Telephone Game with Multilingual LLMs

David R. Miller

{dave}@millermattson.com

September 6, 2025

Abstract

This is an experiment inspired by the children's game "telephone." In that game, a secret message is whispered to the first person, then whispered from person to person, and then the original and final messages are compared and everyone giggles. In this experiment, the whisperers are large language models (LLMs) and the message is translated to a different human language in each step, except for the final step in which the message is translated back to the starting language. We then measure the semantic drift using a sentence embedding model. We also explore the effects of different temperature and top_p settings. This game allows us to explore the accuracy and failure modes of LLM natural language translation.

1 Introduction

One use case for LLMs is for translating natural language text. This report is about an experiment to test the translation capabilities of several LLMs. The set of LLMs include a large frontier model, a small 8-billion parameter model fine-tuned for translations, and a few general-purpose open-weight models that run on consumer hardware.

Each step of our "telephone" game requires the LLM under test to translate from one language to another, finally ending at the original language where the accumulated semantic errors can be measured by cosine similarity of vector embeddings of the seed sentence and final sentence at the end of the round. We let each LLM play three rounds with each of seven different seed sentences and two different language sequences. Or dreadfully more formally:

Let $\mathcal{S} = \{S_0^{(1)}, S_0^{(2)}, \ldots, S_0^{(k)}\}$ be our set of k seed sentences, and let $\mathcal{L} = \{\mathbf{L}^{(1)}, \mathbf{L}^{(2)}, \ldots, \mathbf{L}^{(m)}\}$ be our collection of m language sequences, where each $\mathbf{L}^{(j)} = (L_0^{(j)}, L_1^{(j)}, \ldots, L_{n_j}^{(j)})$ represents a sequence of languages with $L_{n_j}^{(j)} = L_0^{(j)}$.

For each combination of seed sentence $S_0^{(i)} \in \mathcal{S}$ and language sequence $\mathbf{L}^{(j)} \in \mathcal{L}$, we conduct $R = 3$ independent rounds to account for stochastic variation in the LLM outputs. Each round r produces a final translated sentence $S_n^{(i,j,r)}$.

The overall semantic adherence score is computed as:

$$\text{Score} = \frac{1}{|\mathcal{S}| \cdot |\mathcal{L}| \cdot R} \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{L}|} \sum_{r=1}^{R} \text{sim}(\mathbf{e}(S_0^{(i)}), \mathbf{e}(S_n^{(i,j,r)}))$$

where $\text{sim}(\mathbf{e}, \mathbf{e})$ is cosine similarity.

1.1 Models

I tested the following models:

- Claude Sonnet 4 (frontier model)
- aya-expanse-8B Q8 (fine-tuned for translation)
- Mistral-7B-Instruct-v0.3 fp16
- gemma-3-27B-it-QAT Q4
- gpt-oss-20b MXFP4
- gemma-3-270m-it Q8 (tiny tiny tiny model)

Claude Sonnet 4 is the only commercial frontier model in the set and is known for its ability to translate human text between common languages. The Aya-Expanse model is a relatively small 8-billion parameter model that is specifically fine-tuned for translating text among common and uncommon languages. The Mistral 7B, Gemma3 27B, and gpt-oss 20B models are general-purpose, open-weight chat models that run on consumer hardware. The last model in the set, Gemma3 270M, is an extremely tiny model intended for embedding in consumer devices after fine-tuning for a specific purpose. Out of the box, it performed so poorly in these tests that its results are not included.

1.2 Prompting

The prompt was constant in all tests:

```
Role: You are a professional translator. Translate idiomatically,
preserve meaning and tone, and do not add commentary.
Task: Translate from {source_lang} to {target_lang}. Return JSON with
'translation' only. Text: {text}
```

1.3 LLM parameters

The LLM temperature was set to 0.2 and top_p to 0.88 in all tests. These values were determined empirically by human evaluation of the average collective performance of the LLMs. See Section 4 for more information of the effects of these two parameters.

