Lost in Translation: Repurposing Semantic Similarity Benchmarks for Evaluating Lexical-Semantic Consistency in LLM-Based Machine Translation

Quin Ye[♦], **Jelke Bloem**[♣♠]

♦ Data Science (Information Studies), University of Amsterdam
♣ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
quin.ye@student.uva.nl, j.bloem@uva.nl

Abstract

We propose and demonstrate a repurposing of the lexical similarity benchmark Multi-SimLex and the SimLex-999 family of resources for assessing the cross-lingual lexical-semantic consistency of multilingual large language models. While originally gathered for evaluating word embedding models, the parallel nature of the word pairs enables their use in machine translation settings. Using a manually verified subset of 500 word pairs from the Multi-SimLex dataset, we evaluate models' ability to assess semantic similarity and perform translation between English and Mandarin through zero-shot prompting. We compare BLOOMZ and GPT-4's similarity ratings against human-annotated benchmarks and examine translation consistency using our and other metrics, with GPT-4 showing stronger human alignment. As SimLex-999 and Multi-SimLex together cover a range of at least 25 languages, this approach has the potential to be extended to many language pairs including ones that don't involve English, though it requires some manual checks.

Keywords: semantic similarity, semantic consistency, multilingual language models, cross-lingual evaluation, machine translation

1. Introduction

For the past 10 years, the field of NLP has put significant effort into collecting human similarity ratings of word pairs, as these proved to be effective for benchmarking continuous vector representations of words. While word association benchmarks such as WordSim353 existed previously (Finkelstein et al., 2001), the success of word embedding models such as Word2Vec (Mikolov et al., 2013) led to a focus on evaluating similarity as opposed to association and relatedness, first embodied by the annotation guidelines of SimLex-999 (Hill et al., 2015). Given human similarity ratings of a word pair and the cosine similarity of the embeddings of those words in the embedding space of a model, it was expected that better models would yield better correlations with the human ratings.

This form of benchmarking was widely adopted and it led to the creation of SimLex-999 resources for a range of languages, where the English word pairs would be translated and rated by native speakers of that language. Subsequently, Multi-SimLex (Vulić et al., 2021) was introduced to extend this effort across typologically distinct and under-resourced languages, formalizing the previously accidental alignment of word pairs as concept pairs.

While Multi-SimLex has been used to explore the embedding spaces of multilingual static and contextualized word embedding models (Vulić et al., 2021), an underexplored area is how multilingual generative decoder LLMs evaluate and preserve semantic relationships across languages. Recent studies have shown that semantic similarity ratings prompted from LLMs are highly correlated with human ratings (Trott, 2024; Snelder et al., in press; Brans and Bloem, in press). This raises the question of whether they are consistent in this across languages and whether testing this can yield insight into the lexical-semantic consistency of multilingual LLMs across languages.

Our analysis combines (i) agreement with human similarity scores, (ii) translation/back-translation consistency, and (iii) surface and semantic-level diagnostics, including Levenshtein distance (Yujian and Bo, 2007) as a surface-form stability proxy and embedding-based measures as a semantic drift proxy. Rather than proposing a machine translation metric, we treat translation as a controlled stress test for semantic stability: a way to reveal model biases (e.g., score compression) and failure modes in cross-lingual meaning preservation.

2. Related Work

As far as we were able to establish, SimLex-based benchmarks are currently available for 25 languages. Multi-SimLex (Vulić et al., 2021) contains 15 languages (originally 12) and Brans and Bloem (2024) list 13 languages for which SimLex-999 variants exist. Some languages have both, so without overlap, that's 24 languages, with a 25th (SimLex-

Proceedings of The Fourth Workshop on the Role of Resources in the Age of Large Language Models (RESOURCEFUL 2026), pages 1–12 May 11, 2026.

999 for Modern Greek, Mylonadis and Bloem, in press) having been developed since then. Various authors report high correlations between human ratings for one language and another (e.g. between 0.627 and 0.861 for the Multi-SimLex languages), indicating good cross-lingual alignment of the datasets per language. Beyond benchmarking models, these datasets have also been used to demonstrate how human ratings and model predictions differ (e.g. Snelder et al., in press for English and Mandarin), contributing to model development and tuning by identifying gaps between human and machine understanding across languages.

In the same vein, Trott (2024) and Snelder et al. (in press) have explored using instruction-tuned LLM to generate similarity scores directly through prompting several LLMs in English, Dutch and Mandarin to rate semantic similarity. These finding both point towards the challenges LLMs face in capturing semantic distinctions in non-English, lower-resource contexts.

There is a need for cross-lingual assessment of multi-lingual LLM's, how they represent and reflect each language differently, and whether these representations align with human judgments shaped by linguistic and cultural contexts. The present study extends this direction by analyzing whether BLOOMZ (Scao et al., 2023), an open instruction-tuned LLM with exceptionally broad language coverage, can rate similarity across different languages consistently and in alignment with human expectations. Multi-SimLex provides a valuable multilingual dataset with aligned similarity ratings across languages, allowing for comparative evaluation. This allows us to assess whether BLOOMZ perceives semantic shift across English and Mandarin in ways that align with humans and cultural context, to further examine the model's reasoning in cross-lingual tasks.

