Review Article

An Empirical Comparison of Claude, Llama, and Gemini for Aspect-Level Sentiment

Pınar Ersoy[1*], Mustafa Erşahin[2]

[1] Department of Research & Development, Commencis, Istanbul, Turkey, Orcid ID: https://orcid.org/0000-0001-9591-3037, E-mail: pinar.ersoy@commencis.com
[2] Department of Research & Development, Commencis, Istanbul, Turkey, Orcid ID: https://orcid.org/ 0000-0003-4318-8288, E-mail: mustafa.ersahin@commencis.com
[*] Correspondence: pinar.ersoy@commencis.com; Tel.: +90 533 934 78 71

Received: 12 June 2025
Revised: 11 October 2025
Accepted: 18 November 2025
Published: 23 November 2025

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Reference: Ersoy, P., & Erşahin, M. (2025). An empirical comparison of Claude, Llama, and Gemini for aspect-level sentiment. The European Journal of Research and Development, 5(1), 149–163.

Abstract

Aspect-based sentiment analysis provides granular insights into customer feedback by identifying discrete aspects, such as features or topics, and assigning a corresponding sentiment to each. This study assesses three large language models, hereafter referred to as LLMs, namely Google Gemini 2.5 Flash-Lite, Anthropic Claude Sonnet-4 delivered through AWS Bedrock, and Meta LLaMA 3.3 70B delivered through AWS Bedrock, using a real-world multilingual corpus of 7,841 Turkish mobile banking app reviews from İşbank in Turkey. We employ a prompt-based tagging protocol to extract aspect–sentiment pairs from every review, and we compare accuracy, F1-score, inference cost, and latency. The results show that all three LLMs can execute multilingual aspect extraction and sentiment categorization without task-specific fine-tuning. Claude Sonnet-4 attains the highest F1 for aspect extraction and the highest sentiment accuracy, although it incurs a markedly higher inference cost. Gemini 2.5 Flash-Lite achieves competitive accuracy at a fraction of the price, making it well-suited for high-volume analytics. Meta LLaMA at the 70B scale accessed through AWS Bedrock exhibits intermediate performance with moderate cost and latency. We provide detailed performance tables and figures, along with best-practice guidance for enterprise deployment. AWS Bedrock enables the strategic selection of Claude and LLaMA 3.3 70B for

multilingual sentiment analysis, offering valuable insights from app reviews within scale, accuracy, and budget constraints.

Keywords: Natural Language Processing, Sentiment Analysis, Generative AI, Large Language Models

1. Introduction

Online app store reviews contain information that far exceeds a single star rating. Users frequently refer to several attributes within a single comment; for example, they may praise the visual design while criticizing recurring login failures in the same review. Aspect-based sentiment analysis identifies the specific aspects mentioned in a text and determines the sentiment associated with each element. By moving beyond coarse whole-review polarity, aspect-level analysis enables product teams to understand why customers are satisfied or dissatisfied. This fine-grained feedback is crucial for banking applications, which encompass multiple facets, including interface quality, performance, security, and payment processing. A single positive or negative label per review cannot reveal which features delight users or which ones cause friction. Aspect-level insight, by contrast, supports targeted improvements, for example, identifying login reliability as a recurring pain point, even when the overall rating appears neutral (M. Pontiki, D. Galanis, 2014)[1].

Conducting aspect-based sentiment analysis on real-world data is a challenging task. Customer reviews are unstructured, often include multiple aspects with mixed sentiments, and in many cases are written in languages other than English. Early industrial solutions relied on keyword lists and predefined tags. These methods are transparent, yet they struggle with synonyms and contextual meaning, for instance, distinguishing secure login from login failure, and they are unable to capture new or unexpected topics. Supervised machine learning pipelines have also been developed, separating aspect extraction from sentiment classification, but they require substantial annotated data. Creating significant, high-quality annotations for each new domain, such as Turkish banking applications, is costly and time-consuming. Unsupervised topic models, such as LDA and NMF, can discover latent themes without labels and group terms that often correspond to aspects. For example, clustering words like “yavaş”, “donuyor”, and “ağır” can indicate performance and speed (D. M. Blei, A. Y. Ng, M. I. Jordan, 2003)[2], (D. D. Lee, H. S. Seung, 1999)[3]. These models are helpful for exploration; however, they typically require a manual step to interpret and name clusters, and they can conflate aspect and sentiment unless a dedicated sentiment component is added.

