Cross-Provider Consistency of LLM-Based Semantic Enrichment for e-Science Workflows

Tiago de Melo

Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brazil

tmelo@uea.edu.br

Abstract. *Integrating large language models (LLMs) into data-intensive scientific workflows requires understanding their robustness, efficiency, and cross-provider consistency. This paper evaluates five LLMs (ChatGPT, Gemini, DeepSeek, Mistral, and Claude) as interchangeable components of a semantic enrichment pipeline, using 2,387 Billboard Year-End Hot 100 songs (2000–2023) as a case study. All models had near-perfect success rates (99.7–100%) but over 20-fold cost differences ($0.46–$10.70). Thematic agreement was substantial (Cohen's Kappa = 0.70–0.81), and temporal trajectories were consistent across providers (Pearson r = 0.65–0.82). Thus, provider choice minimally affects aggregate analytical conclusions, allowing workflow designs that prioritize cost or latency without sacrificing reproducibility.*

1. Introduction

The fourth paradigm of science, data-intensive discovery, has transformed how researchers analyze large-scale textual collections across the humanities, social sciences, and cultural analytics. In this context, song lyrics are particularly valuable, as they encode emotional expression, personal narratives, and cultural preferences in a form amenable to longitudinal analysis [Watanabe and Goto 2020, Choi et al. 2014, Papazoglou and Gaizauskas 2021, Mihalcea and Strapparava 2012].

Transforming large lyric collections into structured semantic representations remains challenging. Standard NLP methods designed for prose text are not always effective for lyrics, which have unique linguistic properties, such as emotional, concise, and intentionally obscure expressions constrained by melody and rhyme [Watanabe and Goto 2020]. Traditional approaches rely on topic modeling and text mining techniques that often require substantial manual interpretation and may produce latent topics with limited semantic transparency [Hu et al. 2026]. These limitations are particularly relevant in e-Science scenarios, where the objective is to build semantic enrichment workflows that are robust, scalable, reproducible, and transferable across datasets and application domains.

Large language models (LLMs) offer a promising alternative for data-intensive scientific workflows. Rather than inferring themes from lexical co-occurrence alone, LLMs can directly assign interpretable labels to documents, enabling systematic aggregation and cross-model comparison. However, integrating LLMs as components within computational workflows raises questions that extend beyond semantic plausibility. Specifically, it is necessary to assess the robustness of different models across large numbers of instances, to characterize the associated latency and cost trade-offs, and, critically for reproducibility, to determine the extent to which models from different providers yield comparable conclusions.

This paper addresses these questions through a comparative evaluation of five LLMs in a large-scale semantic enrichment pipeline. We analyze 2,387 songs from the

Billboard Year-End Hot 100 chart (2000–2023), using a fixed thematic taxonomy and shared prompting strategy to compare ChatGPT, Gemini, DeepSeek, Mistral, and Claude. The longitudinal corpus provides temporal depth to investigate thematic evolution while enabling evaluation of aggregate pattern stability across models.

We investigate four research questions: (**RQ1**) How robust are different LLMs when used for large-scale thematic enrichment? (**RQ2**) What are the latency and cost trade-offs across LLMs? (**RQ3**) To what extent do different LLMs produce similar thematic distributions? (**RQ4**) How consistently are temporal thematic patterns identified across models?

Our evaluation reveals that all five models achieved near-perfect execution rates (99.7–100%), demonstrating strong operational robustness. Latency and cost varied substantially, with Mistral offering the best cost-performance balance (total cost under $0.50) and Gemini the lowest latency (under 1 second per song). Thematic consistency was high across providers: all models identified the same dominant theme, with song-level Cohen's Kappa ranging from 0.70 to 0.81. Temporal trajectories were also consistent, with mean Pearson correlations between 0.65 and 0.82, suggesting that provider choice does not substantially affect analytical conclusions.