2.1. Machine Translation

While BLOOMZ is multilingual and instruction-tuned, it is not specifically fine-tuned for machine translation. There are more established models specifically designed for high quality translation, such as Google's Transformer-based Neural Machine Translation system (Wu et al., 2016), Facebook's M2M100 model (Fan et al., 2021), and Meta AI's NLLB-200 (No Language Left Behind, NLLB Team et al., 2022), which are explicitly trained on large-scale multilingual parallel corpora. These models are optimized for translation across many languages and often used in translation systems. In contrast, LLMs like BLOOMZ are trained with instruction on a broad range of tasks, and their translation ability is a byproduct of general language modeling rather than as a specifically designed functionality.

Based on these observations, we are expecting the translation quality of our study to vary depending on language pair, context and prompt formulation. Our study does not aim to benchmark BLOOMZ as a state-of-the-art translation system but instead focuses on how it reflects its representations of semantic relationships during translation.

2.2. Prompt Engineering

Prompt engineering has become a dominant method for aligning LLM behavior with human tasks. Instruction-tuned models like BLOOMZ (Scao et al., 2023) have been specifically trained to respond to a variety of natural language instructions, but their ability to handle abstract or scalar tasks like similarity rating remains limited. Prompt engineering directly influences the effectiveness and quality of the model's output (Marvin et al., 2024). Various studies have tested different prompting strategies, such as zero-shot, few-shots, specification of targeted languages, and prompting in different languages (Nair et al., 2024; Bawden and Yvon, 2023; Chen et al., 2023; Robinson et al., 2023).

These results collectively suggest that prompting outcomes diverge depending on the models, the task type, and even the linguistic characteristics of the input. There is no one size fits all strategy, and identifying optimal prompting techniques often requires task-specific trial and errors. In this study we will thereby take a more exploratory approach to prompt engineering, using the established strategies in related work of similarity rating task and BLOOMZ translations as our starting point. (Snelder et al., in press; Brans and Bloem, in press; Nair et al., 2024)

2.3. Translation Fidelity and Evaluation

Translation quality in LLMs is typically evaluated through BLEU scores, trained metrics such as COMET and BEER, BERTScore, human judgments, or back-translation consistency (Papineni et al., 2002; Stanojević and Sima'an, 2015; Edunov et al., 2018; Zhang et al., 2020; Rei et al., 2020) However, the dataset we use in this study is on the single-word level. BLEU requires long sequences while COMET is trained on longer sequences and less likely to work in a decontextualized setting.

In this study we focus on word-level translation fidelity using a Levenshtein distance metric, which quantifies surface-level character change (Yujian and Bo, 2007). While limited in capturing deep semantic drift, it provides a useful approximation of form preservation, particularly in short word-level translations. We also include a multilingual SentenceBERT-based metric as an alternative to BERTScore to capture semantic consistency while keeping the same model across languages.

A recent application of back-translation in the LLM era is demonstrated by Weigang and Brom (2025), who use it for cross-lingual semantic alignment of technical terms. Based on results with GPT-4, DeepSeek and Grok, they suggest that back-translation can not only be used to generate more data, but also to evaluate translation quality, By doing so, they implicitly assess cross-lingual semantic similarity (Weigang and Brom, 2025). This emphasis on semantic alignment is useful for measuring how well models preserve meaning across languages. Weigang and Brom (2025) note that intermediate languages are treated as semantic projection spaces, suggesting we can use back-translation paths as a diagnostic tool for multilingual embedding misalignment or ambiguity in the model.

3. Methodology

3.1. Data Selection

We use the English and Mandarin subsets of the Multi-SimLex dataset (Vulić et al., 2021), a large multilingual benchmark designed to evaluate lexical semantic similarity across languages. Each word pair in Multi-SimLex is annotated with a human-assigned similarity score on a 0–6 scale. The dataset contains 1,888 word pairs per language and enables aligned cross-lingual comparison.

While word pairs in Multi-SimLex are parallel between languages, it is important to note that the Mandarin pairs are not assumed to be direct translations of the English ones. For some pairs, cultural adaptation took place (e.g. Imperial metric units which are not used in Chinese), and sometimes words were substituted if direct translation would result in having duplicate word pairs among the 1,888 pairs. This generally also applies to the language-specific SimLex-999 variants, which were adapted based on the word pairs in the original SimLex-999.

To avoid inclusion of pairs that aren't direct translations of each other, we carried out manual verification to isolate word pairs suitable for evaluating translation fidelity and semantic similarity consistency.

3.1.1. Sampling and Manual Annotation

From the English–Mandarin subset, we randomly sampled 500 word pairs and manually annotated them based on the degree of translational equivalence. The annotation was performed by a coder proficient in both English and Mandarin. The annotation scheme included three categories:

Out of 500 sampled pairs, 423 were judged as accurate translations (Label 1), 53 were ambiguous or polysemous (Label 2), and 24 were incorrect (Label 0). Only the 423 pairs with direct and unambiguous translations were retained for the main analysis. These verified pairs serve as the core evaluation set for comparing LLM performance in both similarity rating and translation tasks.

Label	Description
1	Accurate translation – included in main analysis.
2	Valid but polysemous – may reflect divergent senses.
0	Incorrect translation – excluded from evaluation.

Table 1: Manual annotation scheme for English–Mandarin word pairs.

