Compositional generalisation in LLMs through multilingual lenses

Dieuwke Hupkes
Meta

Compositionality decomposed, how do neural networks generalise?

https://www.jair.org/index.php/jair/article/view/11674

IJCAI-JAIR Awards

The Annual IJCAI-JAIR Best Paper Prize is awarded to an outstanding paper published in JAIR in the preceding five calendar years. The prize committee is comprised of associate editors and members of the JAIR Advisory Board; their decision is based on both the significance of the paper and the quality of presentation. The recipient(s) of the award receives a prize of US$500 (to be split amongst the authors of a co-authored paper). Funding for this award was provided by the International Joint Conferences on Artificial Intelligence.

Compositionality Decomposed: How do Neural Networks Generalise?

PDF

2025 Prize

Dieuwke Hupkes, Verna Dankers, Mathijs Mul and Elia Bruni

Citation: *This work addresses the fundamental question of whether neural networks can exhibit compositional generalization, a cornerstone of human cognition and symbolic reasoning. Bridging philosophical and linguistic theories with contemporary machine learning, the authors propose a suite of five rigorous, task-independent tests that define and probe compositional generalization along multiple dimensions. Their methodology reveals key strengths and weaknesses in widely-used neural architectures and has become a touchstone for research on the limits of deep learning. This work has helped establish compositional generalization as a central empirical topic in AI and continues to influence both theoretical inquiry and practical model evaluation.*

Compositionality decomposed, how do neural networks generalise?

https://www.jair.org/index.php/jair/article/view/11674

Verna Dankers

Mathijs Mul

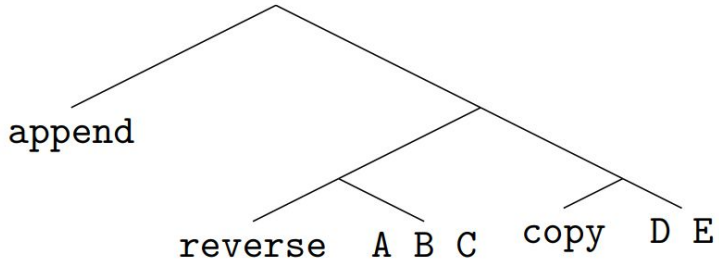
Elia Bruni

The data

PCFG SET = PCFG-generated String Edit Task

Unary functions: reverse, swap, copy, ...
Binary functions: prepend, append, remove_first, ...
Characters: A, B, C, ...

append reverse A B C , copy D E ⇒ C B A D E



The tests

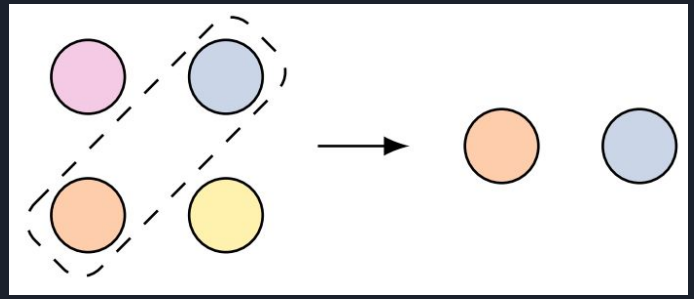
Systematicity

Productivity

Substitutivity

Localism

Overgeneralisation



withhold function

compositions from training

```
swap repeat
append remove_second
repeat remove_second
append swap
```