Large language models introduce a new paradigm. Instead of constructing a pipeline, a general-purpose model can be prompted in a zero-shot or few-shot fashion to return aspect and sentiment pairs directly. Recent studies report that frontier models can approach human annotators on aspect-sentiment tasks, which suggests that prompt-driven analysis can quickly unlock multilingual customer insights without task-specific training [6]. Other investigations have shown that zero-shot and few-shot settings often lag fine-tuned models on formal benchmarks, particularly in multilingual settings and when structured outputs are required (C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, Y. Xue, 2024)[4] (P. F. Simmering, R. Werkmeister, L. Di Stasio, 2023)[5] (J. Šmíd, M. Bělohlávek, and T. Brychcín, 2024)[7]. Open-source families that are fine-tuned for aspect-based tasks can surpass prior results on English datasets. At the same time, general models may struggle without careful adaptation and prompt design (P. F. Simmering, R. Werkmeister, L. Di Stasio, 2023)[5] (J. Šmíd, M. Bělohlávek, and T. Brychcín, 2024)[7]. Variation across models remains significant with respect to language coverage, consistency of structured outputs, and handling of domain-specific vocabulary, which motivates an empirical comparison tailored to a well-defined aspect-based task (C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, Y. Xue, 2024)[4] (P. F. Simmering, R. Werkmeister, L. Di Stasio, 2023)[5] (J. Šmíd, M. Bělohlávek, and T. Brychcín, 2024)[7].

This study compares three large language models, namely Gemini 2.5 Flash-Lite, Claude Sonnet-4, and LLaMA, by analyzing Turkish banking application reviews. These models represent state-of-the-art technology from major providers, including Google, Anthropic through AWS Bedrock, and Meta through AWS Bedrock. Gemini 2.5 Flash-Lite is designed for speed and cost efficiency in high-throughput classification workloads. Claude Sonnet-4 is engineered to provide an optimal balance of quality and responsiveness for enterprise-level applications, accommodating extensive contexts that enable thorough analyses. LLaMA serves as an open-source language model, available in various sizes on AWS Bedrock, and demonstrates robust multilingual capabilities and adaptability. By comparing these three models, the study quantifies the accuracy of aspect extraction and sentiment assignment. It examines the practical trade-offs in inference cost and latency that matter for enterprise adoption, thereby guiding when and how to use AWS Bedrock with Claude and LLaMA for scalable multilingual sentiment analysis in production settings, supported by evidence from controlled experiments on authentic customer feedback [(C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, Y. Xue, 2024)[4] (P. F. Simmering, R. Werkmeister, L. Di Stasio, 2023)[5] (M. Água, P. Pina, and B. Ribeiro, 2025)[6] (J. Šmíd, M. Bělohlávek, and T. Brychcín, 2024)[7].

2. Materials and Methods

Related Work and Literature Review

Online app store reviews convey information that far exceeds a single star rating. Users frequently reference several attributes within a single comment; for example, they may praise the visual design while criticizing recurring login failures in the same review. Aspect-based sentiment analysis identifies the specific aspects mentioned in a text and determines the sentiment associated with each aspect. By moving beyond coarse whole-review polarity, aspect-level analysis enables product teams to understand why customers are satisfied or dissatisfied with specific aspects of their experience. This fine-grained feedback is crucial for banking applications, where the user experience encompasses multiple facets, including interface quality, performance, security, and payment processing. A single positive, negative, or neutral label per review cannot reveal which features delight users or which ones cause friction. Aspect-level insight, by contrast, supports targeted improvements, for example, isolating login reliability as a recurring pain point, even when the overall rating appears neutral.

