The Flow of LLM-Generated Poetry

Patrícia Ferreira
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
patriciaf@dei.uc.pt

Ana Alves
University of Coimbra, CISUC/LASI, DEI
Polytechnic Institute of Coimbra
ana@dei.uc.pt

Catarina Silva
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
catarina@dei.uc.pt

Hugo Gonçalo Oliveira
University of Coimbra, CISUC/LASI, DEI
DEI, University of Coimbra
hroliv@dei.uc.pt

Abstract

We demonstrate how dialogue flows can be adopted for representing transitions between speech acts and their probabilities in sets of generated poems, contributing to the analysis of the narrative structure. By considering poems generated by different Large Language Models (LLMs), open and proprietary, we take conclusions on the impact of size, and focus on the temperature hyperparameter, known as a proxy for creativity. Flow analysis revealed that different LLMs prefer different speech acts and confirmed that higher temperatures increase the set of used transitions, especially for the open models.

Introduction

The flexibility of general-purpose instruction-tuned Large Language Models (LLMs) makes them a convenient tool for zero- and few-shot generation of creative textual artefacts, including poems (Sawicki et al. 2023; Qu, Yuan, and Färber 2026). From natural language instructions, LLMs can produce text that reflects a given intent while exhibiting poetic features, such as a regular form or usage of figurative language. However, deeper analysis noted that there are limitations on meeting a specific number of characters per line (Yu et al. 2024) or of lines and syllables per line (Agirrezabal and Gonçalo Oliveira 2024); as well as on diversity, with some models preferring specific words, meanings or forms (Chen et al. 2024; Walsh, Preus, and Gronski 2024).

Although previous works capture quantifiable aspects of poetry, they prioritize surface-level patterns, highly-relevant in poetry, over the more abstract narrative. To uncover this structural aspect, we propose to explore the construct of dialogue flows and adapt them to represent the flow of speech acts in LLM-generated poems. Several works rely on flow induction from human dialogues where the identification of behavioral patterns contributes to the development of dialogue systems (Martinez and Nugent 2022) or improvements in customer-support (Ferreira et al. 2024). Recently, flow discovery was also applied for analyzing LLM reasoning patterns (Matos, Silva, and Gonçalo Oliveira 2025). We argue that such flows are a useful tool for studying the diversity of poetry at a more abstract level.

To test this, flows are extracted from English sonnets generated by open and proprietary LLMs of different sizes. To enable this, each line of a sonnet is labeled with a poetry-oriented speech act, allowing these sequences to characterize the flow of each LLM.

Moreover, to revisit previous work on the relation between the temperature hyperparameter and creativity (Peeperkorn et al. 2024), poems are generated with two different values of temperature, and its impact on the structural organization of generated poems is examined. While creativity cannot be reduced to diversity alone, we focus on structural diversity as one of its relevant dimensions. Our main contributions are:

- An adaptation of dialogue flows for representing the sequence of speech acts in poetry, thus enabling a rich analysis of the underlying sequence of actions;
- The application of the previous to sets of poems by different LLMs, which confirms that flows are a visual means for analyzing how structural diversity is affected by features like model size and temperature.

This aims to broaden the evaluation of AI-generated poetry, offering a new lens to understand how LLMs navigate the complex balance between structural flow and creativity.

Speech Act Flows

We propose to model the narrative structure of poems by flows connecting speech acts, i.e., directed graphs $G = (N, E)$ where: the set of nodes N includes the possible speech acts, augmented by nodes to mark the start (SOP) and end of poem (EOP); the set of edges E represents the transitions $n_i \rightarrow n_j : n_i, n_j \in N$, meaning that a line labeled with speech act n_j follows another labeled with n_i.

In a set of lines, we can compute the transition probability $P(n_j \mid n_i)$, which quantifies the likelihood of moving from speech act n_i to n_j. If all the poems have the same number of lines, the flow can be organized according to line number. Highlighting frequent speech acts per line provides a visual cue for both common sequences and diversity.

A critical step is the classification of the speech act (Austin 1975), which can be defined as a general action performed by the speaker of an utterance. Different speech act taxonomies exist but, in order to find one more aligned with poetry and, particularly, with our data, we defined one with the help of LLMs. Specifically, we used a random sample of 200 lines from the poems in our dataset (see next section) and prompted GPT-4o to propose the ideal semantic

categories for classification. The zero-shot prompt used for this discovery is in the Appendix (Fig. 3).

The taxonomy includes five categories: **Dynamics (DYN)**, for dynamic movement, actions, or specific intentions performed by an agent; **Evaluation (EVAL)**, for explicit positive assessment that emphasizes greatness, superiority, or enduring impact; **Figuration (FIG)**, for figurative transformations, such as metaphors, similes, or symbolic equivalences; **Representation (REP)**, for concrete depiction of physical settings, landmarks, or objects, focusing on sensory or spatial details; and **Subjectivity (SUBJ)**, for expression of affective states (e.g., joy, sorrow) and the revelation of the entity's inner life or psychological depth.