The tests

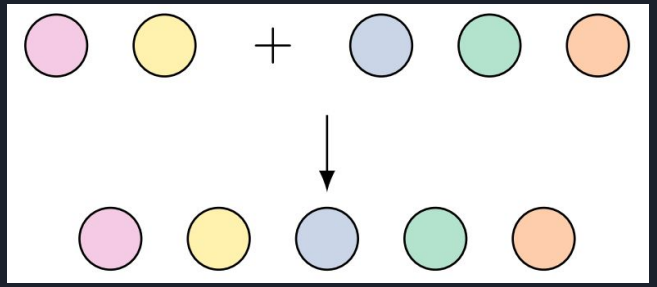
Systematicity

Productivity

Substitutivity

Localism

Overgeneralisation



```
train: up to 8
functions
test: 9+ functions
```

The tests

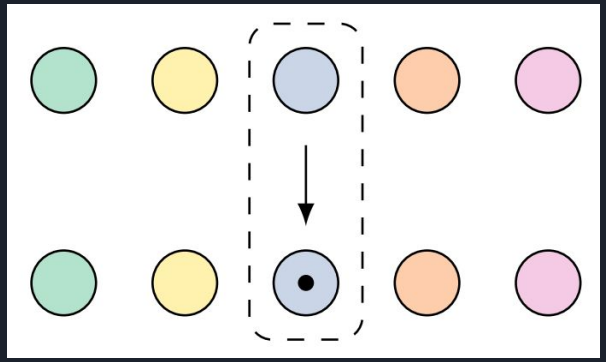
Systematicity

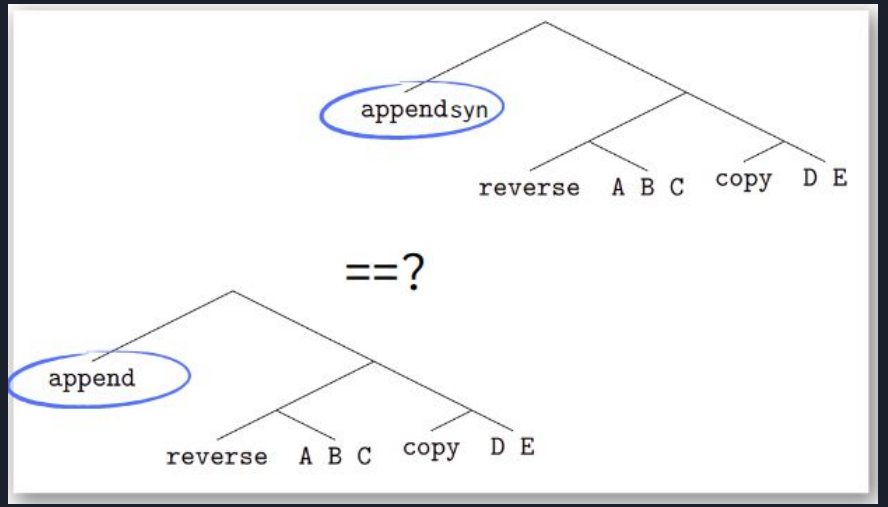
Productivity

Substitutivity

Localism

Overgeneralisation





The tests

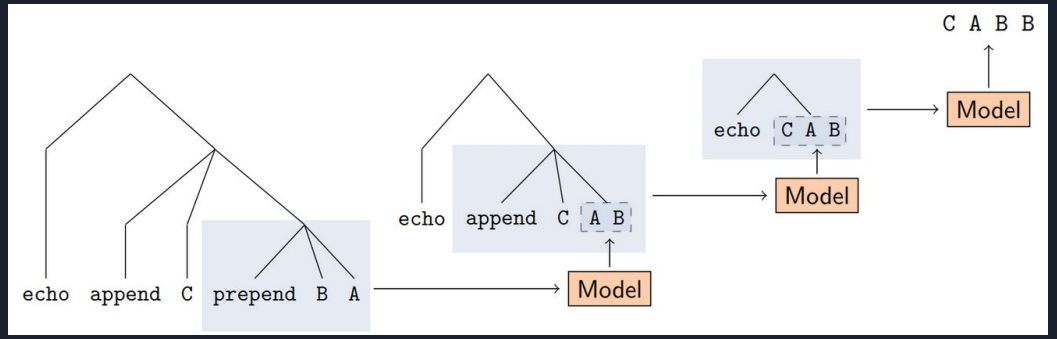
Systematicity

Productivity

Substitutivity

Localism

Overgeneralisation



The tests

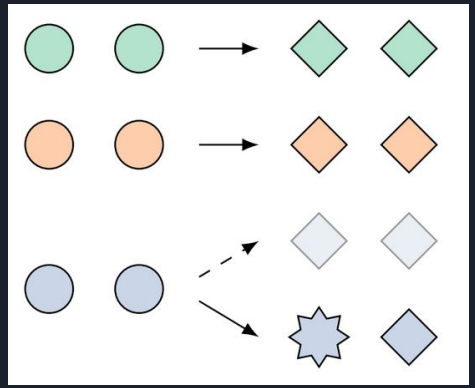
Systematicity

Productivity

Substitutivity

Localism

Overgeneralisation



introduce non-compositional exceptions to study generalisation vs memorisation

Input	Remapped to
reverse echo A B C	echo copy A B C
prepend remove_first A , B , C	remove_second append A , B , C
echo remove_first A , B C	copy append A , B C
prepend reverse A B , C	remove_second echo A B , C

Experiment	LSTMS2S	ConvS2S	Transformer
Task accuracy*	0.79 ± 0.01	0.85 ± 0.01	0.92 ± 0.01
Systematicity*	0.53 ± 0.03	0.56 ± 0.01	0.72 ± 0.00
Productivity*	0.30 ± 0.01	0.31 ± 0.02	0.50 ± 0.02
Substitutivity, *equally distributed*†	0.80 ± 0.00	0.95 ± 0.00	0.98 ± 0.00
Substitutivity, *primitive*†	0.60 ± 0.01	0.58 ± 0.01	0.90 ± 0.00
Localism†	0.46 ± 0.00	0.59 ± 0.01	0.54 ± 0.02
Overgeneralisation*	0.68 ± 0.04	0.79 ± 0.06	0.88 ± 0.07

Compositional evaluation shows large performance drops compared to i.i.d. evaluation...

...and models don’t behave local, don’t treat synonyms as equals, and can memorise exceptions to rules

Ripple effects: wide adoptation of tests

2020

Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language?

2025

Pretraining Frequency Predicts Compositional Generalization of CLIP on Real-World Tasks

Thaddäus Wiedemer[1,2,3*] Yash Sharma[1,2,3*] Ameya Prabhu[2,3] Matthias Bethge[2,3,4] Wieland Brendel[1,2,4]

To measure compositional generalization, we follow Hupkes et al. [12] and retain only test samples containing multiple concepts $o_1, ..., o_n$, where

(i) the model has been familiarized with each concept o_i only in the absence of $o_{j \neq i}$,

(ii) the combination $o_1, ..., o_n$ is plausible.

2020

2025

Towards a Comparative Framework for Compositional AI Models.

Tiffany Duneau

Quantinuum, University of Oxford

July 8, 2025

4 Compositional generalisation

4.1 Productivity . . . . . 4.2 Substitutivity . . . . . 4.3 Systematicity . . . . . 4.4 Overgeneralisation . . . . .

2025

Out-of-distribution generalisation is hard: evidence from ARC-like tasks

George Dimitriadis
Sainsbury Wellcome Centre
Gatsby Computational Neuroscience Unit
University College London
London
g.dimitriadis@ucl.ac.uk

Spyridon Samothrakis
School of Computer Science and Electronic Engineering
University of Essex
Colchester
ssamot@essex.ac.uk

Systematicity in representation

In this work, we focus exclusively on systematicity as defined by Hupkes et al. "This ability concerns the recombination of known parts and rules: anyone who understands a number of complex expressions also understands other complex expressions that can be built up from the constituents and syntactical rules employed in familiar expressions". We argue that an algorithm using systematicity

