Semantic Stability

November 26, 2025

1 Introduction

This projects aims at testing and visualizing the responses, and more importantly the differences between responses to slightly different, but similar prompts.

2 Theory

A simple base prompt p, and a set of variant p^* prompts are chosen. Using the *all-MiniLM-L6-v2* model, their embedding vectors are created. Only those variant p^* prompts are selected, which surpass the **0.85** similarity (calculated using *cosine similarity*).

Base prompt. `Why do humans need sleep?`

Prompt variants.

- What makes sleep essential for humans?
- How does sleep benefit the human body and mind?
- What role does sleep play in human health and functioning?
- Why is it necessary for people to sleep?
- In what ways is sleep crucial to human well-being?
- What are the reasons humans can't function without sleep?
- Why is getting enough sleep important for humans?
- What happens to the human body and brain that makes sleep a necessity?

Filtering prompt variants. To control prompt diversity, we compute the semantic similarity between each variant and its base prompt. The procedure is as follows:

1. Encode the base prompt p and all variants p_i^* using a SentenceTransformer model.
2. Compute cosine similarities $s_i = \cos\_\mathrm{sim}(p, p_i^*)$ for each variant.
3. Construct a data frame containing each variant, its similarity score, and a Boolean flag indicating whether it meets a minimum threshold (e.g. $s_i \geq 0.85$).
4. Sort the data frame in descending order by similarity.

Table 1: Variant prompts with their similarity scores and if they remain in the experiment.

ID	Variant	Similarity	Keep
3	Why is it necessary for people to sleep?	0.918 821	True
6	Why is getting enough sleep important for humans?	0.882 275	True
0	What makes sleep essential for humans?	0.850 024	True
7	What happens to the human body and brain that makes sleep a necessity?	0.825 465	False
5	What are the reasons humans can't function without sleep?	0.817 788	False
4	In what ways is sleep crucial to human well-being?	0.753 535	False
1	How does sleep benefit the human body and mind?	0.747 996	False
2	What role does sleep play in human health and functioning?	0.693 460	False

To determine the base response, the base prompt is sent 10 times to our chosen LLM (*gpt-5-mini*) - from now on referred to as *model* -, and KMeans is applied to the responses' embeddings.

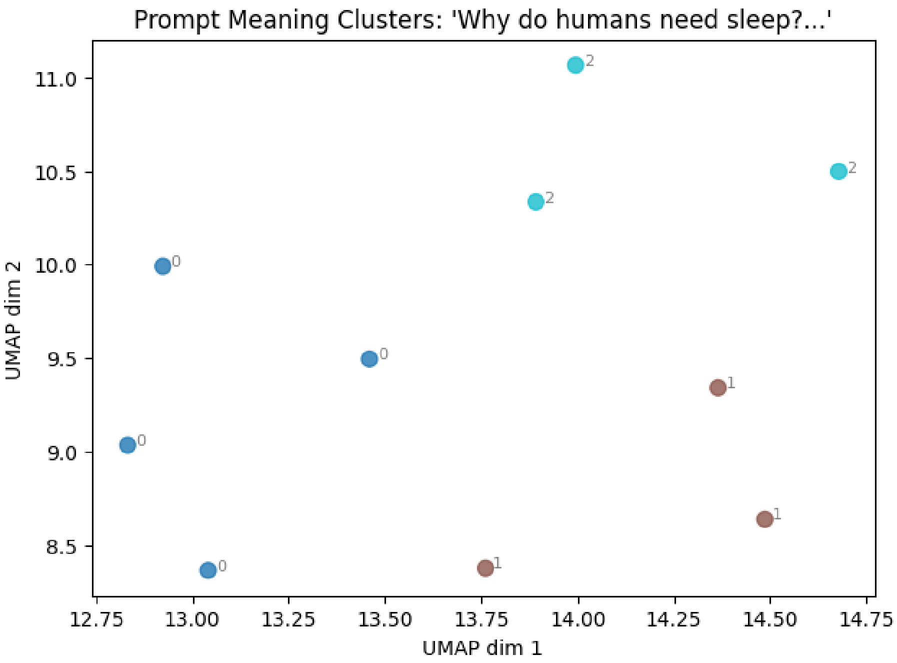

Figure 1: Base Responses' Embeddings Cluster

To choose a cluster, we calculate the cosine similarity between the cluster members and choose the cluster with the most internal similarity, which in this case is *Cluster 0* with a **0.936** similarity. The cluster's centroid is calculated, and from now on it acts as the *base response embedding.*

The following 3 variant prompts were kept and tried in the project.

Table 2: Selected variant prompts with their similarity scores.

ID	Variant	Similarity
3	Why is it necessary for people to sleep?	0.918 821
6	Why is getting enough sleep important for humans?	0.882 275
0	What makes sleep essential for humans?	0.850 024

Every variant prompt is sent 10 times to the *model* and the embedding vector, token number and similarity compared to the *base response embedding* is saved.

Table 3: Responses for *Why is it necessary for people to sleep?*.

Response Index	Text	Similarity	Token Count
1	Sleep is not optional for good health — it’s a...	0.917 627	427
2	Sleep is essential because it performs many cr...	0.946 412	550
3	Sleep is essential because it supports many cr...	0.894 599	331
4	Sleep is essential because it lets your body a...	0.945 880	590
5	Sleep is not optional for the brain and body—i...	0.923 881	597
6	Sleep is not optional for the body and brain —...	0.899 376	616
7	Sleep is essential because it supports many cr...	0.934 993	516
8	Sleep is essential because it supports many bi...	0.918 472	371
9	Short answer: sleep is essential because it le...	0.946 763	570
10	Sleep is essential because it lets your brain ...	0.925 013	397

Table 4: Responses for *Why is getting enough sleep important for humans?*.