Conducting aspect-based sentiment analysis on real data is challenging. Customer reviews are unstructured, often include multiple aspects with mixed sentiments, and in many cases are written in languages other than English. Early industrial solutions relied on keyword lists and predefined tags. These methods are transparent, yet they struggle with synonyms and contextual meaning, for instance, distinguishing between secure login and login failure, and they are unable to capture new or unexpected topics. Supervised machine learning pipelines have also been developed, separating aspect extraction from sentiment classification, but they require substantial annotated data. Creating significant, high-quality annotations for each new domain, such as Turkish banking applications, is costly and time-consuming. Unsupervised topic models, such as LDA and NMF, can discover latent themes without labels and group terms that often correspond to aspects. For example, clustering words like “yavaş”, “donuyor”, and “ağır” can indicate performance and speed. These models are helpful for exploration; however, they typically require a manual step to interpret and name clusters, and they can conflate aspect and sentiment unless a dedicated sentiment component is added.

Large language models introduce a new paradigm. Instead of constructing a pipeline, a general-purpose model can be prompted in a zero-shot or few-shot fashion to return aspect and sentiment pairs directly. Recent studies have reported that frontier models can approach human annotators on aspect-sentiment tasks, suggesting that prompt-driven analysis can quickly unlock multilingual customer insights without requiring

task-specific training. Other work notes that zero-shot and few-shot settings may still lag fine-tuned models on formal benchmarks. Open-source families that are fine-tuned for aspect-based tasks can surpass prior results on English datasets. At the same time, generative language models may struggle without proper adaptation, particularly in terms of language coverage, consistency of structured output, and domain-specific vocabulary. This study compares three large language models, Gemini 2.5 Flash-Lite, Claude Sonnet-4, and LLaMA, by analyzing Turkish banking application reviews. These models represent state-of-the-art technology from major providers, including Google, Anthropic through AWS Bedrock, and Meta, also through AWS Bedrock. Each is associated with distinct strengths. Gemini 2.5 Flash-Lite is designed for speed and cost efficiency in high-throughput classification workloads. Claude Sonnet-4 is engineered to provide an optimal balance of quality and responsiveness for enterprise-level applications, accommodating extensive contexts that enable thorough analyses. Meta's LLaMA serves as an open-source language model offered in various sizes on AWS Bedrock, demonstrating robust multilingual capabilities and adaptability. By comparing these three models, the study quantifies the accuracy of aspect extraction and sentiment assignment. It examines the practical trade-offs in inference cost and latency that matter for enterprise adoption. The objective is to guide when and how to use AWS Bedrock with Claude and LLaMA for scalable multilingual sentiment analysis in production settings, supported by evidence from controlled experiments on authentic customer feedback.

Dataset

We use a proprietary dataset of 7,841 application store reviews for the İşbank mobile banking app, collected in early 2025. This dataset represents a multilingual sentiment analysis scenario, where the reviews are primarily in Turkish, with occasional code-mixing, and each review can mention multiple distinct aspects of the app. The review time span spans late 2024 to 2025, reflecting user feedback after major app updates introduced new features.

Data Characteristics

The aspects discussed in these reviews include both high-level features and specific functionalities. Based on an initial reading and prior analysis in 2021, common aspect categories are: UI/Design (e.g. “arayüz çok güzel” – “the interface is very nice”), Performance (speed and stability; “uygulama ağır çalışıyor” – “the app works slowly”), Login/Authentication (“QR ile giriş çalışmıyor” – “login via QR does not work”), Payments/Transfers (“para gönderme işlemi takılıyor” – “money transfer operation