Since taxonomy induction and annotation rely on GPT-4o, some annotation bias may exist. However, the adopted categories are intentionally high-level and model-agnostic. Table 1 shows the taxonomy in action, with one of the previous acts assigned to each line of a poem in our dataset.

Lines	Act
Stone whispers tales of ages long since past,	FIG
A Seine-kissed city, bathed in golden hue,	FIG
Where lovers stroll and shadows softly cast,	REP
And history breathes in every avenue.	REP
The Eiffel Tower, reaching for the sky,	REP
A wrought-iron dream against the fading light,	FIG
While Notre Dame, with stories in her eye,	FIG
Remembers echoes of a sacred rite.	SUBJ
Cafés hum with chatter, sweet and low,	REP
And artists paint on canvases so bold,	DYN
A symphony of life begins to flow,	FIG
In cobbled streets, a story to unfold.	REP
The scent of bread, a promise warm and true,	REP
Parisian magic, ever fresh and new.	EVAL

Table 1: Sonnet about Paris by Gemma-3-27B, $T = 0.8$, with each line classified.

Experimental Setup

Our main goal is to use speech act flows in the analysis of LLM-generated poems, and further take conclusions on their structural diversity. For this, a dataset was created with poems generated by different LLMs, and complementary metrics were defined.

Models

The LLMs selected for our study represent diverse sizes and access types. They include two proprietary models by OpenAI (GPT-4o and GPT-4o-mini), both accessed through their official API[1]; and two open models by Google (Gemma-3-4B and Gemma-3-27B), both accessed through Ollama[2], in Q4_K_M quantization, running in a local machine equipped with an NVIDIA RTX A6000 GPU (48GB VRAM).

Data

The selected LLMs were instructed to generate sonnets in English about a target entity, using the zero-shot prompt in Appendix (Fig. 2). To cover a broad, but controlled, number of targets, 15 entities were selected across three domains: (people in) Sports, (people in) Music, and Cities. We selected globally-recognized entities based on objective and renowned rankings, namely: Forbes' highest-paid athletes[3], Spotify's top-ranked artists[4], and the Global Cities Index by Oxford Economics[5]. The complete list is in Table 2.

Domain	Target entities
Sports	Cristiano Ronaldo, Stephen Curry, Tyson Fury, Dak Prescott, Lionel Messi
Music	Bad Bunny, Taylor Swift, The Weeknd, Drake, Billie Eilish
Cities	New York, London, Paris, San Jose, Seattle

Table 2: Selected target entities across the three domains.

Moreover, poems were generated using two opposing temperature values (T) to evaluate their impact. We used $T = 0.0$ to capture the model's standard and more deterministic behavior, and $T = 0.8$ to force higher diversity.

Generating 10 poems for each of the 15 entities across two temperatures and four models resulted in a total of 1,200 sonnets, which is publicly available at `https://tinyurl.com/rspz2h66`. Since they consist exclusively of sonnets, depicted flows will have 14 levels, mapping the exact progression of the poem from start to finish.

To classify each line into one of the five speech acts, we resorted to GPT-4o, with $T = 0.0$. The prompt used for this is in the Appendix (Fig. 4). Out of curiosity, Table 3 details the speech act distribution per LLM, entity domain and temperature.

We immediately see that the distribution is driven by the domain of the target entity. For example, DYN acts are the most common in poems about *Sports*, naturally reflecting movement and action. Conversely, poems about *Cities* prioritize REP to describe physical settings and landmarks. The *Music* domain elicits higher rates of SUBJ and FIG.

Furthermore, the analysis confirms that each LLM has its own stylistic signature (Walsh, Preus, and Gronski 2024). The smaller Gemma-3-4B diverges notably from the others, relying heavily on SUBJ across all domains (over 61% in *Music*), whereas the GPT models demonstrate a robust preference for figurative language (FIG). Interestingly, increasing the temperature from $T = 0.0$ to $T = 0.8$ produces only marginal fluctuations in these frequencies. This suggests that, while temperature might alter the sequence of acts, the baseline semantic profile is an intrinsic property of each LLM, remaining robustly stable.

Metrics

To quantify the structural complexity, semantic focus, and generation stability of the poems across different model sizes and temperatures, we employ three sets of metrics.

Graph Metrics inform about structural coverage and path complexity, i.e., number of explored states. They include the total number of unique nodes and directed transitions in the speech act flows.