2022

for Compositional Generalization on gSCAN

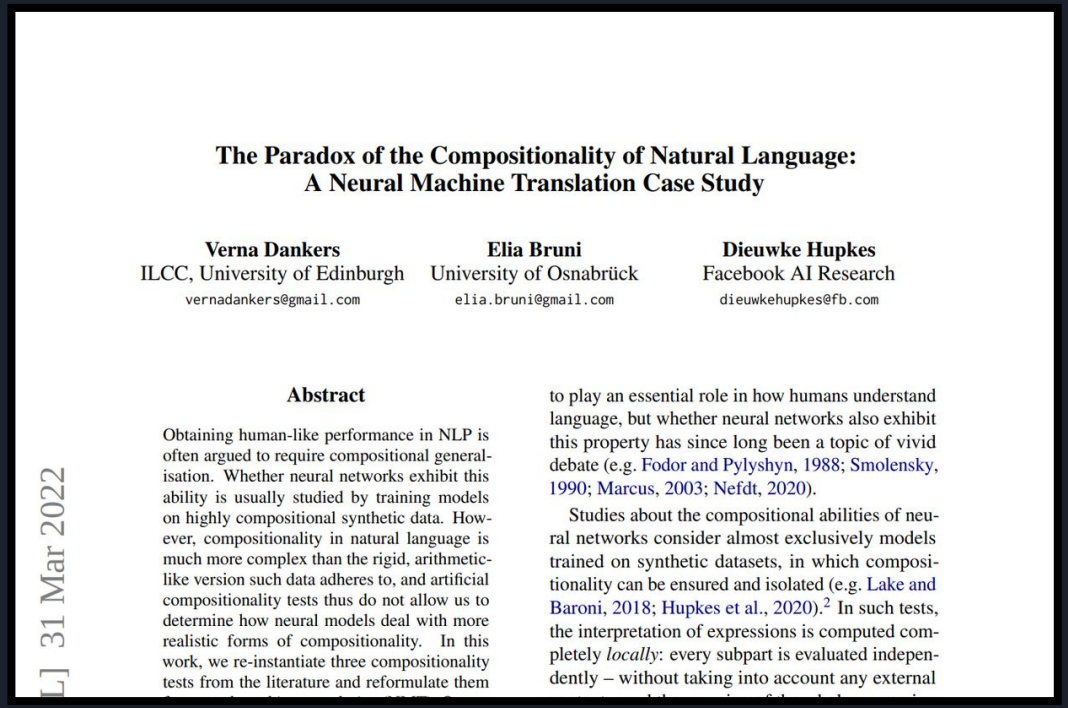
2023

How Compositional is a Model?

Parikshit Ram[1], Tim Klinger[1], Alexander G. Gray[1]
[1]IBM Research
parikshit.ram@ibm.com, tklinger@us.ibm.com, alexander.gray@ibm.com

sequences. Other properties are localism, substitutivity and over-generalization, but we will focus on systematicity and productivity here. Note that, while compositionality is a prop-

And a successful follow-up, rethinking strict compositional generalisation for natural language



The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study

Verna Dankers
ILCC, University of Edinburgh
vernadankers@gmail.com

Elia Bruni
University of Osnabrück
elia.bruni@gmail.com

Dieuwke Hupkes
Facebook AI Research
dieuwkehupkes@fb.com

Abstract

Obtaining human-like performance in NLP is often argued to require compositional generalisation. Whether neural networks exhibit this ability is usually studied by training models on highly compositional synthetic data. However, compositionality in natural language is much more complex than the rigid, arithmetic-like version such data adheres to, and artificial compositionality tests thus do not allow us to determine how neural models deal with more realistic forms of compositionality. In this work, we re-instantiate three compositionality tests from the literature and reformulate them

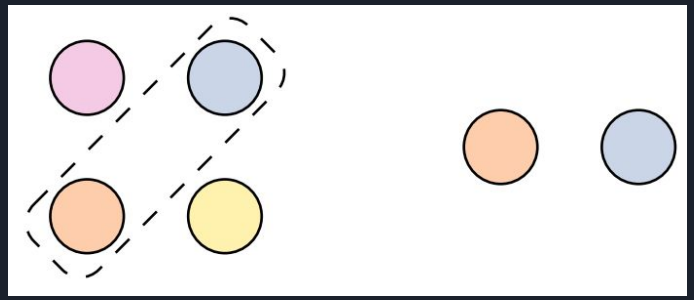
to play an essential role in how humans understand language, but whether neural networks also exhibit this property has since long been a topic of vivid debate (e.g. Fodor and Pylyshyn, 1988; Smolensky, 1990; Marcus, 2003; Nefdt, 2020).

Studies about the compositional abilities of neural networks consider almost exclusively models trained on synthetic datasets, in which compositionality can be ensured and isolated (e.g. Lake and Baroni, 2018; Hupkes et al., 2020).[2] In such tests, the interpretation of expressions is computed completely *locally*: every subpart is evaluated independently – without taking into account any external

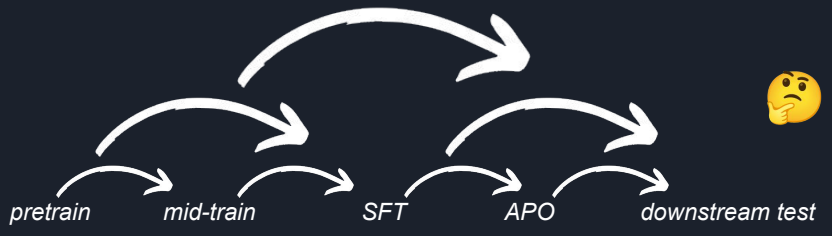
But...
compositional evaluation
has changed