To complement the human ratings in Multi-SimLex, we later collect model-generated similarity scores from BLOOMZ and GPT-4 (via API), and examine their cross-lingual stability relative to human-labeled baselines.

3.2. Experiment Setup

We focus on BLOOMZ-7b1 (Scao et al., 2023), an open-source, multilingual, instruction-tuned LLM which is widely used in multilingual NLP research and resource creation, comparing it with GPT-4 (OpenAI, 2023), a widely used commercial LLM at the time of carrying out this experiment. Using the same prompt and translation cycle, BLOOMZ provides a baseline for open source multilingual LLMs, while GPT-4 provides a strong reference model as a commercial LLM. This pairing is intended as a contrastive case study between these two models.

The experiment is designed to evaluate how multilingual LLMs preserve and reflect lexical semantic similarity across languages through translation. As a reference point, we first compute human-rated semantic drift (Δ_{Human}) from the Multi-SimLex dataset by calculating the difference in similarity scores between English (`score_EN_human`) and Mandarin (`score__human`) of the same word pairs. This serves as a reference point for expected cross-lingual shifts in human perception.

To evaluate the models, we obtain self-reported similarity scores from BLOOMZ and GPT-4 for the same word pairs in both languages using structured prompts, which will be discussed in detail in the next subsection. These scores generated by LLMs are used to compute Δ_{LLM}, the model's perceived cross-lingual shifts across English and Mandarin. By comparing Δ_{LLM} with Δ_{Human}, we can assess whether the model captures the cross-lingual shifts in semantic similarity in a similar way as human raters.

In addition, we include a forward and backward

translation task to analyze semantic fidelity. Translation steps are evaluated both quantitatively (via match rates and Levenshtein distance) and semantically (via back-translated similarity scores).

The experimental workflow is structured as follows:

- **Step 1**: Prompt the model (BLOOMZ or GPT-4) to rate the similarity of English and Mandarin word pairs from Multi-SimLex (`ENG_1 + ENG_2` → `score_EN_llm`; `CMN_1 + CMN_2` → `score_CMN_llm`).
- **Step 2**: Translate each English word into Mandarin (`CMN_llm_1` and `CMN_llm_2`). Outputs are normalized to Simplified Chinese (e.g., punctuation, script).
- **Step 3**: Prompt the model to rate the similarity of its own translated Mandarin pair (`score_CMN_translation`) for internal consistency evaluation.
- **Step 4**: Perform back-translation of the Mandarin outputs into English (`EN_back_1`, `EN_back_2`).
- **Step 5**: Compare source and generated word pairs across both directions: (`ENG_1, ENG_2` vs. `EN_back_1, EN_back_2`) and (`CMN_1, CMN_2` vs. `CMN_llm_1, CMN_llm_2`).
- **Step 6**: Compute a similarity score for the back-translated English pair (`score_btEN_llm`) and evaluate its semantic drift:

  Δ_{tLLM} = `score_EN_llm−score_btEN_llm`

  A small Δ_{tLLM} suggests high semantic stability in translation loops. To supplement this, we calculate lexical match rates and Levenshtein distance between source and translated terms, providing both symbolic and semantic fidelity measures.
- **Step 7**: Compute cosine similarity between multilingual word embeddings of the original and translated/back-translated word pairs using Sentence-BERT (paraphrase-multilingual-MiniLM-L12-v2) (Reimers and Gurevych, 2019, 2020). These embeddings offer a semantic-level comparison independent of scalar ratings or lexical overlap.

3.2.1. Embedding-based semantic drift evaluation.

To further analyze lexical-semantic preservation, we apply an embedding method using the multilingual Sentence-BERT model `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2020). For each word pair, we construct concatenated phrases from both the original English terms and their corresponding Mandarin translation, both the gold translations from Multi-SimLex and the LLM-generated ones. This approach allows us to put the word pairs into a multilingual embedding space, enabling comparison between original English pairs and gold standard Mandarin pairs and LLM-translated Mandarin pairs in a shared embedding space.

This is an atypical use of contextual embedding models as words are embedded without any context or sentence structure, however, it is similar to how static embeddings are derived from contextual embeddings (Bommasani et al., 2020). As the SimLex-999 datasets consist of decontextualized words that were also rated by humans without context, models are inevitably also evaluated on it in this decontextualized manner, including contextual embedding models (Vulić et al., 2021). Our approach is comparable to Brans and Bloem's (in press) "joint" embedding condition for evaluating word pair similarity, though they embed one word with the other word as its context in a contextual embedding model rather than embedding them together in a SentenceBERT model.

We compute contextualized embeddings for:

- `ENG_combined` = `ENG_1` + `ENG_2`
- `CMN_combined` = `CMN_1` + `CMN_2` (gold standard)
- `CMN_llm_combined` = `CMN_LLM_1` + `CMN_LLM_2` (LLM-generated)

Cosine similarity is then used to compute:

- `embedding_similarity_eng_gold` between English and gold standard Mandarin;
- `embedding_similarity_eng_llm` between English and LLM translated Mandarin.

We also compute pairwise similarities within each language for the original English word pairs and their Mandarin counterparts:

- `sim_eng_pair`: similarity between `ENG_1` and `ENG_2`;
- `sim_cmn_gold_pair`: similarity between `CMN_1` and `CMN_2`;
- `sim_cmn_llm_pair`: similarity between `CMN_GPT_1` and `CMN_GPT_2`.