stalls"), Card Services (issues with credit/debit card features), Customer Service ("müşteri hizmetleri dönüş yapmadı" – "customer service did not respond"), and Notifications, among others. Each review in the dataset is accompanied by a star rating (1–5) as provided on the App Store; however, these star ratings often do not directly correlate with the sentiment of each aspect mentioned (a 3-star review might praise one aspect but criticize another). We do not use star ratings in our analysis, except as context; instead, we focus on extracting textual aspect sentiments. No ground-truth aspect annotations were available for the 2025 dataset. To evaluate model performance, we constructed a ground-truth test set by manually annotating a subset of 500 reviews. Three native Turkish speakers labeled each review in this subset with all explicit aspect–sentiment pairs. The inter-annotator agreement on aspect and sentiment identification was high. This gold-standard set is used to calculate accuracy and F1 scores for the models' predictions. Additionally, we leverage a previously tagged 2021 dataset of 9,454 İşbank reviews for prompt development and double-checking model outputs (the 2021 data had been annotated with aspect categories and sentiment for an earlier study, achieving about 91.6% sentiment classification accuracy with classical methods).

4. Models and Prompting

We evaluated three large language model types: Gemini 2.5 Flash-Lite, Anthropic Claude Sonnet-4, and Meta LLaMA as detailed below.

Gemini 2.5 Flash-Lite (Google DeepMind)

This is Google's latest-generation LLM (Gemini) in its Flash-Lite variant, optimized for high-throughput and low-latency tasks. It supports multimodal input, but we used it in text mode only. According to Google's model card, 2.5 Flash-Lite is the most cost-efficient model in the Gemini family, with pricing of $0.10 per 1M input tokens and $0.40 per 1M output – approximately 30 times cheaper on input and 37 times cheaper on output than GPT-4 or Claude's top model. We accessed Gemini through the Google Cloud Vertex AI API. The context length of this model is up to 1 million tokens (with sparse attention), but our use did not approach that limit.

Anthropic Claude Sonnet-4 (via Amazon Bedrock)

Claude 4 is Anthropic's flagship LLM series akin to GPT-4. The Sonnet-4 model is described as a "mid-size model with a balance of quality, cost-effectiveness, and responsiveness" built for high-volume applications. It is essentially a distilled version of

Claude Opus (the largest model) that still offers frontier performance on a varying range of tasks while being cheaper and faster per call. Importantly for our use, Claude Sonnet-4 supports up to a 200k token context, which is beneficial for reading long documents or batches of text. In our case, we input one review at a time, so context length was not a limiting factor, but the model's ability to handle long prompts allowed us to include extensive instructions and examples. We accessed Claude Sonnet-4 via AWS Bedrock, a fully managed service for foundation models. Bedrock's Claude integration ensured data residency in our chosen region and easy scaling via API. According to Anthropic's pricing, Claude Sonnet-4 costs $3 per million input tokens and $15 per million output tokens. Claude is known for its strong multilingual comprehension and advanced reasoning abilities. This made it a convincing choice for aspect mining in Turkish reviews.

Meta LLaMA (70B Instruct) (via Amazon Bedrock)

LLaMA is an open-source family of models released by Meta AI, with versions ranging from 7B to 70B. On AWS Bedrock, various fine-tuned Instruct variants of LLaMA are available; we selected the LLaMA 3.3 70B Instruct model to maximize accuracy. This model has been refined to follow instructions and supports multiple languages due to the multilingual data in LLaMA's training set. One advantage of LLaMA on Bedrock is its flexible scaling and cost, as smaller sizes are also offered. This allows an enterprise to trade off some accuracy for a much lower price. In our evaluation, we focused on the 70B to see the maximum capability of LLaMA. Like Claude, LLaMA was accessed through Bedrock's unified API, which allowed us to easily call the model without managing any custom infrastructure. We did not apply any fine-tuning; the model was used as provided. Pricing for the LLaMA-70B was approximately $0.002 per 1,000 tokens (input or output), meaning it is substantially cheaper than Claude for output tokens ($2 vs $15 per million). However, it is also slightly more affordable for input. LLaMA's context window on Bedrock was 4k tokens for the version we used, which was sufficient for our prompt lengths.

