UNIVERSITEIT VAN AMSTERDAM

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

Quin Ye (quin.ye@student.uva.nl)

Jelke Bloem(J.bloem@uva.nl)

University of Amsterdam

UNIVERSITEIT VAN AMSTERDAM

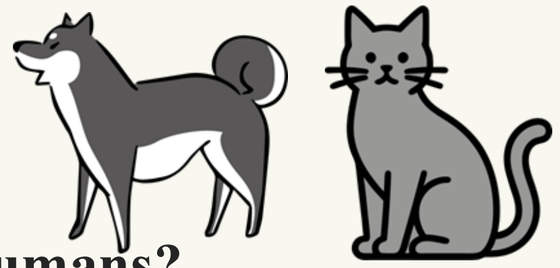
INTRODUCTION & MOTIVATION

- SimLex-999 and Multi-SimLex were originally built for evaluating semantic representations
- Their parallel concept structure makes them reusable for cross-lingual LLM evaluation.
- This lets us test whether models preserve lexical-semantic relations through translation
- The approach can potentially extend to 25 languages covered by SimLex-style resources and be applied to different Multilingual LLMs.

UNIVERSITEIT VAN AMSTERDAM

INTRODUCTION & MOTIVATION

Human: “dog” and “cat”=2 , “猫” and “狗” = 3



Δhuman=1

Do multilingual LLMs judge them the same as humans?

After LLM translation, dog -> hound, cat -> kitten,

Is the semantic relationship preserved, or does it drift in semantic space?

UNIVERSITEIT VAN AMSTERDAM

CONTRIBUTION

- A way to repurpose Multi-SimLex / SimLex-style resources for evaluating cross-lingual lexical-semantic consistency in LLMs
- A manually verified English–Mandarin subset for translation-sensitive analysis
- A diagnostic framework combining: human alignment, cross-lingual shift analysis translation/back-translation fidelity, surface and embedding-based metrics
- A comparative case study: BLOOMZ vs GPT-4

UNIVERSITEIT VAN AMSTERDAM

INTRODUCTION & MOTIVATION

Why BLOOMZ?

Open source

Multilingual

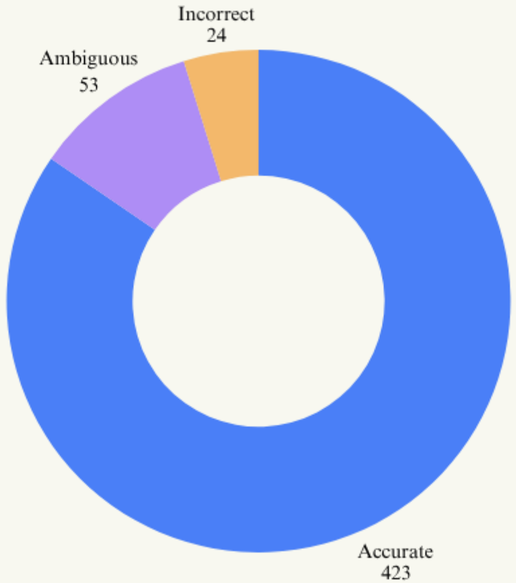
Instruction Tuned

Broad language coverage

UNIVERSITEIT VAN AMSTERDAM

DATA & MANUAL VERIFICATION

- **Start with English and Mandarin subsets of Multi-SimLex**
- **Randomly sample 500 word pairs**
- **Manually annotate translational equivalence**
- **Retain only the 423 verified pairs for the main analysis**



UNIVERSITEIT VAN AMSTERDAM

METHOD PIPELINE

INPUT

Verified Multi-SimLex EN–CMN pairs
human scores + manual verification

LLM PIPELINE

Rate original pairs similarity
LLM rates EN pair and CMN pair

LLM Translate EN → CMN
normalize to Simplified Chinese

Rate translated CMN pair
internal consistency check

Back-translate CMN → EN

Evaluation

Compare source vs generated pairs
EN source vs EN back-translation
gold CMN vs LLM CMN

Evaluate
ΔHuman vs ΔLLM
human alignment
match rate + Levenshtein
Sentence-BERT embedding

UNIVERSITEIT VAN AMSTERDAM

METRIC CHOICE

Why not rely only on BLEU / COMET / BERTScore?

- **the data are single-word, decontextualized pairs**
- **standard MT metrics are mainly designed for longer sequences**

We need diagnostics for:

- **semantic similarity alignment**
- **cross-lingual rating shift**
- **translation/back-translation stability**
- **lexical and embedding-based drift.**

UNIVERSITEIT VAN AMSTERDAM

PROMPT ENGINEERING

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: "**{word1}**" and "**{word2}**"
Answer with a single number only.