[1] `https://openai.com/`
[2] `https://ollama.com/`
[3] `https://www.forbes.com/lists/athletes/`
[4] `https://tinyurl.com/yke49wkt`
[5] `https://tinyurl.com/4k94xb55`

Model	Domain	T	DYN	EVAL	FIG	REP	SUBJ
Gemma-3-4B	Sports	0.0	28.7	23.3	10.0	3.0	35.0
		0.8	26.9	21.7	12.3	6.6	32.6
	Music	0.0	5.7	8.7	13.0	11.0	61.6
		0.8	5.6	8.0	15.6	9.4	61.4
	Cities	0.0	4.0	10.0	12.1	49.3	24.6
		0.8	5.0	6.4	9.6	52.4	26.4
Gemma-3-27B	Sports	0.0	52.1	15.3	13.3	9.6	9.7
		0.8	49.0	20.0	12.9	8.6	9.6
	Music	0.0	32.9	14.3	30.3	3.1	19.4
		0.8	25.9	15.0	30.0	5.9	23.3
	Cities	0.0	27.1	8.4	24.4	35.7	4.3
		0.8	22.7	7.6	28.3	35.9	5.6
GPT-4o-mini	Sports	0.0	48.6	23.4	17.4	2.6	8.0
		0.8	45.7	21.9	20.8	1.9	9.7
	Music	0.0	30.6	6.6	36.9	2.0	24.0
		0.8	28.9	8.9	36.0	2.1	24.1
	Cities	0.0	20.1	1.9	34.0	32.6	11.4
		0.8	16.9	4.0	34.6	33.4	11.1
GPT-4o	Sports	0.0	38.4	33.4	17.9	1.3	9.0
		0.8	41.3	29.3	17.0	2.6	9.9
	Music	0.0	24.4	9.9	40.4	3.9	21.4
		0.8	25.7	8.9	39.3	4.3	21.9
	Cities	0.0	15.6	7.3	30.0	39.4	7.7
		0.8	13.9	5.7	31.6	40.4	8.4

Table 3: Distribution of speech acts (%) across models, domains, and temperatures (T).

For the **Semantic Metrics**, the poems and the entity names are embedded with `all-MiniLM-L6-v2`[6]. *Topic Similarity* measures the cosine between the target entity and each generated poem. *Stability* aims to assess how different poems generated with the same model, temperature and target vary among them. It is given by the cosine between a randomly selected prototype against the remaining poems for the same entity, following the exemplar-based evaluation methodology of (Peeperkorn et al. 2024).

As a **Divergence Metric**, the Mean Absolute Difference (MAD) measures the impact of temperature as the difference between the probability transition matrices with $T = 0.0$ and $T = 0.8$, scaled by a factor of 100 for readability.

Results and Analysis

To analyze the structural progression of the generated poems, we examine the distribution of speech acts across the 14 lines. Figure 1 (a–d) illustrates the flows for all the poems about *Cities* by Gemma-3-27B and GPT-4o, with $T = 0.0$ and $T = 0.8$. Flows for all LLMs and domains are in Appendix (Figs. 5 and 6). Node colors denote speech acts, from SOP to EOP, and directed edges represent transitions with variable thickness and darkness, depending on probability.

With $T = 0.0$, Gemma exhibits a nearly linear structure with restricted transitions. As temperature increases, the flow becomes more dispersed, showing greater sequential variation. In contrast, GPT-4o generates a dense and diverse structure even with $T = 0.0$. Increasing to $T = 0.8$ adds only subtle variations, keeping the base flow intact.

These flows provide a visual anchor for the statistical distributions reported in Table 3. For *Cities*, the REP act (blue) serves as the foundational backbone across all 14 levels. This is evidenced by the primary transition paths, represented by thicker edges, and suggests that the domain's descriptive nature acts as a central narrative axis, maintaining a stable core even when higher temperature introduces a wider variety of alternative paths. To quantify these observations, Table 4 reports the metrics described in the previous section.

[6] `https://tinyurl.com/5cdp746d`

Model (T)	#N	#E	Topic Sim	Stability	MAD
Gemma-3-4B (0.0)	67	167	0.32 ± 0.11	0.90 ± 0.07	6.10
Gemma-3-4B (0.8)	72	302	0.30 ± 0.11	0.79 ± 0.07	
Gemma-3-27B (0.0)	63	137	0.31 ± 0.10	1.00 ± 0.00	5.85
Gemma-3-27B (0.8)	72	300	0.31 ± 0.10	0.82 ± 0.06	
GPT-4o-mini (0.0)	72	258	0.41 ± 0.09	0.87 ± 0.07	4.32
GPT-4o-mini (0.8)	72	292	0.39 ± 0.07	0.74 ± 0.06	
GPT-4o (0.0)	71	278	0.45 ± 0.09	0.87 ± 0.05	3.38
GPT-4o (0.8)	72	299	0.41 ± 0.07	0.76 ± 0.04	

Table 4: Metrics for the analyzed models across all domains at $T = 0.0$ and $T = 0.8$. #N and #E indicate the number of nodes and edges.MAD quantifies the difference between the transition matrices at these two temperatures.

The aggregate data in Table 4 show that the previous behavior is not restricted to *Cities*, but systemic across all domains: overall, the transitions of Gemma-3-4B almost double with $T = 0.8$. In contrast, GPT-4o undergoes just a marginal increase. A breakdown per domain is in the Appendix (Table 5).

On the number of nodes, full structural coverage is reached at higher temperatures, with all models attaining the maximum graph size, indicating that higher temperature promotes using all speech acts across all poem positions. However, proprietary GPTs already achieve full coverage with $T = 0.0$, confirming that, unlike Gemma, their baseline structural diversity is independent of temperature.