Prompt Design

All models were given a very similar instruction prompt tailored for aspect extraction. We crafted the prompt in English, as all three LLMs have English instruction-following capabilities, and requested the output in a structured JSON format.

We included a few-shot example in the prompt for Gemini and Claude, as we found this improved their extraction consistency. Due to LLaMA's shorter context limit, we gave it

a zero-shot or one-shot prompt to avoid truncation. An example embedded in the prompt was as follows.

```
Input Review (Turkish):
"Arayüz çok güzel ama ödeme sürekli hata veriyor."

Expected Output:
[{"Aspect": "UI", "Sentiment": "Positive"},
 {"Aspect": "Payments", "Sentiment": "Negative"}]
```

Figure 1: The Input Review and Expected Output in JSON format

This example demonstrates a case with two aspects in one sentence (UI positive, Payments negative). By doing so, we signaled to the model how to handle contrastive opinions in a single review. We also provided an example with implicit aspect sentiment: e.g., "Uygulama bazen donuyor." with expected aspect "App Performance" and sentiment "Negative", to guide models in naming the aspect of the app freezing issue. Prior internal tests guided the inclusion of such examples: without examples, we noticed some models might output aspects in Turkish or provide a single overall sentiment. With examples, all models reliably produced the desired JSON with English aspect labels and correct sentiment tags. We note that prompt tokens contributed significantly to the total token count per query. This large prompt size was deliberate to maximize accuracy.

Each review was applied to the models independently. For Claude and LLaMA, we used the AWS Bedrock invoke API in us-east-1, with the models running under on-demand inference. For Gemini, we used Google's Vertex AI endpoint in their europe-west4 region. We did not impose a hard temperature or randomness setting. By default, Claude and LLaMA on Bedrock use a temperature around 0.2 for deterministic instruction following, and we set Gemini's parameters similarly to prioritize correctness over creativity. The outputs were parsed and normalized. In a few cases, the models returned synonyms (e.g., "Customer Support" vs. our ground truth "Customer Service") or slightly differing aspect granularity (e.g., "Mobile App" vs. "App Performance"). We defined simple mappings so that, for evaluation, these would be considered correct matches.

Evaluation Metrics

We evaluated aspect extraction and sentiment classification jointly using pair-wise metrics. A predicted aspect–sentiment pair was considered correct if the element matched a gold aspect from the human labels, and the sentiment polarity was also correct. From these counts, we computed:

- **Precision**: the fraction of aspect–sentiment pairs output by the model that were correct. This measures how precise the model's extractions are (hallucinated or incorrect aspects reduce precision).
- **Recall**: the fraction of gold aspect–sentiment pairs that the model successfully extracted. This indicates how accurately the model captured information.
- **F1-Score:** the harmonic mean of precision and recall (F1 = 2·P·R/(P+R)). We report F1 as an overall measure of extraction quality. Importantly, this F1 is instance-based, not per-aspect category; it treats each pair in each review as a separate item to retrieve. This aligns with our goal, as missing an aspect in one review is just as much an error as missing it in another.

In addition to F1, we report an Accuracy metric. Here, we define accuracy as the percentage of sentiment labels that were correctly assigned to the aspects the model extracted. In other words, accuracy focuses on sentiment classification for a given element. We calculate it as: (number of extracted pairs with correct sentiment) / (number of extracted pairs). This is akin to precision, but specifically in the context of sentiment polarity assignment. We found it helpful to separate errors of aspect identification from errors of sentiment labeling. A model might correctly find an aspect but assign the wrong sentiment.

For completeness, we also measured the Exact Match Accuracy at the review level (did the model exactly reproduce the complete set of aspect-sentiment pairs for the review, with no extras or misses). This is a strict metric, and we report general trends rather than using it as the primary metric, since a single missed minor aspect would render the entire review a failure. Exact matches were highest for Claude (~78% of reviews were tagged), slightly lower for Gemini (~75%), and even lower for LLaMA (~70%), indicating that most reviews are handled well, with differences primarily evident in multi-aspect cases.