Finally, we define an embedding-based semantic drift score:

$$\Delta_{\text{Emb}} = \cos(\texttt{ENG\_combined}, \texttt{CMN\_llm\_combined}) - \cos(\texttt{ENG\_combined}, \texttt{CMN\_combined})$$

This measures how closely the LLM translated meaning aligns with human gold-standard translations in semantic space. These embedding scores offer a complementary diagnostic to similarity ratings by capturing semantic preservation from a continuous, contextualized perspective.

3.3. Prompt Design and Engineering

3.3.1. Similarity Rating Task

We experimented with several zero-shot prompt variants to elicit semantic similarity judgments, adapted from prior studies using instruction-tuned models (Snelder et al., in press). Surprisingly, prompts that performed well on GPT-4 (e.g., short format, decimal-only instructions) returned either empty responses or default scores (e.g., consistently "5") on BLOOMZ. This confirms how model-specific such prompts are and suggests a limitation in BLOOMZ's ability to follow formatting instructions for scalar rating tasks.

After testing multiple variants (see Appendix A for variations), we selected the following prompt (Prompt 2 from the appendix) for its relatively stable performance and broader score variance:

> Rate the similarity between the following two **{language}** words on a scale from 0 to 6:
>
> 0 = completely not similar 1 = barely similar 2–3 = weak similarity 4–5 = strong similarity 6 = nearly identical
>
> Words: "猫" and 狗" Answer with a single number only.

While this prompt still underrepresented extreme values (0 and 6), it outperformed others in producing variance aligned with human judgments.

As with any task, LLMs can be inconsistent in generating ratings, even when prompted multiple times with the same word pair. A more representative rating can be established by prompting for each pair multiple times and averaging the ratings, as was done by Snelder et al. (in press). We did not do this as it is computationally costly, and Snelder et al.'s (in press) results show that standard deviations of multiple ratings by the same model with the same prompt are lower than standard deviations of multiple humans rating.

3.3.2. Translation Task

For word translation, we used a straightforward instruction applicable to both forward (EN → ZH) and back (ZH → EN) translation:

> *Translate the following* **{source language}** *word into* **{target language}**: '**{word}**' *Translation:*

This prompt generally returned accurate translations. However, using "Mandarin" as the target language led to several undesirable outputs, such as:

- Responses in Traditional Chinese script,
- Regional or cultural vocabulary (e.g. Taiwanese/Cantonese),
- Bracketed or punctuated outputs (e.g. 「狗」 or [猫]).

To improve output quality and standardization, we replaced "Mandarin" with "Simplified Chinese" as the target language. This adjustment increased the likelihood of retrieving Mainland-standard Mandarin Chinese word forms. Inconsistent characters and formatting were then resolved through a final normalization step, converting all Chinese text to simplified characters and removing extraneous punctuation or brackets.

A similar normalization procedure was applied to back-translated English outputs to facilitate alignment and evaluation with original tokens.

3.4. Evaluation Metrics

To fully utilize the potential of SimLex-style datasets, we assess not only rating alignment with human judgments but also fidelity and semantic drift during translation.

3.4.1. Semantic Similarity Correlation

We first assess how well LLM-generated similarity scores align with human-labeled scores from Multi-SimLex. For both English and Mandarin, scores are normalized with z-scores. We compute Spearman's rank correlation between LLM and human scores to evaluate alignment in semantic judgment.

3.4.2. Cross-Lingual Rating Shift (Δ Analysis)

To evaluate the consistency of similarity perception across languages, we compute the cross-lingual difference in ratings for each word pair:

$$\Delta_{\text{Human}} = \texttt{score\_ZH\_human} - \texttt{score\_EN\_human},$$
$$\Delta_{\text{LLM}} = \texttt{score\_ZH\_llm} - \texttt{score\_EN\_llm} \quad (1)$$

We then compute the Spearman correlation between these two Δ distributions. This tests whether the LLM mimics the human-perceived semantic shift between English and Mandarin for the same pair of concepts.

3.4.3. Rating Difference Disagreement

We define a metric called *difference disagreement* to measure the divergence between human and model perceptions of cross-lingual change:

$$\texttt{difference\_disagreement} = \Delta_{\text{LLM}} - \Delta_{\text{Human}}$$

This allows us to detect specific word pairs where the model exhibits abnormal cross-lingual shifts and prioritize these for manual error analysis.

3.4.4. Translation Fidelity: Match Rate and Levenshtein Distance

To assess translation accuracy, we combine the surface metrics with word embedding. For the surface metrics, we compute:

- **Forward match rate:** Proportion of LLM-generated Mandarin terms that match gold-standard Mandarin from Multi-SimLex.
- **Backward match rate:** Proportion of back-translated English terms matching original English terms from Multi-SimLex.
- **Levenshtein distance:** String-edit distance between translated and gold-standard terms, capturing near-synonymy or morphological variations.

String-edit distances are not comparable across languages due to different spelling conventions and writing systems, especially in the case of English-Mandarin translation. However, our application of it only compares terms within a language (e.g. Mandarin translated to Mandarin gold). Naturally, edit distances for Mandarin will be lower due to the use of fewer characters per word, so differences in Levenshtein distance between English and Mandarin do not have a meaningful interpretation.