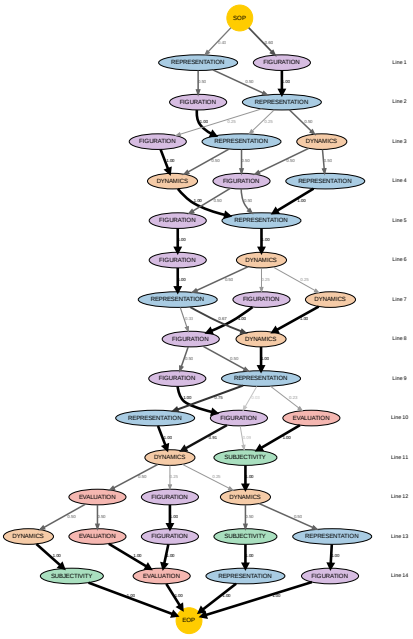
Metrics further highlight that Gemma-3-27B reaches a maximum *Stability* at $T = 0.0$, meaning that it always generates the exact same poem. Increasing to $T = 0.8$ successfully breaks this determinism, dropping internal stability to 0.82. This suggests the correlation between temperature and creativity, as it signals a move away from rigid repetition toward structural variety. In this case, the gain in diversity occurs without sacrificing thematic focus, as *Topic Similarity* remains stable across all models, regardless of temperature, ensuring that the poems explore new structural paths while remaining anchored to the target entity.

Finally, GPT LLMs are more stable across different temperatures, with more divergence (MAD) in the smaller model, but still lower than in the Gemma LLMs, where the smaller model is also the one where temperature has greater impact. This suggests that, regardless of the family, larger models are less affected by temperature.

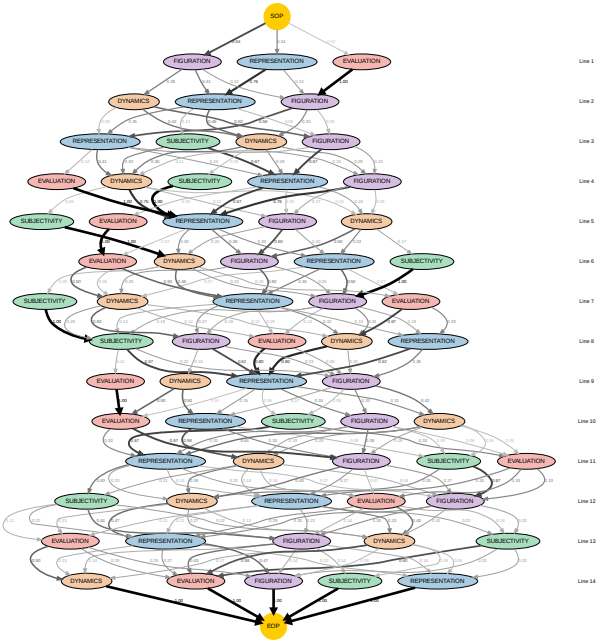
Overall, results suggest that smaller open quantized LLMs lack baseline stability and rely heavily on temperature to force structural diversity.

Conclusion

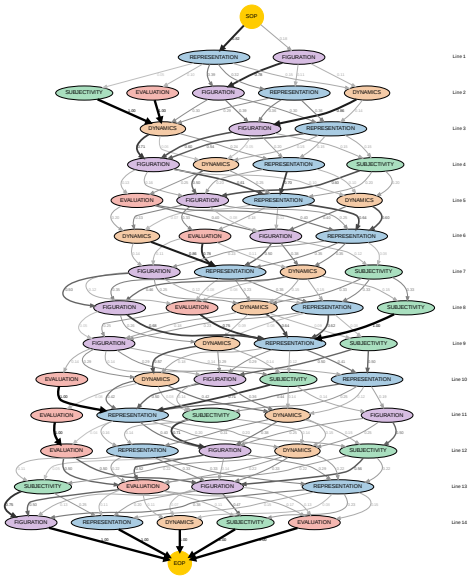
We showed how speech act flows can be useful for analyzing poetry and used them for studying the impact of model size and temperature in LLM-generated poems. These flows showed that higher temperatures successfully increase structural variety by reducing stability, without sacrificing thematic focus. Additionally, our study revealed clear differ-



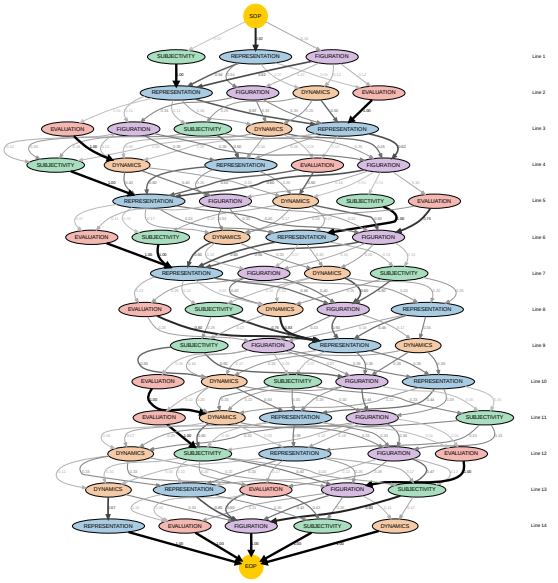
(a) Gemma-3-27B ($T = 0.0$)