Efficiency Metrics

We tracked two key efficiency metrics during inference: latency and cost. Latency was measured as the end-to-end time to get a model response for one review (excluding network overhead as much as possible). We recorded the average latency per review by timing batches of requests. On an AWS C5 instance, using Bedrock synchronous calls, Claude Sonnet-4 averaged about 1.0 seconds per review (with std ~0.2s for typical length reviews), LLaMA-70B was around 0.8s/review, and Gemini Flash-Lite was fastest at approximately 0.5s s/review. These times include model processing of ~1,000 input tokens and generating ~50-100 output tokens. All three models demonstrated low variability in latency for inputs of similar size, which is expected since the input lengths were not varied significantly. It is worth noting that Claude's latency can be tuned by using its "near-instant" mode versus "extended reasoning" mode; our use of primarily straightforward extraction meant we effectively received near-instant-style responses.

Bedrock's infrastructure also offers a batch mode for Claude and LLaMA, which can amortize overhead when processing multiple inputs. This could reduce per-item latency and cost if we concatenated multiple reviews; however, we treated each one independently to match a real-time analysis scenario. For inference cost, our prompts averaged ~1000 tokens and generated around 60 tokens per review.

Additionally, as mentioned, batch processing on Bedrock can cut costs by up to 50% for large jobs by amortizing overhead. In a production setting, where analyzing millions of reviews is necessary, these features can be effectively utilized.

Finally, to ensure fairness, we did not count any manual post-processing or external sentiment tools. The evaluation code computed metrics by aligning model outputs to ground truth pairs, allowing for the small synonyms mapping mentioned

3. Results

The performance of each model on the annotated test set, in terms of Precision, Recall, F1-score for aspect-sentiment extraction, and sentiment Accuracy on extracted pairs. Table 1 visualizes the comparative Accuracy and F1 scores of the models.

Table 1: The Performance of Accuracy, Precision, and Recall

Model	Aspect-Sentiment Precision	Recall	F1-Score	Sentiment Accuracy	Inference Cost (USD)	Latency (sec/review

Gemini 2.5 Flash-Lite	0.87	0.82	0.85	0.90	~$0.98	~0.5
Claude Sonnet-4	0.91	0.87	0.89	0.92	~$30.6	~1.0
Meta LLaMA 70B	0.83	0.78	0.80	0.88	~$16.6	~0.8

As we can see, Claude Sonnet-4 achieved the highest F1-score. This indicates Claude was the most effective in extracting the correct aspect-sentiment pairs, likely owing to its larger model size or more extensive training in understanding context. Claude's higher precision (0.91) means it very rarely hallucinated or introduced incorrect aspects – most of what it output was valid. Its recall (0.87) was also highest, suggesting it found slightly more of the true aspects mentioned than the others. Gemini Flash-Lite also performed strongly, with an F1 score of ~0.85; its precision and recall were 0.87 and 0.82, respectively. In practical terms, Gemini might miss more than Claude (slightly lower recall), possibly skipping some subtle aspect mentions, and it had a few more mistaken extractions than Claude. LLaMA 70B trailed with F1 ~0.80, primarily due to lower recall (0.78). We observed that LLaMA sometimes missed aspects that were expressed in more implicit ways, or it would combine two related aspects into one. For example, in one review, a user mentioned issues with both the credit card and the money transfer features. LLaMA outputs a single aspect, "Banking transactions," with negative sentiment, whereas Claude and Gemini correctly separated "Card" (Negative) and "Transfer" (Negative). This counts as a recall miss for LLaMA since it did not explicitly list both.