The contribution is twofold. Operationally, we provide a systematic comparison of LLMs, examining robustness, latency, and cost, which enables informed workflow design decisions. Analytically, we investigate whether different providers yield convergent thematic interpretations and stable temporal trajectories, a key requirement for reproducible scientific findings. The paper uses a culturally meaningful dataset[1] as a case study to examine how LLM-based semantic enrichment can support reliable e-Science workflows.

2. Related Work

The integration of LLMs into research pipelines has raised questions about robustness, reproducibility, and scalability. Studies show that LLM outputs exhibit task-dependent consistency, with binary classification achieving near-perfect reproducibility while complex tasks show greater variability [Wang and Wang 2025]. Comparative evaluations reveal meaningful differences across providers in precision, recall, latency, and cost [Gilardi et al. 2023, Guo et al. 2024]. However, most studies focus on benchmark datasets with ground-truth labels rather than open-ended semantic enrichment.

Song lyrics have served as a testbed for NLP methods in cultural analytics. Prior work employed topic modeling to uncover thematic evolution [Hunke et al. 2025] and sentiment analysis to track emotional trends [Foramitti et al. 2025]. Studies on Billboard charts examined lyrical novelty and cultural dynamics [Askin and Mauskapf 2017]. While valuable, these approaches typically rely on unsupervised methods that may produce latent topics with limited semantic transparency.

Few studies have systematically evaluated whether different LLMs produce convergent semantic interpretations on the same large-scale corpus. Existing benchmarks assess instance-level accuracy but do not examine whether aggregate patterns and temporal trends remain stable across providers. This could be considered a critical gap for e-Science applications where reproducibility is paramount. Our work addresses this gap by comparing five LLMs under identical conditions, evaluating both operational trade-offs and analytical consistency.

[1]Dataset available at `https://github.com/tmelo-uea/billboard-hot100-lyrics`.

3. Methodology

3.1. Data Collection

We collected songs from the Billboard Year-End Hot 100 chart for each year between 2000 and 2023. Lyrics were retrieved through the Genius API, and only English-language songs were retained for linguistic consistency. The final corpus contains 2,387 songs (13 entries were unavailable due to missing data in certain years).

3.2. Evaluated Models

We selected five LLMs from different providers to ensure diversity across architectural families (Table 1). All models were accessed through their official APIs during March 2026 using default temperature settings. Selection criteria prioritized models offering structured JSON output, commercial API accessibility, and distinct provider ecosystems. Notably, ChatGPT (`gpt-5-mini`) applied low-effort reasoning by default, as evidenced by its API response metadata (`reasoning_tokens` > 0); all other models were queried without explicit reasoning activation.

Table 1. Large language models evaluated in this study.

Model	Provider	Model Identifier	Param.	Reasoning
ChatGPT	OpenAI	`gpt-5-mini`	Undisclosed	Yes (low)
Gemini	Google	`gemini-2.5-flash`	Undisclosed	No
DeepSeek	DeepSeek	`deepseek-chat`	$\sim$671B*	No
Mistral	Mistral AI	`mistral-small-2603`	$\sim$24B	No
Claude	Anthropic	`claude-sonnet-4-6`	Undisclosed	No

* Mixture-of-Experts; $\sim$37B active per token.

Experiments were run on Google Colaboratory (Linux, Python 3.12, 2 CPU cores, 12.67 GB RAM). All models were accessed via cloud APIs, so no GPU was used. Each song was processed once without retries; responses failing JSON schema validation were counted as failures. No automated JSON repair or post-processing was applied.

3.3. Prompt Design

To ensure comparability, we employed a unified prompt design. Each model processed one song at a time and returned structured JSON containing: (i) a free-text thematic field (`main_theme`), (ii) a categorical field from our predefined taxonomy (`normalized_theme`), (iii) three keywords, (iv) a brief description (up to 12 words), (v) a sentiment label, and (vi) a confidence score (0.0–1.0). The confidence score records each model's self-reported certainty in its thematic assignment; we collect it for completeness and potential downstream filtering, but it is not analyzed in this study. Figure 1 shows the system prompt.