Compositional generalisation in LLMs

Controlled train-test splits ❌



train **test**

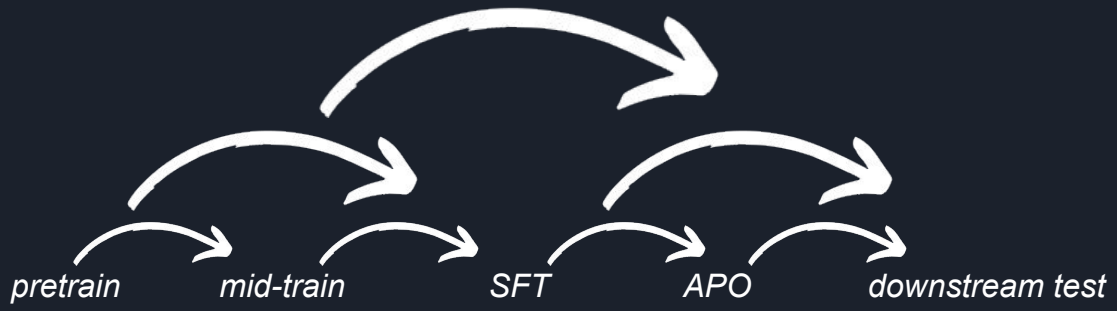


🤔

Compositional generalisation in LLMs

Controlled train-test splits ❌

Access to or knowledge about the training data ❌✅



Compositional generalisation in LLMs

Controlled train-test splits ❌

Access to or knowledge about the training data ❌✅

Compositional generalisation in LLMs

Controlled train-test splits	❌
Access to or knowledge about the training data	❌✅
Tools to analyse the training data	?

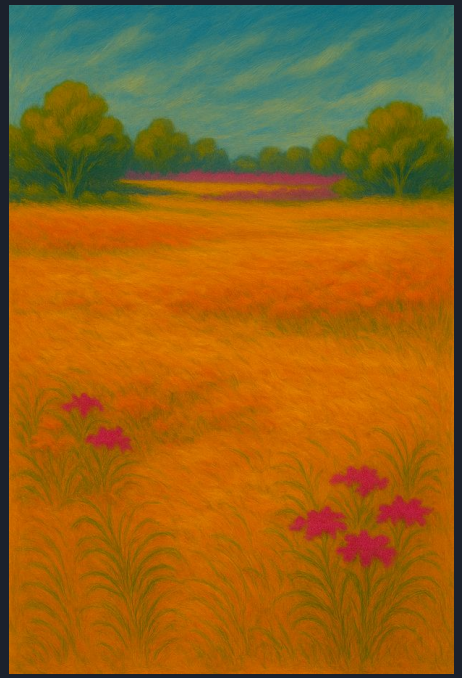
Compositional generalisation in LLMs

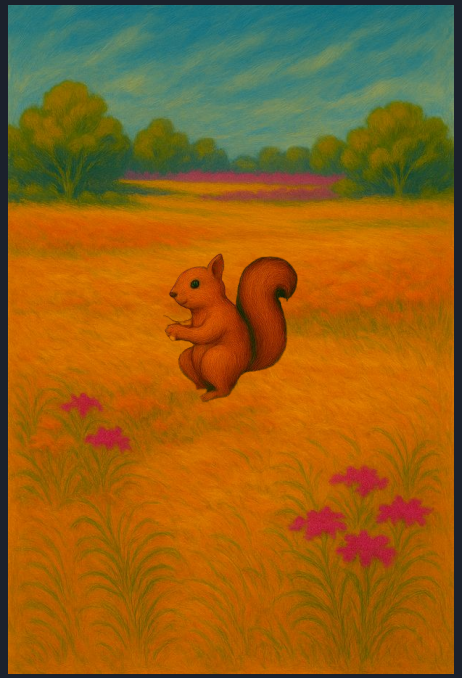
Controlled train-test splits	❌
Access to or knowledge about the training data	❌✅
Tools to analyse the training data	?
Clear separation of form and meaning	❌

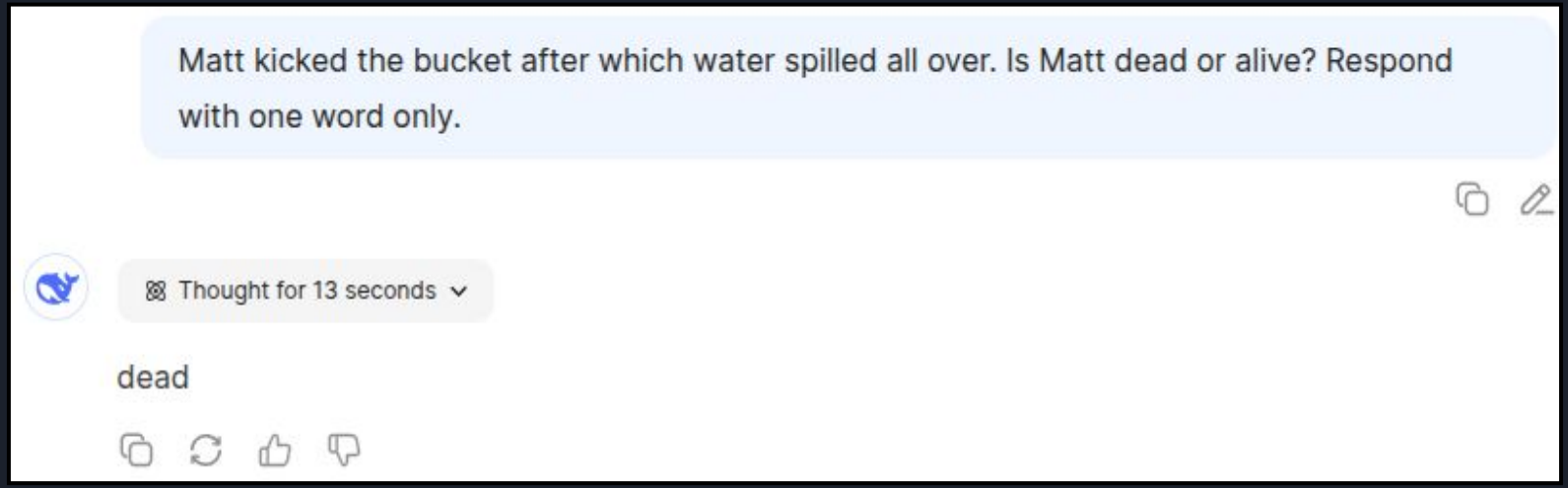
So... how do we evaluate (compositional) generalisation in LLMs?