BLOOMZ struggled with fine-grained scalar output

prompts that worked for GPT-4 did not necessarily work for BLOOMZ.

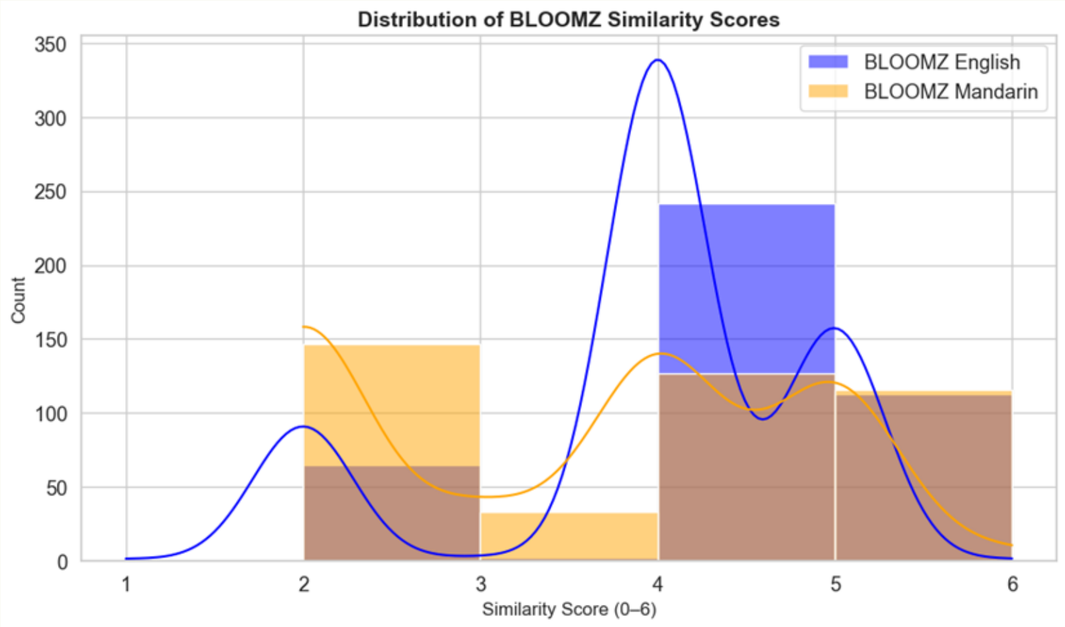
Thus, Specification of the scales needed.

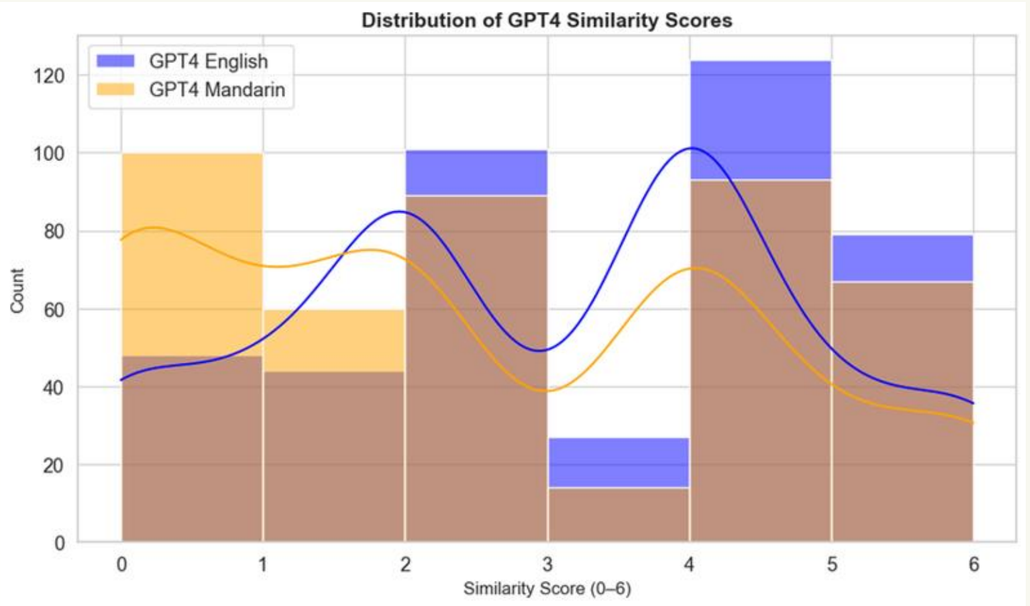
For translation, specifying Simplified Chinese helped standardize output.