1.4 Language sequences

The game was conducted in two sequences of languages—a short sequence of languages that most LLMs were trained on, and a longer sequence that includes languages less well represented in some LLM training data. The two sequences are shown in Table 1.

Table 1: Translation sequences

Short Trip

Step	Language
1	English
2	German
3	Japanese
4	French
5	Chinese
6	Spanish
7	English

World Tour

Step	Language
1	English
2	Polish
3	Hindi
4	German
5	Korean
6	Turkish
7	Japanese
8	Russian
9	French
10	Chinese
11	Spanish
12	English

1.5 Seed texts

The game was repeated for each of the seed texts shown in Table 2.

Each LLM got to play three rounds of each seed phrase through each language sequence, then all of its scores were averaged separately for the short language sequence and long sequence.

1.6 Measure of semantic errors

To measure the accumulated semantic drift through each language sequence round, we measure the similarity of the embedding vectors of the beginning and ending texts using the 3072-dimension OpenAI semantic embedding model text-embedding-3-large.

Table 2: Seed texts

1.	"That's how the cookie crumbles."
2.	"All our dreams can come true, if we have the courage to pursue them."
3.	"When it comes to AI, the pressure from the top isn't on caution, it's on speed."
4.	"The history of philosophy is primarily concerned with theories based on rational inquiry and argumentation; some historians understand it in a looser sense that includes myths, religious teachings, and proverbial lore."
5.	"In the hush of small moments, waltz with quiet beauty—feel it bloom beneath your touch. Herein lie resplendent jewels of existence, bravely outshining distant promises of tomorrows unknown."
6.	"They gossiped with the speed of a prairie fire in a high wind, consuming every fact and fiction in its path."
7.	"The meeting was a dumpster fire, but we'll play it by ear tomorrow."

2 Results

Figure 1 shows the results of all the games averaged.

3 Example

This is an example of a typical round through the long language sequence. This is the output of one round of Mistral 7B fp16:

1. **English seed:** "The meeting was a dumpster fire, but we'll play it by ear tomorrow."
2. **English to Polish**
   Spotkanie było totalną katastrofą, ale zobaczymy, jak potoczy się jutro.
   Back translation: The meeting was a total disaster, but we'll see how things play out tomorrow.
3. **Polish to Hindi**
   मीटिंग पूरी तरह से बर्बाद हो गई, लेकिन कल क्या होता है देखते हैं।
   Back translation: The meeting was a complete waste, but let's see what happens tomorrow.