Do we still need to evaluate compositionality?

In an orange meadow, a squirrel is driving a carriage with six wheels






Matt kicked the bucket after which water spilled all over. Is Matt dead or alive? Respond with one word only.
Thought for 13 seconds
dead

Compositional generalisation in LLMs

Controlled train-test splits	❌
Access to or knowledge about the training data	❌✅
Tools to analyse the training data?	?
Clear separation of form and meaning	❌

So… how do we evaluate (compositional) generalisation in LLMs?

Using multilingual consistency to evaluate (compositional?) generalisation

- Disentanglement of form and meaning
- Natural distribution shifts
- “Free” meaning preserving transformations

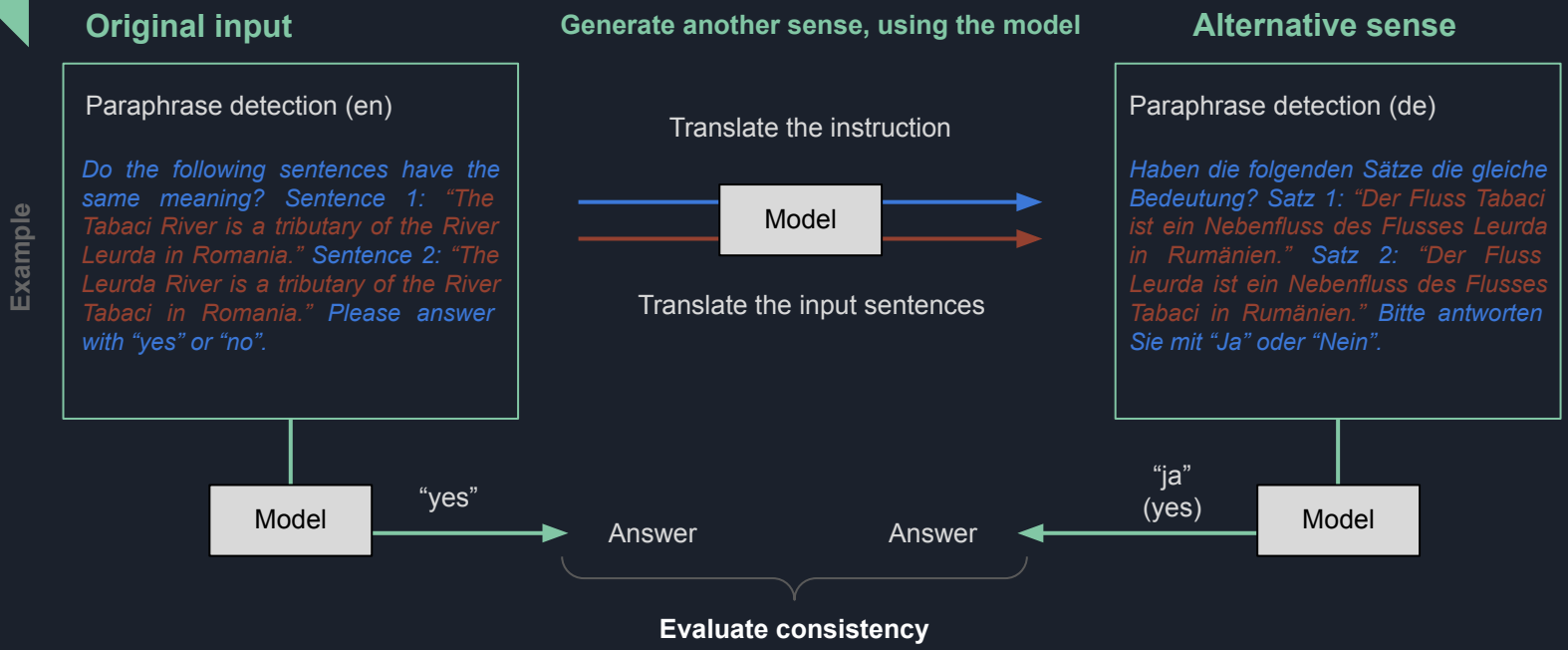
From Form(s) to Meaning: Probing the Semantic Depths of Language Models Using Multisense Consistency

https://direct.mit.edu/coli/article/50/4/1507/123794/From-Form-s-to-Meaning-Probing-the-Semantic-Depths

Xenia Ohmer

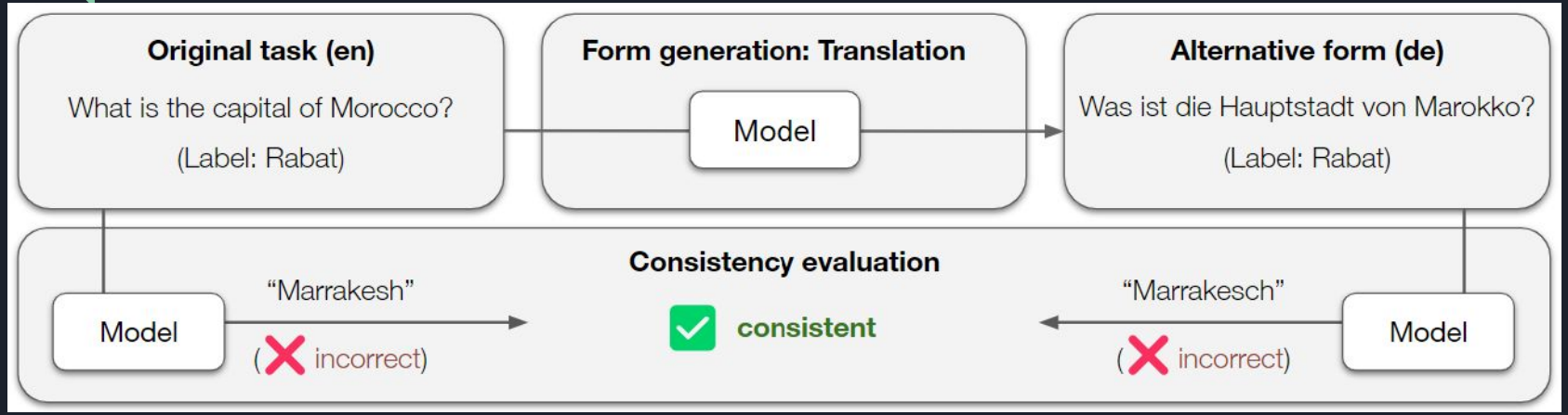
Elia Bruni

Consistency across ‘representations’