The normalized taxonomy, informed by prior work [Watanabe and Goto 2020, Choi et al. 2014, Papazoglou and Gaizauskas 2021, Mihalcea and Strapparava 2012], comprises ten categories: *love_relationships*, *heartbreak_loss*, *sex_desire*, *self_reflection_growth*, *empowerment_confidence*, *party_celebration*, *lifestyle_success*, *social_commentary*, *faith_spirituality*, and *other*. The residual *other* category accommodates songs that do not fit the predefined labels, reducing the risk of forced assignments.

System Prompt

You are an expert analyst of song lyrics and cultural themes. Analyze ONE song lyric and return ONLY valid JSON.

Rules

1. Read lyric as a whole
2. `main_theme`: free-text
3. `normalized_theme`: from list
4. `keywords`: exactly 3
5. `theme_description`: $\leq$12 words
6. `sentiment`: pos — neg — neu — mix
7. `confidence`: 0.0–1.0
8. Return ONLY valid JSON

Output Schema

```
{
 "main_theme": "string",
 "normalized_theme": "string",
 "keywords": ["...","...","..."],
 "theme_description": "...",
 "sentiment": "pos",
 "confidence": 0.85
}
```

Figure 1. System prompt used for thematic enrichment.

3.4. Evaluation Metrics

RQ1 (Robustness): Success rate, defined as the proportion of songs returning valid JSON conforming to the expected schema. **RQ2 (Efficiency):** Average and median response times per song, average cost per song, and total execution cost based on each provider's API pricing. **RQ3 (Agreement):** Evaluated at three levels: (i) aggregate distributions showing percentage of songs per theme; (ii) pairwise divergence using Jensen-Shannon Divergence (JSD) and L1 distance; (iii) song-level agreement using exact match percentage and Cohen's Kappa [Cohen 1960]. **RQ4 (Temporal stability):** Yearly time series of thematic proportions (2000–2023), with cross-model consistency quantified through Pearson correlation between model-specific trajectories.

4. Results

4.1. RQ1 – Robustness

All models showed strong robustness. ChatGPT, DeepSeek, Mistral, and Claude achieved 100% success rates. Gemini reached 99.71% (7 failures out of 2,387), with failures involving songs with highly explicit sexual lyrics, such as *Peaches & Eggplants*, *Throat Baby (Go Baby)*, and *Freek-a-Leek*, suggesting, but not confirming, that explicit content may have contributed to these isolated issues. These results demonstrate that execution robustness is not a limiting factor for using the five evaluated LLMs in large-scale thematic analysis.

4.2. RQ2 – Latency and Cost Trade-offs

Table 2 shows substantial operational differences. Mistral and Gemini were the fastest (average latencies under 1 second), while Claude, DeepSeek, and ChatGPT required 3.0–3.7 seconds per song. Cost differences were stark. Processing the entire corpus cost \$0.46 for Mistral versus \$10.70 for Claude. Mistral offered the best cost-performance balance, while Gemini excelled in latency.

4.3. RQ3 – Thematic Distribution Agreement

Figure 2 shows thematic distributions across models. All five identify *love_relationships* as predominant (26.9%–33.0%), indicating robust preservation of the corpus's overall thematic structure. Secondary themes showed more variation: ChatGPT and Gemini ranked *sex_desire* second, while Claude, DeepSeek, and Mistral assigned this position to

Table 2. Latency and cost comparison across models.