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

UNIVERSITEIT VAN AMSTERDAM

RESULTS: SIMILARITY RATING





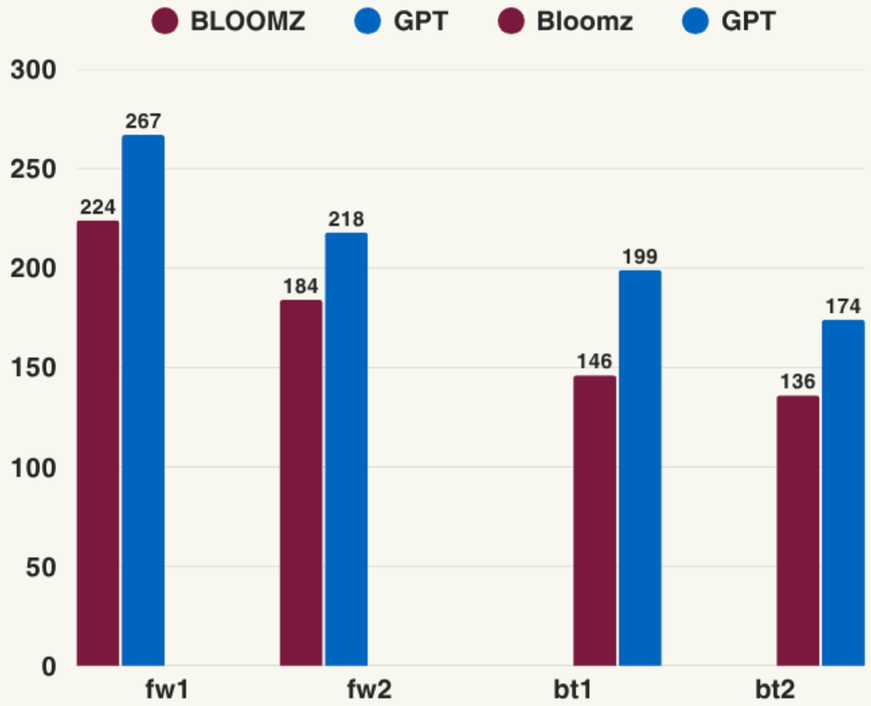
- **BLOOMZ ratings cluster in the mid-range and avoid extremes, GPT-4 uses the scale more meaningfully**
- **BLOOMZ shows weak correlation with human ratings AND no meaningful cross-lingual shift tracking (Δ)**
- **GPT-4 aligns much better with human judgments in both rating and behavior moderately similar to human in cross-lingual shift**

UNIVERSITEIT VAN AMSTERDAM

RESULTS: TRANSLATION

Exact Match Rates

- BLOOMZ forward: 224 / 184
- BLOOMZ back: 146 / 138
- GPT-4 forward: 267 / 218
- GPT-4 back: 199 / 174



Edit distance: GPT-4 had lower Levenshtein distance

Directionality: both models performed better in EN→ZH than ZH→EN

RESULTS: SEMANTIC PRESERVATION

Embedding-based semantic consistency

- **compares original and translated pairs in multilingual Sentence-BERT space**
- **embedding-based comparison shows whether semantic relations between the two words across translations are preserved, not just whether strings match**

Metric	BLOOMZ	GPT-4
Embedding consistency	0.711	0.782

TAKEAWAYS

- SimLex-style resources can be repurposed for cross-lingual LLM evaluation
- the framework captures different aspects of semantic consistency:
  - similarity alignment
  - shift tracking
  - lexical stability
  - embedding drift
- GPT-4 is substantially stronger than BLOOMZ across these diagnostics
- BLOOMZ shows basic multilingual competence, but weak fine-grained semantic stability.

UNIVERSITEIT VAN AMSTERDAM

LIMITATIONS&FUTURE WORK

Limitations

- **lexical / decontextualized setting only**
- **English–Mandarin case study only**
- **one annotator for manual verification**
- **possible benchmark contamination risk due to SimLex-style datasets being around for a long time**

Future work

- **more annotators**
- **more multilingual LLMs**
- **more language pairs**
- **especially under-resourced and non-English-centered pairs**
- **further investigation of why BLOOMZ compresses its scale.**

THANK YOU