3.4.5. Deviation of Translation in Embedding Space

To complement the surface level evaluation of the translation quality, we use multilingual Sentence-BERT (Reimers and Gurevych, 2019) to analyze the drift in the translation at the embedding level. For each word pair, we compute cosine similarities between:

- The original English pair and its gold Mandarin equivalent
- The original English pair and the BLOOMZ-generated Mandarin translation
- The internal pairwise similarity within each language (e.g., `dog-cat` vs. 猫–狗)

Formally, let E = embedding($\texttt{ENG}_1$ + $\texttt{ENG}_2$) and C = embedding($\texttt{CMN}_1$ + $\texttt{CMN}_2$), then:

$$\text{sim}_{\text{gold}} = \cos(E, C), \quad \text{sim}_{\text{llm}} = \cos(E, \widehat{C})$$

We calculate the embedding based deviation in translation as:

$$\delta_{\text{embedding}} = \text{sim}_{\text{llm}} - \text{sim}_{\text{gold}}$$

This captures how far the BLOOMZ-generated translation deviates semantically from the human translation, using contextual embeddings instead of surface forms. We report average drift and distributional plots in the Results section.

4. Results

4.1. LLM Semantic Similarity Ratings: BLOOMZ vs GPT-4

We evaluated semantic similarity ratings using a manually verified subset of 423 English–Mandarin word pairs from Multi-SimLex. Both BLOOMZ-7b1 and GPT-4 were prompted to rate similarity on a 0–6 scale using standardized zero-shot instructions. Ratings were collected separately for English and Mandarin, and then z-score normalized for a better alignment and comparison with human annotations.

4.1.1. Rating Distributions

BLOOMZ ratings clustered around 4–5, especially in English, with much absence of extreme values (0, 1, 6), suggesting that the model could have a mid-range bias. Mandarin ratings had slightly more variance than English. In contrast, GPT-4 showed a more evenly distributed score usage, but the distribution is still distinct from the distribution of human ratings. See Appendix B for an overview of distributions.

4.1.2. Raw Similarity Correlation

BLOOMZ ratings weakly correlated with human scores (Spearman $\rho = 0.216$, $p = 0.0013$). English scores showed no significant correlation ($\rho = 0.054$). In contrast, GPT-4 achieved strong correlation with human scores in both Mandarin (Spearman $\rho = 0.750$, $p < 10^{-77}$) and English. ($\rho = 0.768$, $p < 10^{-83}$), suggesting it better captured gradations of similarity.

4.1.3. Cross-lingual Difference Correlation.

To assess whether LLMs capture semantic shift across languages, we computed the difference between English and Mandarin similarity scores per

word pair, and correlated these with human Δ values. BLOOMZ failed to capture cross-lingual semantic shifts (Spearman $\rho = -0.05$, $p = 0.447$). GPT-4 showed moderate alignment with human perceived semantc shifts (Pearson $r = 0.425$, Kendall's $\tau = 0.306$, both $p < 0.001$), indicating some similarity to human perception in tracking cross-lingual meaning shifts.

4.1.4. Difference Disagreement Analysis.

We further examined the difference disagreement. This metric reveals how closely each model matches the human-perceived magnitude of semantic shift. BLOOMZ exhibited large disagreements on certain pairs, particularly when the Mandarin rating distribution was flat. GPT-4 showed fewer extreme disagreements and better alignment with human semantic transitions. See Appendix C for top-10 divergence cases.

4.2. Lexical and Semantic Preservation in Translation

To evaluate translation fidelity, we examined both lexical preservation and semantic stability using multiple metrics, including exact match rates, Levenshtein distance, and cosine similarity from multilingual embeddings.

4.2.1. Exact Match Rates.

BLOOMZ preserved 146 and 138 tokens, respectively, during back-translation (ZH → EN), while forward translations (EN → ZH) preserved 224 and 184 tokens. This suggests BLOOMZ performed better on forward translation, possibly due to overliteral rendering of English terms into Mandarin. GPT-4 achieved higher preservation in both directions, with 199 and 174 tokens preserved after back-translation and 267 and 218 after forward translation.

4.2.2. Levenshtein Edit Distance

Edit distance between source and translated terms further reveals translation stability. BLOOMZ's back-translated English terms had an average Levenshtein distance of 3.7 to 3.8 characters, while its forward-translated Mandarin terms differed by less than 1.2 characters on average. GPT-4 showed significantly lower edit distances across both directions, with means of 1.4 (EN) and 0.6–0.7 (ZH), showing better surface-level preservation with GPT-4.[1]

[1] As noted in the methodology section, the differences between the languages are due to the different writing systems, so we can only compare between models here.

4.2.3. Semantic Similarity Scores

We compared LLM-generated similarity ratings before and after translation to measure semantic drift. For BLOOMZ, Pearson correlations between original and back-translated similarity scores were $r = 0.21$ (btEN) and $r = 0.33$ (ftZH), while GPT-4 improved to $r \approx 0.80$ (ftZH) and $r \approx 0.62$ (btEN). These results highlight that GPT-4 has better stability in translating word pairs without context across languages compared to BLOOMZ.