Model	Avg. Lat. (s)	Med. Lat. (s)	Avg. Cost	Total Cost
ChatGPT	3.73	3.45	$0.00069	$1.64
Gemini	0.96	0.85	$0.00050	$1.20
DeepSeek	3.73	3.73	$0.00031	$0.75
Mistral	0.89	0.85	$0.00019	$0.46
Claude	3.04	2.83	$0.00448	$10.70

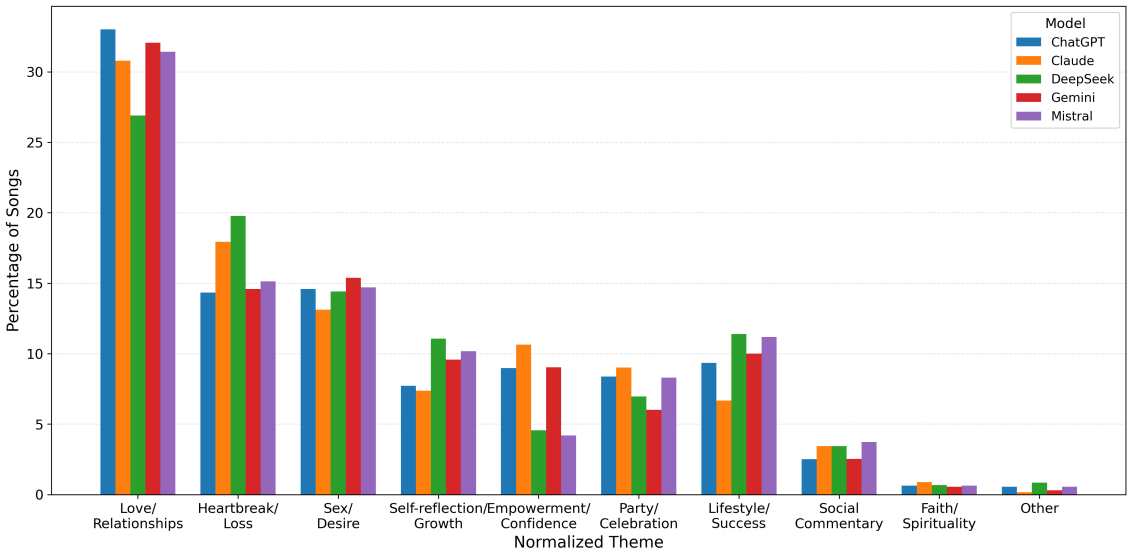

Figure 2. Percentage distribution of normalized themes across models.

heartbreak_loss. The *other* category remained below 1% for all models, supporting the taxonomy's adequacy.

As shown in Tables 3 and 4, all ten pairwise comparisons among the five evaluated models were analyzed. Taken together, these complete pairwise results indicate that inter-model convergence is high at both the distributional and song levels. ChatGPT and Gemini were most similar (JSD = 0.0027), while Claude and DeepSeek showed greatest divergence (JSD = 0.0205). Song-level agreement was substantial across all pairs, with exact agreement ranging from 75.6% to 84.4% and Cohen's Kappa from 0.70 to 0.81 (Table 4). Notably, the three highest Kappa values involve ChatGPT, Gemini, and Claude, indicating that these models form the most semantically aligned group. These results demonstrate that, across the five evaluated models, LLMs yield broadly consistent thematic interpretations at both aggregate and instance levels, providing a reliable basis for temporal analysis.

4.4. RQ4 – Temporal Thematic Patterns

Figure 3 shows the temporal evolution of ten thematic categories. All models follow similar trajectories, indicating consistent trend identification. *Love_relationships* remains dominant, while *heartbreak_loss* rises in later years. *Lifestyle_success* becomes more salient from the late 2010s, and *party_celebration* peaks in the early 2010s. *Sex_desire* and *social_commentary* vary more across models but retain similar temporal shapes.

Cross-model temporal consistency was quantified through mean Pearson corre-

Table 3. Pairwise similarity.

Model 1	Model 2	JSD	L1
ChatGPT	Gemini	0.0027	0.0731
DeepSeek	Mistral	0.0044	0.1290
ChatGPT	Claude	0.0057	0.1416
Claude	Gemini	0.0090	0.1841
ChatGPT	Mistral	0.0093	0.1290
Gemini	Mistral	0.0094	0.1231
DeepSeek	Gemini	0.0125	0.2119
ChatGPT	DeepSeek	0.0148	0.2421
Claude	Mistral	0.0185	0.2044
Claude	DeepSeek	0.0205	0.2447

Table 4. Song-level agreement.