(b) Gemma-3-27B ($T = 0.8$)



(c) GPT-4o ($T = 0.0$)



(d) GPT-4o ($T = 0.8$)

Figure 1: 14-level flows of speech acts for the Cities domain for Gemma-3-27b and GPT-4o at $T = 0.0$ and $T = 0.8$. Node colors represent the speech acts: Representation (blue), Figuration (purple), Evaluation (red), Dynamics (orange), and Subjectivity (green). Yellow nodes mark the start (SOP) and end (EOP) of the poem.

ences between model families: larger proprietary models produce more complex sequences even at zero temperature, whereas open quantized models are highly rigid at baseline and rely entirely on temperature to achieve diversity. By capturing these behaviors, our visually-aided methodology proved to be a powerful tool for better understanding how LLMs balance creativity and structure, and may complement future evaluations on this domain.

Once confirmed the utility of flows in this task, we plan to deepen the analysis of LLM-generated poetry and expand the study to other LLMs, temperatures, poetry styles, and languages, possibly revising the taxonomy of speech acts.

Acknowledgments

This work was financed by the Portuguese Recovery and Resilience Plan (PRR), through project C645008882-00000055– Center for Responsible AI. This work was also supported by FCT– Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra. Patrícia Ferreira was supported by FCT – Foundation for Science and Technology, I.P. through the PhD scholarship with reference 2024.01240.BD.

References

Agirrezabal, M., and Gonçalo Oliveira, H. 2024. Zero-shot Metrical Poetry Generation with Open Language Models: a Quantitative Analysis. In *Proceedings of 15th International Conference on Computational Creativity*, ICCC. Jönköping, Sweden: ACC.

Austin, J. L. 1975. *How to do things with words*. Harvard university press.

Chen, Y.; Gröner, H.; Zarrieß, S.; and Eger, S. 2024. Evaluating diversity in automatic poetry generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 19671–19692. ACL.

Ferreira, P.; Carvalho, I.; Alves, A.; Silva, C.; and Gonçalo Oliveira, H. 2024. Sentiment-aware dialogue flow discovery for interpreting communication trends. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 274–288. Kyoto, Japan: ACL.

Martinez, J. M. S., and Nugent, A. 2022. Inferring ranked dialog flows from human-to-human conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 312–324. ACL.

Matos, J.; Silva, C.; and Gonçalo Oliveira, H. 2025. Cognitive flow: An LLM-automated framework for quantifying reasoning distillation. In *Proceedings of the 18th International Natural Language Generation Conference*, 596–616. Hanoi, Vietnam: ACL.

Peeperkorn, M.; Kouwenhoven, T.; Brown, D.; and Jordanous, A. 2024. Is temperature the creativity parameter of Large Language Models? In *Proceedings of 15th International Conference on Computational Creativity*, ICCC. Jönköping, Sweden: ACC.

Qu, Z.; Yuan, S.; and Färber, M. 2026. Poetone: A framework for constrained generation of structured chinese songci with LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 32764–32772.

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; and Khatun, A. 2023. Bits of grass: Does GPT already know how to write like Whitman? In *Proceedings of the 14th International Conference for Computational Creativity (ICCC'23)*.

Walsh, M.; Preus, A.; and Gronski, E. 2024. Does ChatGPT have a poetic style? In *Proceedings of the Computational Humanities Research Conference 2023*, volume 3558. CEUR-WS.org.

Yu, C.; Zang, L.; Wang, J.; Zhuang, C.; and Gu, J. 2024. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 315–325.

Appendix

Prompts

This section details the prompts used throughout the study, following a structured pipeline from data generation to automated annotation. The process was divided into three main stages:

1. **Sonnet Generation:** Figure 2 shows the prompt used to generate the English sonnets across all models. It includes constraints to ensure the output consists only of the 14 poetic lines.
2. **Speech Acts Discovery:** Figure 3 presents the zero-shot prompt provided to GPT-4o to analyze a sample of the corpus and identify the most recurring functional categories of speech acts.
3. **Automated Classification:** Figure 4 details the final classification prompt. This instruction was used to label every line of the dataset into the five discovered categories.

> Write a sonnet about {`entity`}.
> Rules: Write strictly in English. Do not write introductions or conclusions. Write exactly and only the 14 lines of the sonnet.

Figure 2: Sonnet generation prompt, where {`entity`} represents the target entity.

> You are an expert in analyzing poetry. I am conducting a study on AI-generated Sonnets in English. I have a list of 200 verses below. I need you to categorize them based on what the poetic line is doing (its function). Read the verses and create a list of 5 or 6 simple categories that describe what each verse is doing. Here is the sample of verses: `[Sample list...]`

Figure 3: Prompt used for the inductive discovery of the speech act taxonomy.