4.2.4. Embedding-Based Semantic Consistency

To assess whether translation preserves relative lexical-semantic similarity between word pairs, we compute cosine similarity using multilingual Sentence-BERT (`paraphrase-multilingual-MiniLM-L12-v2`). We then correlate the English pairwise similarities with their translated Mandarin pairs. GPT-4 shows strong preservation of similarity structure across translation, with Pearson $r = 0.782$ between English and Mandarin embeddings. In contrast, BLOOMZ yields a slightly lower correlation ($r = 0.711$), indicating a weaker consistency in the semantic space during translation.

Across lexical and semantic metrics, GPT-4 consistently outperforms BLOOMZ. It is important to note that both models perform better in forward translation (EN → ZH) than in back-translation, indicating that there might be translation asymmetry and instability in the reverse translation. This suggests that meaning can degrade progressively across multiple translation steps in both models.

5. Discussion

5.1. Overview of Findings

We introduced and demonstrated a reproducible framework for probing semantic consistencies in multilingual LLMs by re-purposing existing (Multi)SimLex lexical semantic resources. Due to the focus on lexical semantics of decontextualized words, this is far from a comprehensive approach to MT benchmarking. However, it does provide more detailed insight into lexical-semantic translation than text-based metrics such as BERTScore or COMET. Embedding models such as BERT are often themselves benchmarked using SimLex-style resources, so our approach cuts out the middleman and goes directly to a gold standard of those semantic models, at the cost of lexical coverage.

We have built the following elements on top of the human semantic similarity judgements from the original datasets:

- Cross-lingual similarity alignment metrics,

- Difference disagreement analysis for semantic shift detection,

- Surface-form stability signals (exact match; Levenshtein as an orthographic proxy) alongside embedding-based drift indicators.

Overall, our framework provides an interpretable diagnostic of how lexical semantic relations behave across languages and prompting conditions, and can be applied to any pair of languages out of the 25 languages for which (Multi)SimLex resources exist, to evaluate a broader range of different multilingual LLMs.

This study examined BLOOMZ's capacity to preserve semantic similarity and lexical fidelity in English–Mandarin translation, using a manually verified subset of the Multi-SimLex dataset. Results show that while BLOOMZ produces consistent responses under structured prompts, it lacks fine-grained sensitivity in similarity judgments and semantic consistency during translation. Despite BLOOMZ's state-of-the-art performance in translating under-resourced languages compared to GPT-3.5 (Nair et al., 2024), we observe a significant performance gap to GPT-4.

5.2. Semantic Similarity Ratings

BLOOMZ's similarity ratings weakly correlate with human annotations in Mandarin ($\rho = 0.20$) and English ($\rho = 0.12$), far below reported benchmarks for GPT-4o ($r = 0.86$, Snelder et al., in press) and BERT ($r = 0.476$, Ehrmanntraut et al., 2021) on SimLex-999. The cross-lingual difference correlation ($\rho = -0.05$) suggests that BLOOMZ does not reflect human-like perception of meaning shifts across languages. In contrast, GPT-4 achieved higher alignment with human ratings ($\rho = 0.768$) for English and ($\rho = 0.750$) for Mandarin, demonstrating better interpretability in cross-lingual semantic evaluation.

5.3. Lexical Preservation

Forward translation preserved approximately 50% of tokens, while back-translation accuracy dropped to roughly 30%. Levenshtein distance analysis indicated that forward translations differed by only 1–2 characters on average, but variance increased sharply during back-translation (std. ≈ 3.0–3.8), revealing instability. Rating correlations between original and translated word pairs were weak (Mandarin $r = 0.33$, English $r = 0.21$), confirming that BLOOMZ often fails to maintain semantic coherence across translation steps. GPT-4 showed moderately stronger correlations and more consistent lexical preservation.

5.4. Prompt Responses

Despite being instruction-tuned, BLOOMZ defaulted to narrow integer ranges (2–5) and avoided extreme or fractional scores. Exploratory prompt adjustments improved reliability but exposed sensitivity to task phrasing. While using the same prompt as BLOOMZ, GPT-4 produced finer-grained numeric responses, and a high correlation with human ratings. These findings reinforce the importance of instruction following and scale calibration for multilingual LLMs when performing cross-lingual tasks.

6. Conclusion

This study proposed a framework to evaluate cross-lingual semantic consistency in multilingual LLMs by repurposing existing semantic similarity benchmarks that exist for 25 languages. Our approach combines similarity correlation, translation fidelity, and disagreement metrics. The framework has the potential to be extended to other languages covered by SimLex resources. Of particular interest would be the investigation of under-resourced languages such as Yue Chinese and Welsh, as well as pairs of typologically distinct languages that don't involve English.

Our case study focused on BLOOMZ, which had shown promising performance in translation in a highly multilingual setting in previous work, translating under-resourced languages with good performance. Results indicate that while BLOOMZ demonstrates basic multilingual competence, especially in translation tasks, it lacks the ability to produce human-like semantic similarity ratings, and performs inconsistently especially in back-translation tasks. GPT-4 shows stronger abilities and stability in cross-lingual tasks, and it aligns closely with human perception in a dynamic cross-lingual semantic space. Our findings highlight systematic differences between BLOOMZ and GPT-4 in aligning with human similarity judgments and in maintaining lexical semantic relations under translation cycles, offering practical insight for interpreting and improving multilingual model behavior.