Model 1	Model 2	Exact	κ
ChatGPT	Gemini	84.37	0.8094
Claude	Gemini	83.53	0.8004
ChatGPT	Claude	83.16	0.7955
DeepSeek	Gemini	80.63	0.7670
Claude	DeepSeek	80.56	0.7671
ChatGPT	DeepSeek	79.43	0.7527
DeepSeek	Mistral	78.93	0.7465
Gemini	Mistral	77.61	0.7282
ChatGPT	Mistral	77.34	0.7247
Claude	Mistral	75.58	0.7052

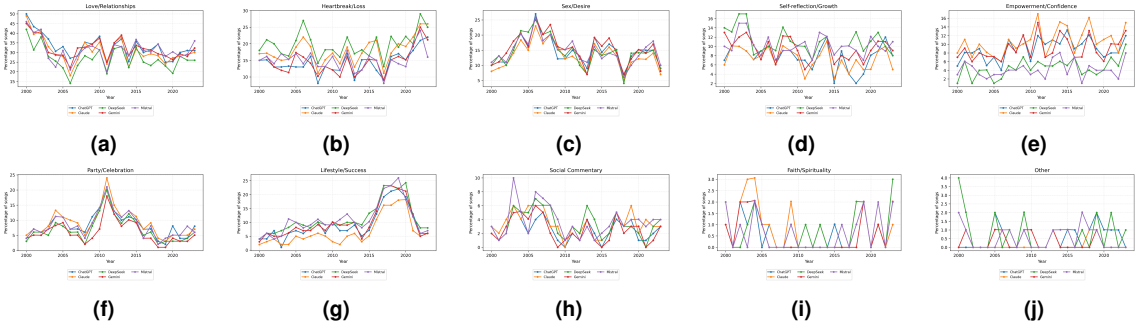

Figure 3. Temporal evolution of the normalized thematic categories across the five evaluated models from 2000 to 2023: (a) Love/Relationships, (b) Heartbreak/Loss, (c) Sex/Desire, (d) Self-reflection/Growth, (e) Empowerment/Confidence, (f) Party/Celebration, (g) Lifestyle/Success, (h) Social Commentary, (i) Faith/Spirituality, and (j) Other.

lations between model-specific trajectories. All pairwise correlations were positive and moderately high (0.65–0.82). The strongest alignment was between Claude and Gemini ($r = 0.82$), followed by ChatGPT–Claude and ChatGPT–Gemini (both $r = 0.79$). DeepSeek and Mistral showed slightly lower correlations ($r = 0.65$–0.75), but all values remained substantial, reinforcing the reliability of LLM-based temporal analysis for the evaluated models.

5. Discussion

Table 5 summarizes each model's strengths. Mistral is the fastest and cheapest, ChatGPT, Gemini, and Claude show the strongest thematic agreement, and Gemini and Claude lead in temporal consistency, informing workflow design based on project priorities. Our findings have important implications for e-Science workflows that employ large language models (LLMs) as semantic enrichment components. The high level of cross-model agreement (Cohen's Kappa = 0.70–0.81) indicates that scientific conclusions are largely reproducible across different providers. This property is particularly critical for the so-called fourth paradigm of science, in which computational and data-intensive workflows must yield consistent results independent of specific infrastructure or platform choices.

The more than 20× cost gap between Mistral and Claude shows that operational efficiency must be weighed alongside analytical quality. In resource-constrained settings, cheaper models can offer similar interpretations at much lower cost. The fixed taxonomy, structured JSON output, and unified prompting methodology can be applied to other domains (e.g., scientific abstracts, news corpora), enabling robust e-Science pipelines

Table 5. Summary of model strengths across evaluation dimensions.