Flows for all Models and Domains

This section presents the complete 14-level flows aggregated across all three evaluated domains (*Sports*, *Music*, and *Cities*). Figure 5 illustrates the overall structural behavior of the open-weight Gemma models, while Figure 6 details the proprietary GPT models. These aggregated graphs confirm that the structural patterns observed in specific domains remain consistent globally.

Metrics

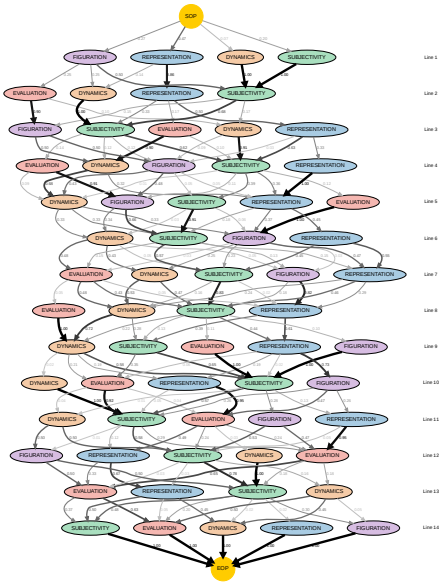
Table 5 provides a detailed breakdown of the structural and semantic metrics across the three evaluated domains. These values confirm that the previously observed trends are systemic behaviors rather than isolated artifacts. Across all domains, increasing the temperature expands the structural network (increasing both nodes and transitions) while simultaneously degrading internal stability and topic similarity.

> Classify the provided poetic line/verse strictly into one of the following 5 categories, based exclusively on the definitions below:
> 1. **Representation:** Concrete depiction of physical settings, landmarks, objects, or visible attributes with focus on sensory or spatial detail. It also includes situating a subject within a cultural, geographic, historical, or social framework.
> 2. **Subjectivity:** Expression or attribution of affective states (joy, sorrow, passion, hope, etc.) and the revelation of inner life (soul, spirit, memory, psychological depth).
> 3. **Evaluation:** Explicit positive evaluation of a person, place, or entity, emphasizing greatness, superiority, or admiration. It includes the construction of permanence, iconic status, or enduring impact.
> 4. **Figuration:** Figurative transformation where something is described as or compared to something else (simile, symbolic equivalence). It also includes explicit or implicit opposition, tension, qualification, or paradox.
> 5. **Dynamics:** Presentation of dynamic movement or event performed by an agent. It also includes the expression of intention, function, effect, or goal.
> Answer only with the category name. Do not write justifications, periods, or introductions.
> Verse to classify: {`verse`}

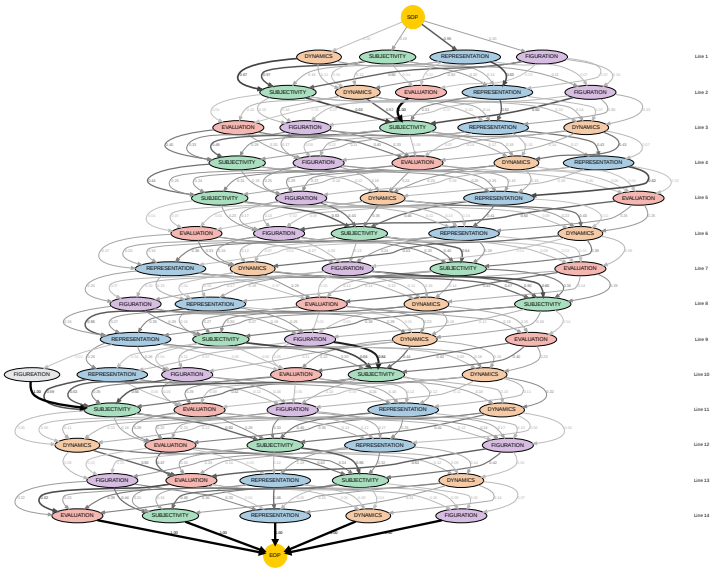
Figure 4: Poetic line classification prompt, where {`verse`} represents the input text.

Furthermore, the domain-specific breakdown reveals distinct semantic challenges. Notably, all models struggle significantly more to maintain topical focus in the *Music* domain compared to *Sports* or *Cities*. This is evidenced by the lower Topic Similarity scores across the board (e.g., GPT-4o drops to 0.37 ± 0.10 at $T = 0.8$), suggesting that generating poetry about musical entities induces a higher degree of abstract or metaphorical semantic drift.