In terms of sentiment accuracy, all models performed relatively well, indicating that when they identified an aspect, they usually got the sentiment right. Claude was best at 92%, followed by Gemini at 90%, and LLaMA at 88%. These differences are slight, all above a reasonable threshold for sentiment classification quality in practice. The more challenging aspect of the task was identifying the aspect, rather than determining whether something was praised or criticized. We noticed that when errors occurred in sentiment, it was often due to nuanced or mixed sentiments. For example, one review said (translated) "The new update is good (👍) in design, but made the login slower (👎)." All models captured both aspects (Design – positive, Login – negative). However, another review said, "The app is easy to use, but I am not happy with it lately," without specifying a reason. The models handled it differently: Claude inferred a generic "Overall" aspect (Negative), Gemini returned nothing (since no specific aspect was mentioned, arguably correctly), and LLaMA returned "Usability" as Positive

(misinterpreting "easy to use" as a standalone positive aspect). In our evaluation, we considered both "Overall" and no aspect as reasonable interpretations; however, LLaMA's output of "Usability: Positive" was deemed incorrect relative to the true intent. Such edge cases had a slightly greater impact on LLaMA's accuracy.

Overall, Claude's lead in F1 was modest, about 4 points higher than Gemini's. Considering Gemini is a much smaller model focused on speed, its performance reinforces findings that even distilled models can perform complex extraction with proper prompting. The difference between these models would likely widen on more complex tasks requiring reasoning, but for straightforward aspect mining, Gemini held up well. LLaMA's performance was slightly lower, which may be due to it not being as finely tuned to instructions as the others. Figure 2 shows the total inference cost and displays the average latency per review.

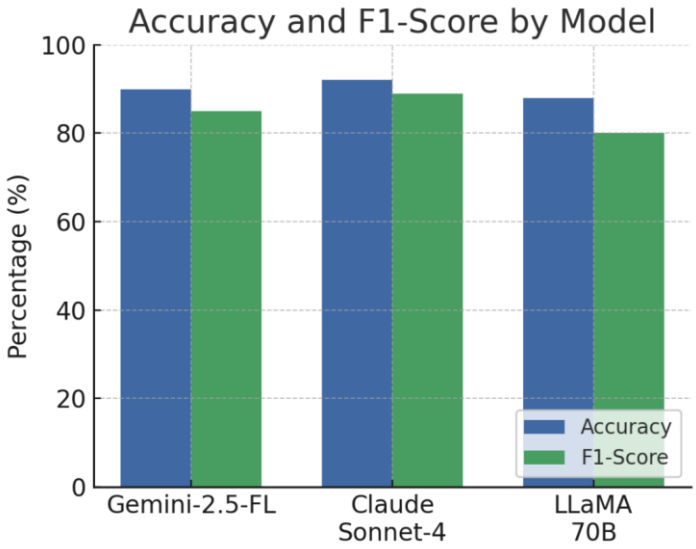

Figure 2: The Input Review and Expected Output in JSON format

4.1.Qualitative Analysis

We examined the sample outputs to identify where the models differed. In straightforward cases, e.g. "Harika bir uygulama, havale işlemleri çok hızlı." ("Wonderful app, transfer transactions are very fast.") – all models performed perfectly, outputting an aspect "Transfers" (or "Money Transfer") with Positive sentiment. For a more complex example: "Bildirimler faydalı ama uygulama bazen donuyor." ("Notifications are useful, but the app sometimes freezes."), as mentioned earlier, the

expected aspects are Notifications: Positive, App Performance: Negative. Here, all three models successfully extracted both aspects with correct sentiments. The Gemini output exactly matches the example we provided in its prompt, with {"Aspect": "Notifications", "Sentiment": "Positive"} and {"Aspect": "App Performance", "Sentiment": "Negative"}, which is reassuring. Claude did similarly and interestingly added a nuance: it originally wrote "App Performance/Stability" as the aspect, which is not wrong but verbose; our evaluation mapping counted that as correct for "App Performance". This shows Claude's tendency to be slightly more verbose or explanatory (perhaps an artifact of its training to be helpful), whereas Gemini was terse. LLaMA's output for this was also correct. However, it phrased it as "App speed," which we considered equivalent to "Performance." These differences did not affect scores after normalization. However, they indicate that the wording of the prompt influenced aspect naming. Since we requested concise English aspect names, none of the models returned Turkish words in the aspect field. Without that prompt constraint, early trials saw, for example, LLaMA returning "donma sorunu" ("freezing problem") as an aspect, which is understandable but less ideal for aggregation. This underscores a best practice: be explicit in the prompt about the format and language of outputs, especially in multilingual contexts.