Model	Robustness	Latency	Cost	Agreement	Temporal
ChatGPT	✓			✓	
Gemini		✓		✓	✓
DeepSeek	✓				
Mistral	✓	✓	✓		
Claude	✓			✓	✓

✓ indicates best or near-best performance in each dimension. For Temporal, ✓ marks the models in the highest-correlation pair (Claude–Gemini, $r = 0.82$).

beyond cultural analytics. However, variation in secondary theme assignments suggests that fine-grained semantic distinctions may depend more on the provider. Researchers needing precise category boundaries should use ensemble methods or human validation.

6. Limitations

This study has several limitations. The corpus includes only English-language lyrics from a single commercial chart, limiting generalizability to other languages, genres, and scientific texts. The ten-category taxonomy, adapted from popular-music research, may not transfer to other domains. Each model was queried once with a fixed prompt; we did not test prompt variations or repeated runs, so individual-level classification stability remains partly unknown. Provider updates may alter model behavior without changing version identifiers, hindering reproducibility. Our metrics capture inter-model consistency, not semantic validity: without a human-annotated gold standard, we cannot verify the correctness of themes or detect hallucinated labels. Temporal consistency is evaluated only with Pearson correlation; rank-based or trend-shape measures might provide a richer characterization.

7. Conclusion

This paper evaluated five LLMs as semantic enrichment components in a large-scale e-Science workflow using a corpus of 2,387 Billboard songs released between 2000 and 2023. Across providers, execution success rates were consistently high (99.7–100%), indicating that all models were operationally reliable for the proposed workflow. At the same time, latency and monetary cost varied substantially, with Mistral providing the most favorable cost–performance trade-off and Gemini achieving the lowest average latency.

Despite these operational differences, the models produced broadly consistent analytical outputs. All providers identified the same dominant theme in the corpus, pairwise song-level agreement remained substantial (Cohen's Kappa = 0.70–0.81), and temporal patterns were moderately to strongly correlated across models (Pearson r = 0.65–0.82). Taken together, these results suggest that, in this case study, provider choice has limited impact on aggregate analytical conclusions, even though model-specific differences remain visible at the level of individual song classifications.

The findings support using LLMs for semantic enrichment in data-intensive scientific workflows, particularly for extracting large-scale thematic patterns under cost and time constraints. However, the study measures consistency across providers rather than semantic validity against a human gold standard. Future work should extend the analysis to multilingual corpora, add expert annotations, test sensitivity to prompt design and repeated

runs, and assess whether cross-provider stability generalizes to other scientific domains and workflow settings.

Acknowledgments

The author acknowledges the support provided by the INCT–TILD–IAR, funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Askin, N. and Mauskapf, M. (2017). What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944.

Choi, K., Lee, J. H., and Downie, J. S. (2014). What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. In *IEEE/ACM joint conference on digital libraries*, pages 453–454. IEEE.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Foramitti, M., Nater, U. M., Lamm, C., and Martins, M. (2025). Societal crises disrupt long-term increases in stress, negativity, and simplicity in us billboard song lyrics from 1973 to 2023. *Scientific Reports*, 15(1):41733.

Gilardi, F., Alizadeh, M., and Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Guo, Y., Ovadje, A., Al-Garadi, M. A., and Sarker, A. (2024). Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association*, 31(10):2181–2189.

Hu, Q. Q., Azmi-Murad, M. A. M., Azman, A. B., and Nasharuddin, N. A. (2026). Revisiting the role of lyrics in music emotion recognition: A critical analysis of the semantic–perceived emotion gap and methodological challenges. *Information Fusion*, page 104303.

Hunke, T., Huber, F., and Steffens, J. (2025). The evolution of song lyrics: An nlp-based analysis of popular music in germany from 1954 to 2022. *Music & Science*, 8:20592043251331155.

Mihalcea, R. and Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 590–599.

Papazoglou, V. and Gaizauskas, R. (2021). Using listeners' interpretations in topic classification of song lyrics. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, pages 22–26.

Wang, J. J. and Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *arXiv preprint arXiv:2503.16974*.

Watanabe, K. and Goto, M. (2020). Lyrics information processing: Analysis, generation, and applications. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 6–12.