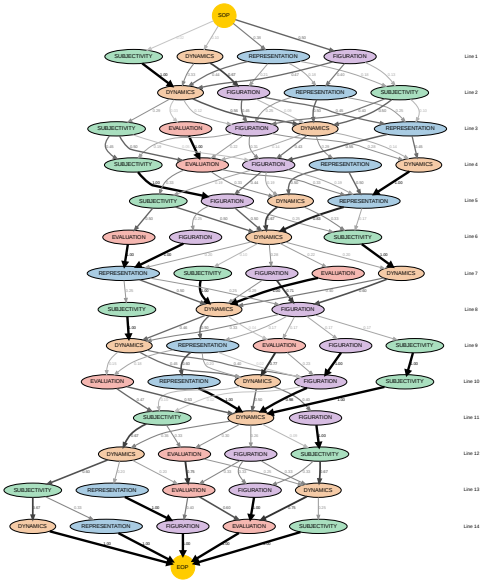
Finally, the distance metrics show how stable the proprietary models are. In the *Cities* domain, GPT-4o has very low distance scores. This suggests that the model has a very strong and clear internal knowledge of these famous locations. Consequently, even when we increase the temperature, the model manages to keep the poem's structure stable.



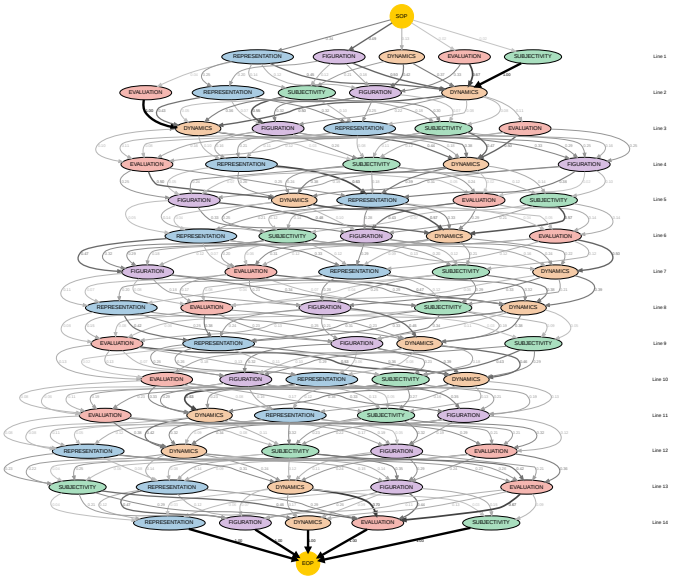
(a) Gemma-3-4B ($T = 0.0$)



(b) Gemma-3-4B ($T = 0.8$)

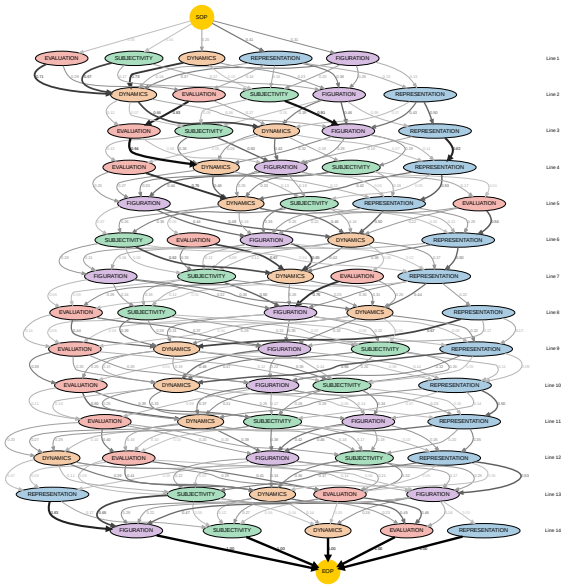


(c) Gemma-3-27B ($T = 0.0$)

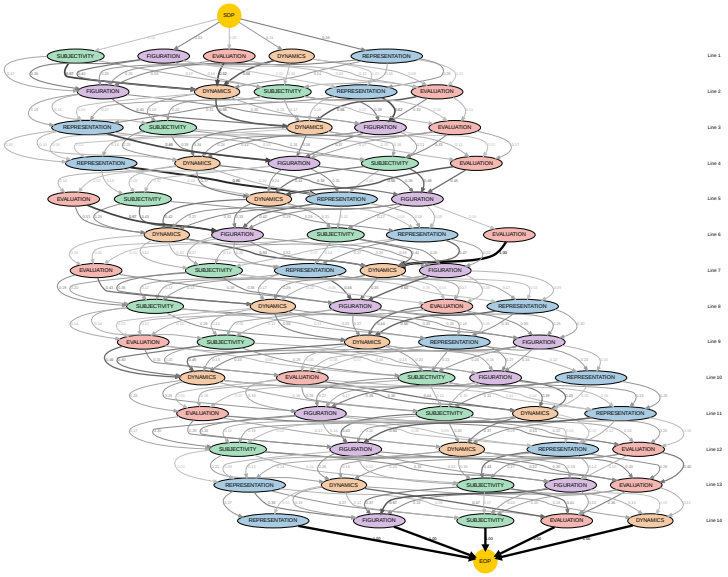


(d) Gemma-3-27B ($T = 0.8$)