From form (s) to meaning: Probing the semantic depths of language models using multisense consistency
Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses
Ohmer et al. (2023, 2024)

Multisense consistency paradigm

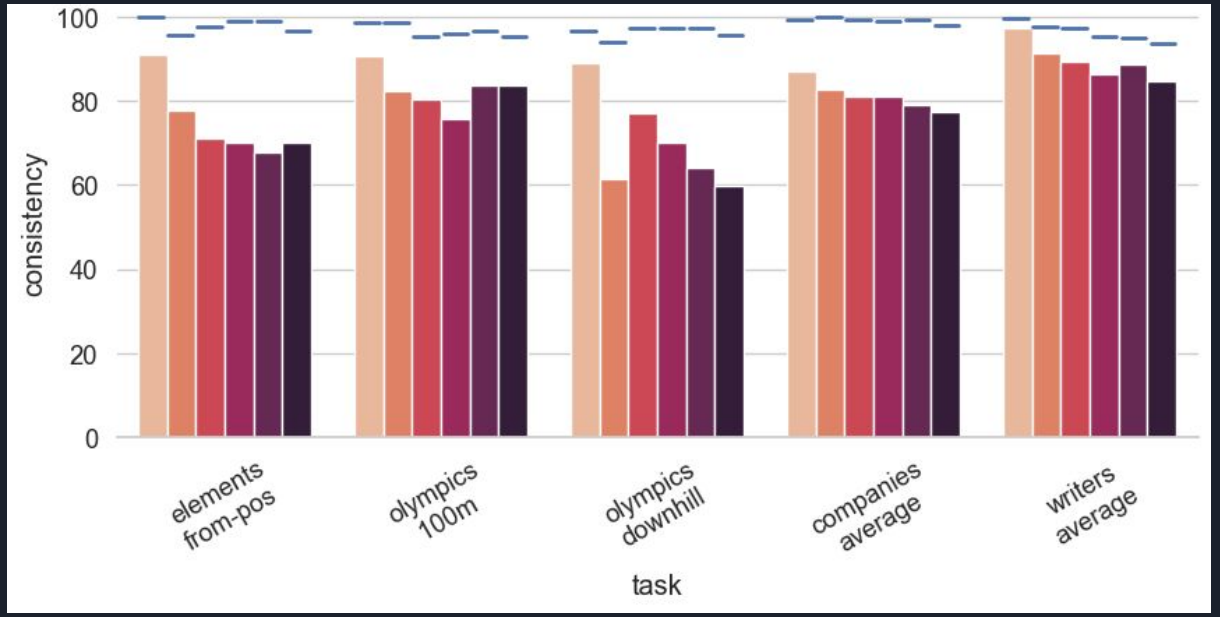


If a model is not self consistent (correct or incorrect) across the two questions, can it be compositional?

Datasets

dataset	subtask	N	template / example
ELEMENTS	FROM-ELEMENT	90	"What is the atomic number of the chemical element **He**?"
	FROM-POSITION	90	"What is the atomic number of the chemical element in period **5** and group **7**?"
OLYMPICS	100M	148	"Who won the **gold** medal in the **men's** 100 meters at the **2000** Summer Olympics?"
	DOWNHILL	117	"Who won the **bronze** medal in the **women's** downhill competition at the **1976** Winter Olympics?"
WRITERS	-	$186 \times 5 = 930$	"In what year was the writer **Friedrich Schiller** born?"
COMPANIES	-	$100 \times 5 = 500$	"In what city does **Airbus SE** have its headquarters?"

ChatGPT results: Simple facts



→ Even on simple factual questions the model generates inconsistent responses.

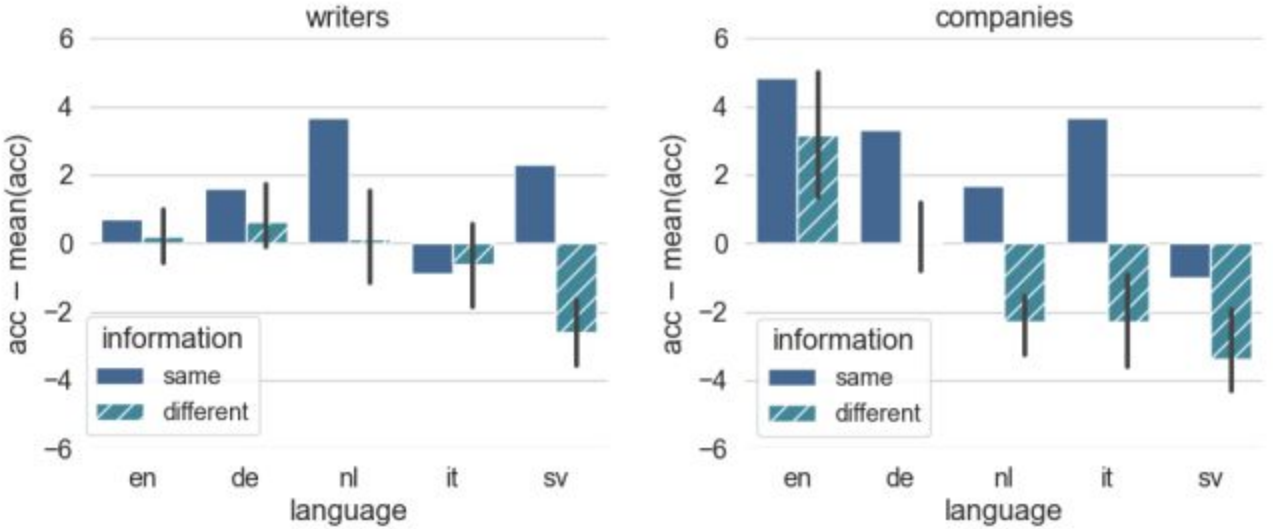


Figure 8: **Language-dependent knowledge for the SIMPLE FACTS dataset.** For each language, we compute how its accuracy when asked asked about information matching that language compares to its accuracy when asked about information not matching that language (e.g. asking about Dutch writers in Dutch vs in Swedish), compared to the overall averages for those groups. Generally, the model has higher accuracy when the prompt language and requested information pertain to the same country (plain bars) than when it is asked in a non-matching language (hatched bars).