6.1. Ethical considerations and limitations

Using a dataset for a purpose different than its intended one can raise ethical concerns. However, we remain in the domain of similarity benchmarking, but in a cross-lingual setting and in combination with a translation task. We do not foresee any additional potential harms stemming from this use, especially as the original dataset is aggregated across participants.

A limitation of using SimLex-999 data is that these datasets have been around for a while and

might be present in LLM pretraining data. LLMs may therefore be better at rating these particular word pairs compared to words different to those in the SimLex-999 benchmark, which would overestimate the similarity of translations that are very similar to the benchmark. In theory, this could also contaminate the translation task itself. However, it is less likely that the translations were memorized as the pairs of the different-language SimLex-999 versions are rarely or never presented in a single document, as far as we are aware.

Manual annotation was performed by a single coder. We acknowledge this as a limitation; future work should involve multiple annotators to assess annotation consistency and reduce potential subjective bias.

One limitation of this study lies in BLOOMZ's restricted ability to process complex prompt structures and produce similarity ratings with meaningful granularity. Moreover, its translation of individual word pairs frequently altered or flattened the semantic relationship between terms, reducing the reliability of cross-lingual comparison. To address these limitations, the study explored prompt engineering, error analysis, and thorough evaluation metrics.

BLOOMZ's tendency to produce only mid range integer scores (2-5) makes it difficult to determine whether the poor correlation with human rating stems from a lack of semantic understanding, a flawed scoring scale and tuning, or whether the output genuinely reflects how the model internally represents semantic relationships. Future research should address this ambiguity by improving instruction tuning and exploring embedding extraction from the model to more accurately assess and refine LLM interpretability in the multilingual spaces of decoder LLMs.

7. References

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: The case of BLOOM.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Lizzy Brans and Jelke Bloem. 2024. SimLex-999 for Dutch. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 14832–14845, Torino, Italia. ELRA; ICCL.

Lizzy Brans and Jelke Bloem. in press. Multi-SimLex for Dutch: Benchmarking embedding- and prompt-based model performance on semantic similarity. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026))*. European Language Resources Association.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type-and token-based word embeddings in the digital humanities. In *Computational Humanities Research Conference*, pages 16–38.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. Prompt engineering in large language models. In *Data intelligence and cognitive informatics*, pages 387–402, Singapore. Springer Nature.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Leonidas Mylonadis and Jelke Bloem. in press. SimLex-999 for Modern Greek. In *Proceedings of the 4th Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL 2026) @ LREC 2026*.

Aarathi Rajagopalan Nair, Deepa Gupta, and B. Premjith. 2024. Investigating translation for Indic languages with BLOOMZ-3b through prompting and LoRA fine-tuning. *Scientific Reports*, 14(1):24202.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. GPT-4 technical report. Accessed October 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Teven Le Scao et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model.

Xander Snelder, Yunchong Huang, and Jelke Bloem. in press. Prompting instruction-tuned LLMs for semantic similarity values. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*. European Language Resources Association.

Miloš Stanojević and Khalil Sima'an. 2015. Evaluating MT systems with BEER. *The Prague Bulletin of Mathematical Linguistics*, 104(1):17–26.

Sean Trott. 2024. Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, pages 1–19.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2021. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Li Weigang and Pedro Carvalho Brom. 2025. LLM-BT-terms: Back-translation as a framework for terminology standardization and dynamic semantic embedding.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 29(6):1091–1095.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A. Alternative Prompts

A.1. Alternative Prompt 1: Role Setting

You are a bilingual speaker. Rate how similar the meanings of the following two `{language}` words are on a scale from 0 to 6:

0 = completely unrelated
1 = barely related
2 = weak similarity
3 = moderate similarity
4 = fairly similar
5 = very similar
6 = synonyms

Words: "`{word1}`" and "`{word2}`"

Choose a single number that best reflects the similarity in meaning. Avoid defaulting to the middle unless it clearly fits. Answer with a number only.

A.2. Alternative Prompt 2: Different Wording

Rate the semantic similarity between the following two `{language}` words on a scale from 0 to 6.

0 = completely unrelated
1 = weak relation
2–3 = loosely related
4–5 = moderately related
6 = near synonyms

Words: "`{word1}`" and "`{word2}`"

Answer with a single number only.

A.3. Alternative Prompt 3: Different Scale

Rate the semantic similarity between the following two English words on a scale from 0.0 to 20.0.

0 represents no semantic similarity and 20 represents perfect semantic similarity. Do not write anything else.

Words: "`{word1}`" and "`{word2}`"

Answer:

B. Similarity Rating Distribution

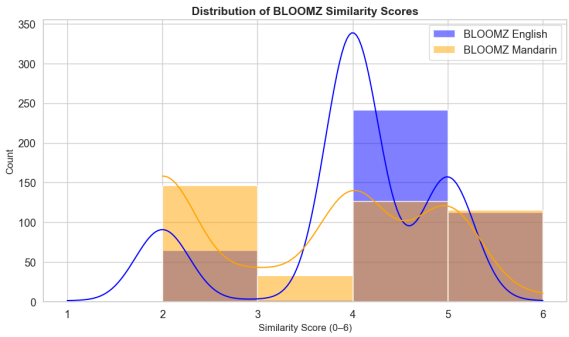

Figure 1: BLOOMZ Similarity Rating Distribution.