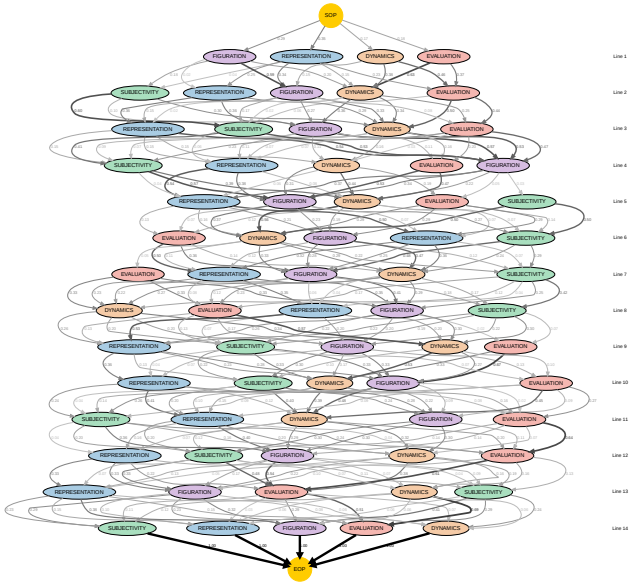
Figure 5: 14-level flows of speech acts across all domains for Gemma-3-4b and Gemma-3-27b at $T = 0.0$ and $T = 0.8$.



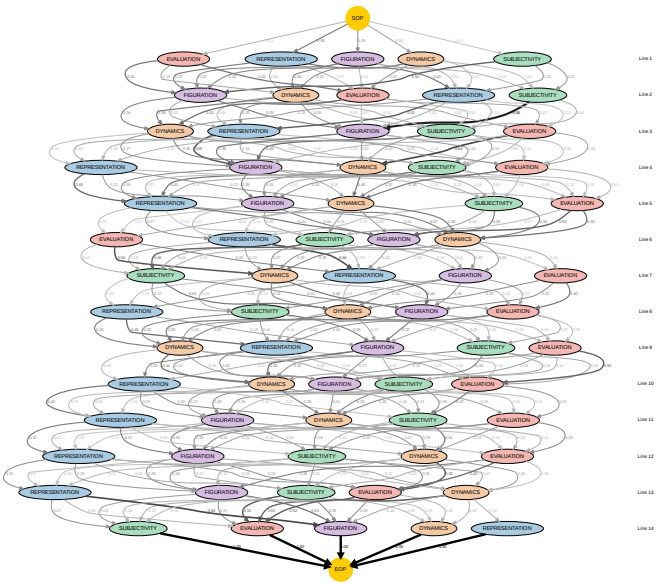
(a) GPT-4o-mini ($T = 0.0$)



(b) GPT-4o-mini ($T = 0.8$)



(c) GPT-4o ($T = 0.0$)



(d) GPT-4o ($T = 0.8$)

Figure 6: 14-level flows of speech acts across all domains for GPT-4o-mini and GPT-4o at $T = 0.0$ and $T = 0.8$.

		# Nodes		# Transitions		Topic Similarity		Stability		
Model	**Domain**	$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	$T = 0.0$	$T = 0.8$	**MAD**
Gemma-3-4B	Sports	49	68	87	204	0.35 ± 0.03	0.34 ± 0.11	0.90 ± 0.01	0.76 ± 0.09	5.77
	Music	46	66	73	170	0.24 ± 0.01	0.18 ± 0.05	0.89 ± 0.01	0.77 ± 0.09	5.72
	Cities	40	68	64	183	0.37 ± 0.03	0.38 ± 0.05	0.90 ± 0.02	0.82 ± 0.04	6.82
Gemma-3-27B	Sports	45	65	67	179	0.36 ± 0.00	0.37 ± 0.08	1.00 ± 0.00	0.81 ± 0.06	4.96
	Music	46	68	65	208	0.26 ± 0.00	0.24 ± 0.06	1.00 ± 0.00	0.79 ± 0.06	6.53
	Cities	40	66	57	195	0.31 ± 0.00	0.33 ± 0.05	1.00 ± 0.00	0.85 ± 0.05	6.06
GPT-4o-mini	Sports	57	65	133	180	0.49 ± 0.09	0.44 ± 0.11	0.84 ± 0.07	0.73 ± 0.08	4.14
	Music	57	59	165	180	0.34 ± 0.06	0.34 ± 0.10	0.88 ± 0.05	0.73 ± 0.07	4.45
	Cities	58	65	139	203	0.40 ± 0.03	0.37 ± 0.06	0.89 ± 0.05	0.77 ± 0.03	4.36
GPT-4o	Sports	57	64	146	191	0.49 ± 0.05	0.44 ± 0.10	0.87 ± 0.04	0.75 ± 0.04	5.22
	Music	60	67	171	191	0.45 ± 0.07	0.37 ± 0.10	0.86 ± 0.05	0.75 ± 0.05	2.85
	Cities	61	67	183	198	0.42 ± 0.03	0.41 ± 0.04	0.87 ± 0.04	0.78 ± 0.03	2.06

Table 5: Domain-specific structural and semantic metrics at $T = 0.0$ and $T = 0.8$. MAD quantifies the difference between the transition matrices generated at these two temperatures.