Response Index	Text	Similarity	Token Count
1	Sleep is essential for nearly every aspect of ...	0.850 531	364
2	Sleep is essential because it supports nearly ...	0.831 646	494
3	Sleep isn’t just “rest”—it’s an active, essent...	0.798 350	413
4	Sleep is a fundamental biological need. Gettin...	0.815 536	504
5	Getting enough sleep is essential because it a...	0.793 406	536
6	Sleep is essential for almost every system in ...	0.791 501	403
7	Sleep is essential because it supports nearly ...	0.861 821	470
8	Sleep is essential for nearly every part of yo...	0.843 667	380
9	Getting enough sleep is essential because slee...	0.839 533	432
10	Sleep is essential because it supports nearly ...	0.838 996	397

Table 5: Responses for *What makes sleep essential for humans?*.

Response Index	Text	Similarity	Token Count
1	Sleep is essential because it’s when many crit...	0.898 335	598
2	Sleep is essential because it’s when the body ...	0.916 517	652
3	Sleep is essential because it is when the brai...	0.912 007	490
4	Sleep is essential because it performs multipl...	0.885 284	557
5	Sleep is essential because it supports multipl...	0.920 379	555
6	Sleep is not just passive rest — it is an acti...	0.920 757	526
7	Sleep is essential because it supports multipl...	0.913 000	566
8	Sleep is essential because it’s when many acti...	0.915 230	664
9	Short answer: Sleep is essential because it su...	0.945 184	611
10	Sleep is not just “time off” — it’s an active,...	0.911 126	566

Variant Prompt Response Map. To better understand how small prompt differences influence the semantic content, a two-dimensional visualization was created (Figure 2). Each point in the plot corresponds to a single response generated from one of the three selected variant prompts. The horizontal axis represents the token count of the response, while the vertical axis shows its cosine similarity to the base response centroid. A small amount of jitter is applied to the token count to avoid marker overlap.

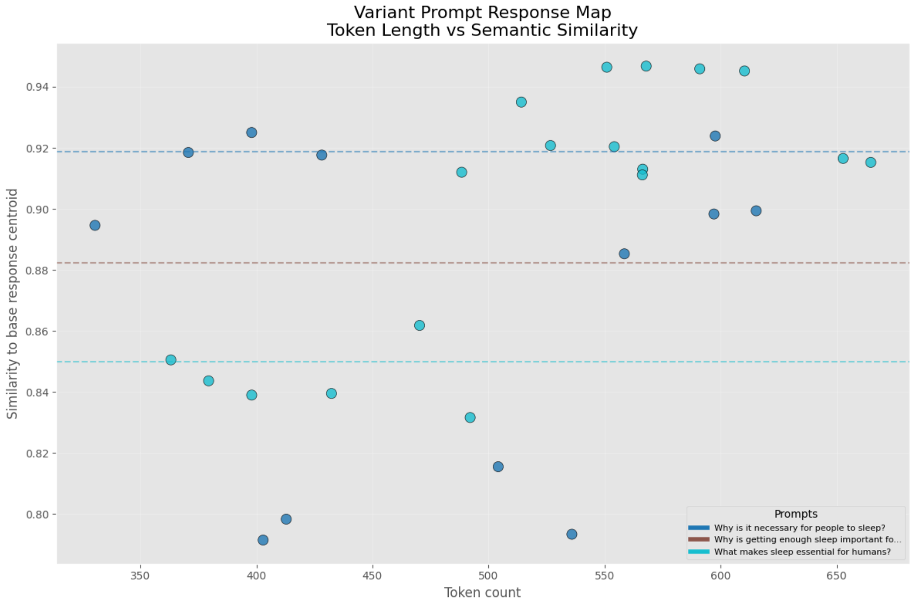

Figure 2: Variant Prompt Response Map showing token length vs. semantic similarity for three selected prompt variants.

The visualization incorporates two types of information: (1) **prompt-level similarity**, represented by dashed horizontal lines, and (2) **response-level behaviour**, represented by the scattered points. Each dashed line marks the semantic distance between a variant prompt and the original base prompt, making it possible to visually compare how far each prompt deviates in meaning before any responses are generated. Prompts with higher similarity scores (lines positioned near the top of the plot) are semantically closest to the base prompt.

For each prompt, the model was queried ten times, and the embedding vectors of these responses were compared to the base response centroid. This results in point clusters around the dashed lines, indicating how close the answers remain to the expected semantic region.

Two notable patterns emerge from the plot:

1. **Prompt similarity strongly predicts response similarity.** The prompt with the highest semantic similarity (*"Why is it necessary for people to*

sleep?”) produces responses tightly clustered near the upper region of the plot. In contrast, the prompt with the lowest similarity among the selected group (*“What makes sleep essential for humans?”*) shows responses shifted downward, indicating greater semantic drift. This confirms that even small changes in phrasing can systematically nudge the model toward different areas of its semantic space.

2. **Token count varies substantially but does not alone explain semantic drift.** Although the model produces responses with token counts ranging roughly from 330 to 650 tokens, the semantic similarity appears only weakly correlated with response length. Short and long responses both occur across the similarity range, implying that verbosity alone is not a strong predictor of semantic stability. However, for some prompts, a mild trend can be observed: longer responses may drift slightly further away from the centroid, possibly due to the model adding additional explanatory content.