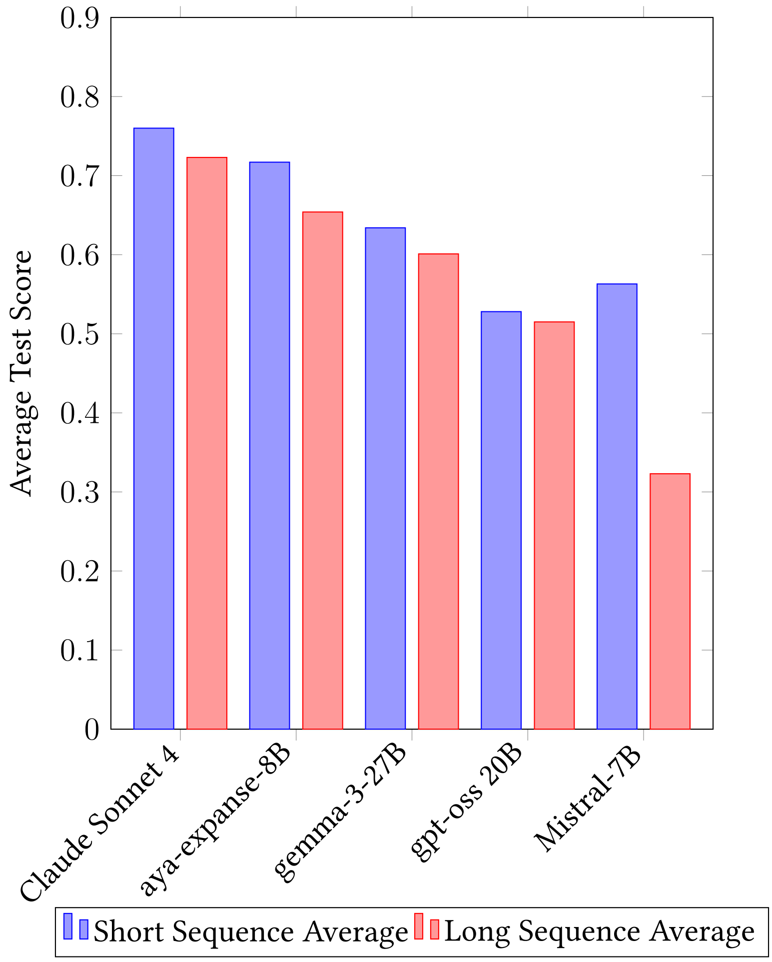

Figure 1: Overall LLM Performance

4. **Hindi to German**
   Das Meeting ist total im Eimer gewesen, aber wir werden sehen, was morgen passiert.
   Back translation: *The meeting was a complete mess, but we'll see what happens tomorrow.*

5. **German to Korean**
   오늘 회의는 완전히 망했지만, 내일 상황을 지켜보자.
   Back translation: *Today's meeting was a complete disaster, but let's wait and see how things look tomorrow.*

6. **Korean to Turkish**
   Bugünkü toplantı tam bir fiyasko oldu ama yarını bekleyelim, ne olur görelim.
   Back translation: *Today's meeting was a complete fiasco, but let's wait for tomorrow and see what happens.*

7. **Turkish to Japanese**
   今日の会議は完全な失敗だったけど、明日を待ちましょう、どうなるか見てみましょう。
   Back translation: *Today's meeting was a complete failure, but let's wait until tomorrow and see what happens.*
8. **Japanese to Russian**
   Сегодняшняя встреча сорвалась, но ничего, посмотрим, что будет дальше.
   Back translation: *Today's meeting fell through, but that's okay—let's see what happens next.*
9. **Russian to French**
   La réunion d'aujourd'hui a été annulée, mais ce n'est pas grave, on verra ce qui se passe ensuite.
   Back translation: *Today's meeting was canceled, but it's no big deal—we'll see what happens next.*
10. **French to Chinese**
    今天的会议取消了，不过没关系，我们走着瞧吧。
    Back translation: *Today's meeting was canceled, but no worries—let's wait and see.*
11. **Chinese to Spanish**
    La reunión de hoy se ha cancelado, pero no pasa nada, ya veremos qué pasa.
    Back translation: *Today's meeting is cancelled, but no worries, we'll see what happens.*
12. **Spanish to English**
    Today's meeting is cancelled, but no worries, we'll see what happens.

The similarity measure of this round is 0.591 as computed by the cosine similarity of the embedding vectors from the embedding model text-embedding-3-large, comparing the starting and ending texts, which are:

- "The meeting was a dumpster fire, but we'll play it by ear tomorrow."
- "Today's meeting is cancelled, but no worries, we'll see what happens."

In this round, the description of the meeting changed subtly from "disaster" to "complete failure" and then to "cancelled." The specificity of "tomorrow" morphed into a vague future.

4 Temperature and top_p tests

The effects of temperature and top_p are highly dependent on the LLM model and the characteristics of the text under translation. A temperature too low can result in translations that lack a certain dynamic quality. A temperature too high can distort semantic meaning which, in our telephone game, cannot be recovered once incurred.

Figure 2 shows the effect of temperature and top_p parameter values on the vector similarity scores for the model Aya-Expanse 8B Q8 using the seed phrase "In the hush of small moments, waltz with quiet beauty—feel it bloom beneath your touch. Herein lie resplendent jewels of existence, bravely outshining distant promises of tomorrows unknown." The values are the average of playing five rounds through the short language sequence for each combination of temperature and top_p values.