Consistency for correct and incorrect examples

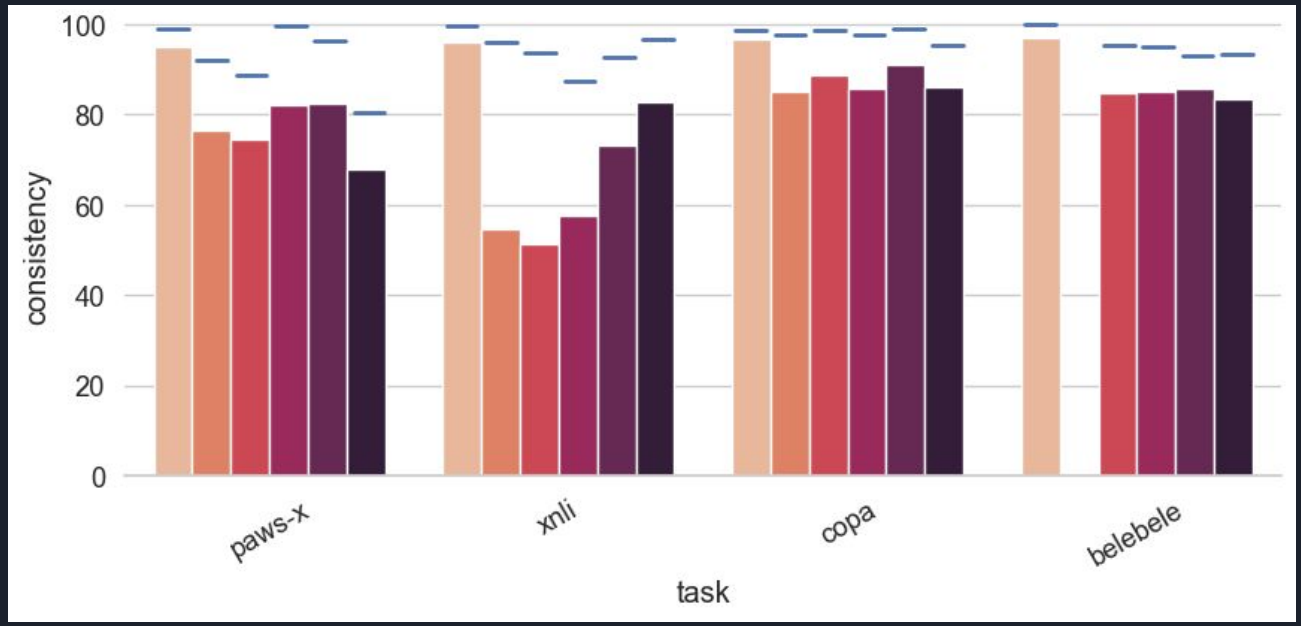
		T_{en}	$T_{en \to de}$	$T_{en \to zh}$	$T_{de \to en}$	$T_{zh \to en}$
PAWS-X	consistency all	0.99	0.84	0.76	0.86	0.70
	consistency correct	0.99	0.89	0.78	0.92	0.82
	consistency incorrect	0.98	0.67	0.71	0.72	0.52
XNLI	consistency all	0.98	0.74	0.67	0.63	0.67
	consistency correct	0.99	0.77	0.71	0.83	0.80
	consistency incorrect	0.96	0.66	0.57	0.45	0.50

Table 5: Detailed consistencies for the core experiment as well as for a baseline of two different runs with T_{en}. Listed are the consistency across all responses (consistency all), as well as the consistency across responses that were correct (consistency correct) and responses that were incorrect (consistency incorrect) on the source task.

Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses
Ohmer et al. (2023,)