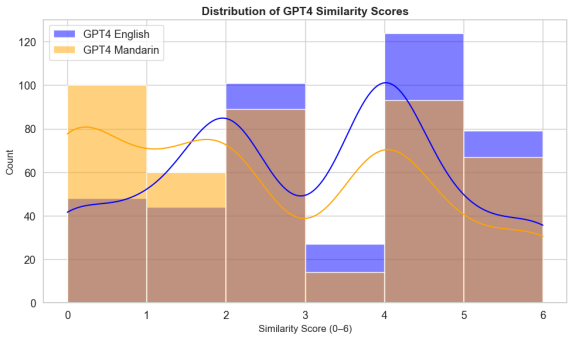

Figure 2: GPT4 Similarity Rating Distribution.

C. Difference Disagreement

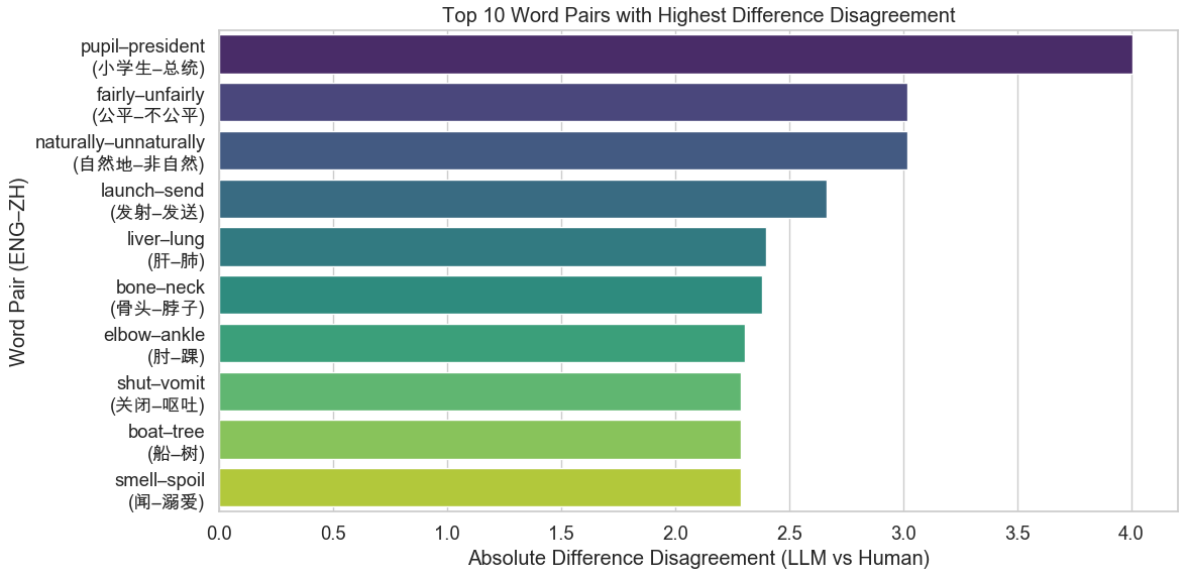

Figure 3: BLOOMZ Top 10 divergence

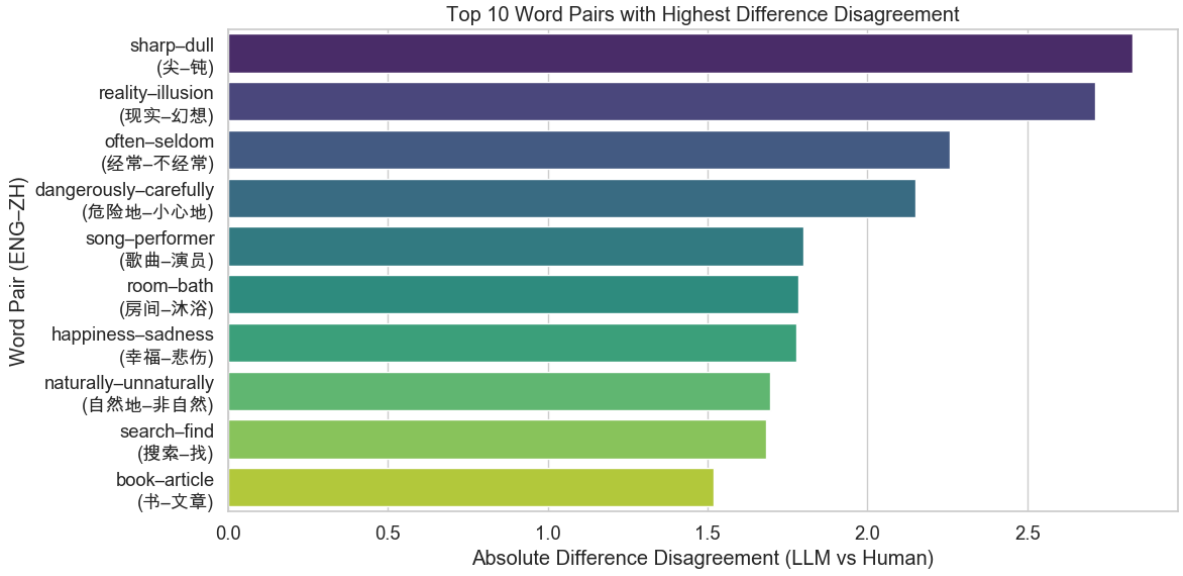

Figure 4: GPT4t Top 10 divergence