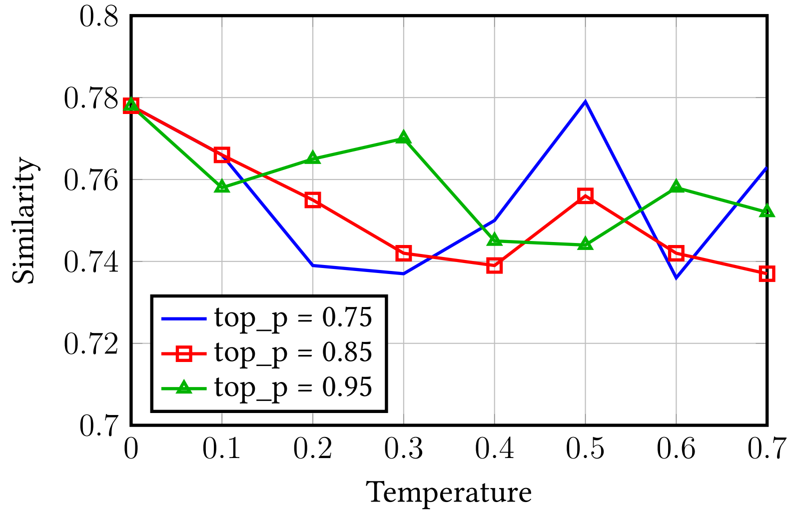

Figure 2: Effect of parameter values on similarity measure for Aya-Expanse 8B Q8

Figure 3 is a similar test using the model Gemma3 27B Q4 and the seed phrase "They gossiped with the speed of a prairie fire in a high wind, consuming every fact and fiction in its path."

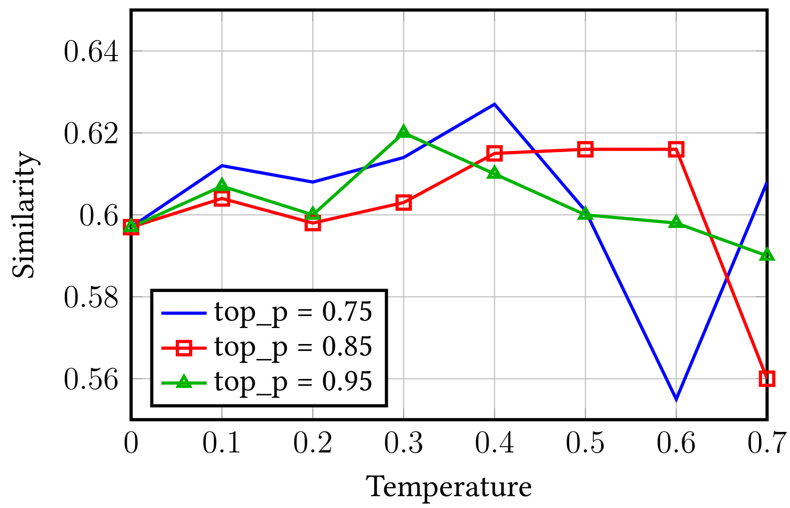

Figure 3: Effect of parameter values on similarity measure for Gemma3 27B Q4

5 Discussion

This experiment tested only short-form translation. Long-form translation will have different characteristics due to the different ways that LLMs are trained on long contexts.

It's not surprising that Claude Sonnet 4 performed the best in this experiment given its estimated hundreds of billions of parameters. It's pleasantly surprising to see Aya-Expanse, a model fine-tuned for language translation, perform nearly as well with only 8 billion parameters.

It is impressive how well these models are able to translate natural language in rough equivalent form. They do sometimes distort subtlety and nuance, but even human translators are unable to consistently maintain exact semantic equivalence.

There is commercial pressure to embed translation capabilities into edge and consumer technology, and that might work for certain use cases with the current crop of open-weight models, but not for sensitive applications that require consistent semantic fidelity. Businesses will try anyway. *Caveat emptor.*

Measuring semantic preservation with an embedding model is not an optimum measure of translation success. A panel of diverse human judges would be better.