receive only customer
comments (internal staff comments are
pre-filtered).

CRITICAL RULES:
1. IGNORE technical content completely:
   error messages, stack traces, config
   snippets, version numbers, file paths,
   log outputs, technical jargon.
2. FOCUS on emotional language and pragmatic
   intent: explicit/implicit frustration,
   polite masking, gratitude.
3. Consider context patterns: repeated
   issues, production outages, long-running
   tickets, quick acknowledgments.

SENTIMENT DEFINITIONS:
- POSITIVE: gratitude, satisfaction, optimism
- NEUTRAL: factual, neither frustrated nor
  pleased
- NEGATIVE: mild dissatisfaction or concern
- FRUSTRATED: strong negative emotion,
  urgency, repeated complaints

DECISION GUIDANCE:
- Uncertain Neutral/Negative: prefer NEGATIVE
- Professional language != satisfied
- Production impact/urgency -> FRUSTRATED

OUTPUT: Sentiment, confidence (0-1),
reasoning, key phrases, escalation
recommendation.
```

F Fine-Tuning Experiments

Table 7 summarizes all 11 fine-tuning experiments on Siebert RoBERTa-large. All used 80/20 train/test splits with the minority class stratified. Every experiment achieved ticket-level MCC$\leq$0 with zero bad CSAT recall, confirming that $\sim$26 minority examples are insufficient for supervised fine-tuning regardless of technique.

#	Technique	Epochs	MCC	Bad Rec.
1	Cross-entropy (baseline)	5	$-.003$	0%
2	Focal loss (γ=2)	5	$-.006$	0%
3	Focal loss (γ=5)	5	$-.008$	0%
4	EDA + cross-entropy	5	.000	0%
5	EDA + focal loss (γ=2)	5	$-.004$	0%
6	LoRA (r=8)	10	$-.002$	0%
7	LoRA (r=16)	10	$-.005$	0%
8	K-fold ensemble (5-fold)	5	$-.001$	0%
9	Self-training (2 iterations)	5	.000	0%
10	DAPT (support corpus)	10+5	$-.007$	0%
11	DAPT + focal loss (γ=2)	10+5	$-.009$	0%

Table 7: All 11 fine-tuning experiments on Siebert RoBERTa-large. MCC and Bad Recall at ticket level. All used lr=2×10^{-5}, batch size 16, 80/20 stratified split. DAPT epochs shown as pre-train+fine-tune. Total compute: 0.53 GPU-hours ($0.19 on spot A100).