We also encountered cases of partial aspect identification. The models responded: Claude perfectly identified both ("Card Payments" and "QR Payments", both Negative). Gemini identified only "Mobile Payment" (Negative), essentially merging the two issues into one aspect. LLaMA similarly gave one aspect, "Payment methods" (Negative). In the context of actionability, one could argue that the user's issue is broadly that the "payment functionality" is broken; however, since they specifically mentioned two methods, we expected to address two aspects. This contributed to Gemini and LLaMA's slightly lower recall. Notably, Claude's cost here is an order of magnitude higher; whether that granularity is worth the cost would depend on the use case.

4. Discussion and Conclusion

In this comparative study, prompt-based large language models proved effective for aspect-based sentiment analysis on real-world multilingual app reviews. Gemini, Claude, and LLaMA each demonstrated the capacity to extract aspect-level sentiments from Turkish texts with high accuracy despite the absence of task-specific or language-specific training. Claude Sonnet 4 on AWS Bedrock delivered the strongest overall performance and produced more comprehensive and precise extractions. This accuracy advantage was accompanied by substantially higher operational costs and longer

inference times. In contrast, Google Gemini 2.5 Flash Lite achieved results that were close to those of the best model, while operating at a small fraction of the cost and with markedly lower latency, which highlights the value of optimization for throughput. Meta LLaMA, with seventy billion parameters on AWS Bedrock, offered a middle ground, as its accuracy was adequate and it executed faster and more cost-effectively than Claude, making it attractive when balancing quality and expenditure.

In conclusion, AWS Bedrock offers powerful tools for enterprise sentiment analysis through Claude and LLaMA, each serving a distinct operational niche. For real-time and large-scale analytics where cost and speed are the primary constraints, LLaMA is well-suited and accurate enough to drive dashboards and alerts about customer pain points at low cost. When deeper accuracy is required or when linguistic complexity is high, Claude Sonnet 4 is an appropriate choice that delivers near-human nuance in sentiment interpretation at the expense of higher compute.

A hybrid strategy can be adopted, in which LLaMA handles the bulk of routine analysis, while Claude is allocated to complex cases or periodic deep dive reports. The comparative evidence indicates that no single model dominates across all dimensions of accuracy and efficiency.

Maintaining both Claude and LLaMA within the same toolkit enables organizations to tailor their pipelines to their specific requirements and timelines. The outcome is an actionable aspect-level customer insight that supports timely and cost-effective product and service improvements.

As large language model technology advances, these systems are expected to achieve higher accuracy at lower cost. The procedures and findings reported here can serve as a baseline for subsequent evaluations and for future production deployments.

References

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," Proceedings of the 8th International Workshop on Semantic Evaluation, pp. 27–35, 2014.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[3] D. D. Lee and H. S. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," Nature, vol. 401, pp. 788–791, 1999.

[4] C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, and Y. Xue, "Evaluating Zero-Shot Multilingual Aspect-Based Sentiment Analysis with Large Language Models," arXiv preprint arXiv:2412.12564, 2024.

[5] P. F. Simmering, R. Werkmeister, and L. Di Stasio, "Large Language Models for Aspect-Based Sentiment Analysis," arXiv preprint arXiv:2310.18025, 2023.

[6] M. Água, P. Pina, and B. Ribeiro, "Large Language Models Powered Aspect-Based Sentiment Analysis for Enhanced Customer Insights," Tourism and Management Studies, vol. 21, 2025.

[7] J. Šmíd, M. Bělohlávek, and T. Brychcín, "LLaMA-Based Models for Aspect-Based Sentiment Analysis," Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 66–78, 2024.