Consistency on NLU benchmarks

ChatGPT results: NLU benchmarks



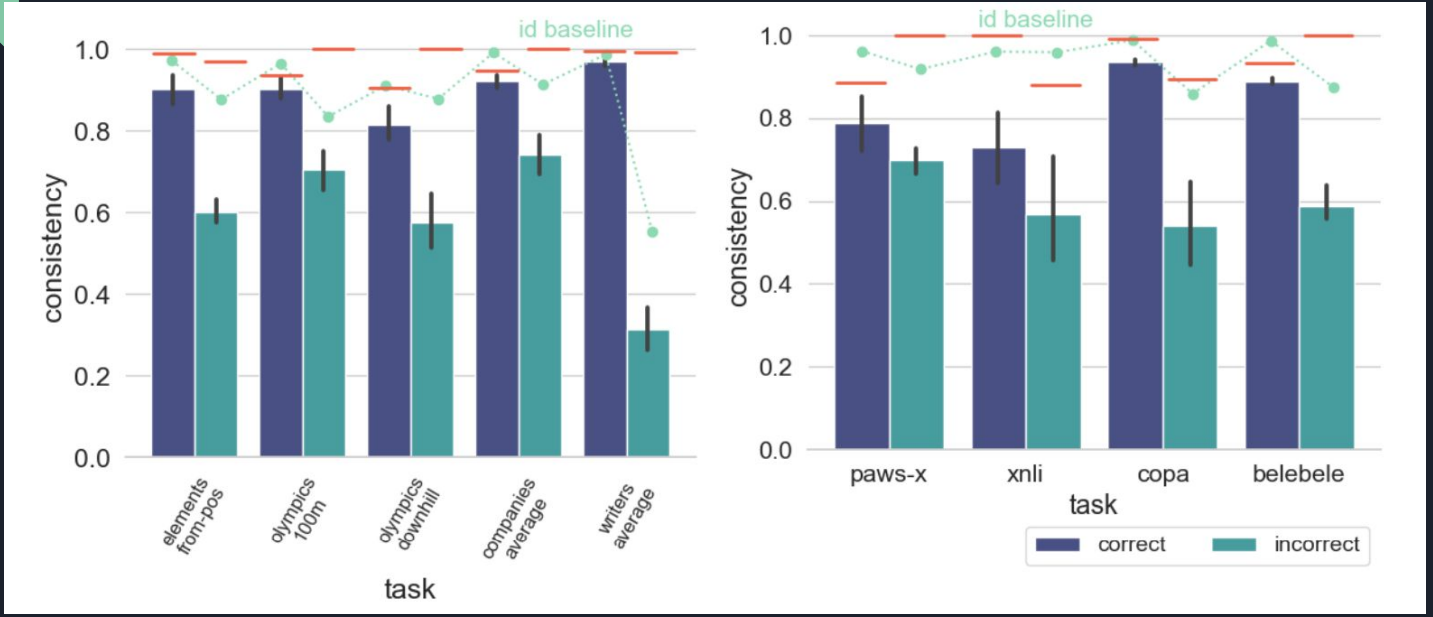
→ These inconsistencies are also evident on standard NLU benchmarks for paraphrase detection, NLI, commonsense reasoning, and knowledge tests.

Consistency and translation quality

		bleu	rouge1	rouge2	rouge-l	comet-22
paws	de^T	57.5	0.81	0.65	0.77	0.85
xnli	de^T	41.9	0.69	0.49	0.66	0.84
copa	it^T	40.9	0.66	0.45	0.64	0.86
belebele	de^T	41.1	0.69	0.46	0.63	0.84
	it^T	38.1	0.69	0.44	0.61	0.85
	nl^T	34.3	0.68	0.40	0.57	0.85
	sv^T	44.0	0.73	0.53	0.68	0.86

We consider the quality of the translation to different senses, according to commonly used metrics. All scores are high, suggesting that the model's inconsistencies are not driven by an inability to translate.

Consistency and correctness



Examples that are consistent and incorrect provide stronger evidence for a form-independent meaning understanding than consistent correct examples. The large difference between consistent correct and consistent incorrect thus indicates that some of the consistent correct examples were correct independently.

MultiLoKo: a multilingual local knowledge benchmark for LLMs

https://arxiv.org/abs/2504.10356

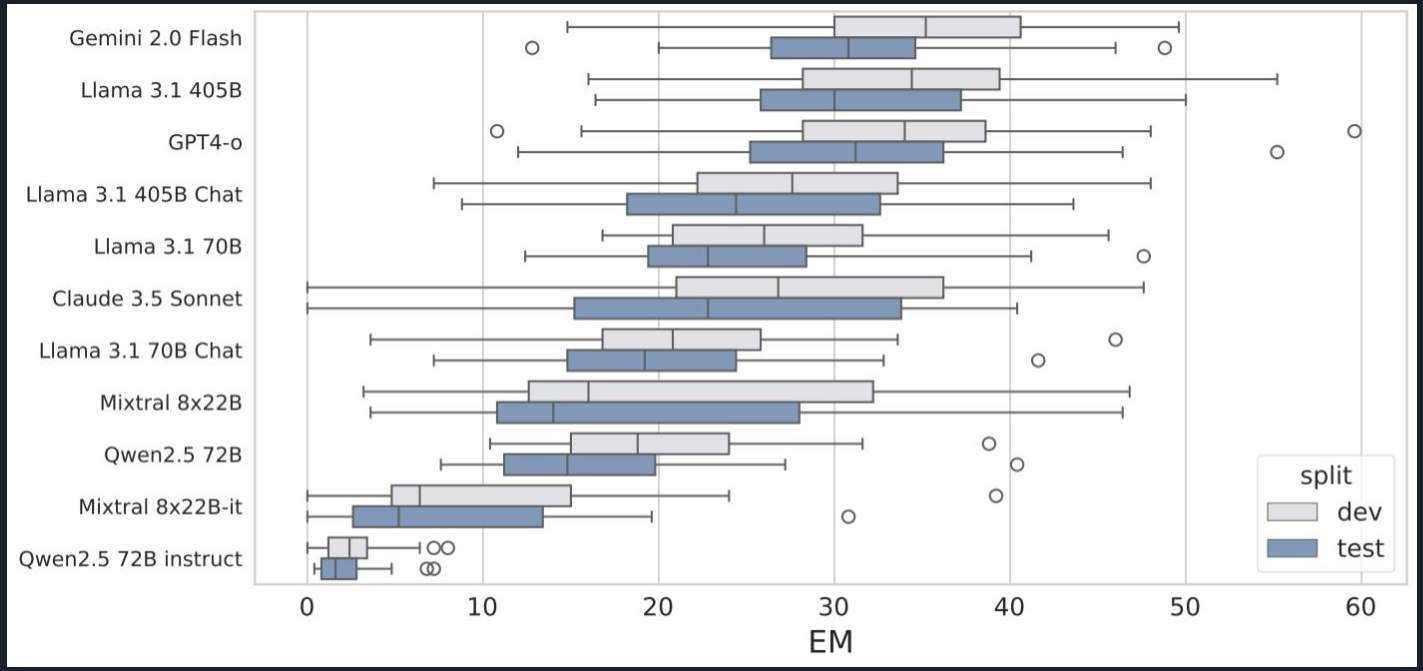
Nikolay Bogoychev

MultiLoKo: a multilingual local knowledge benchmark for LLMs

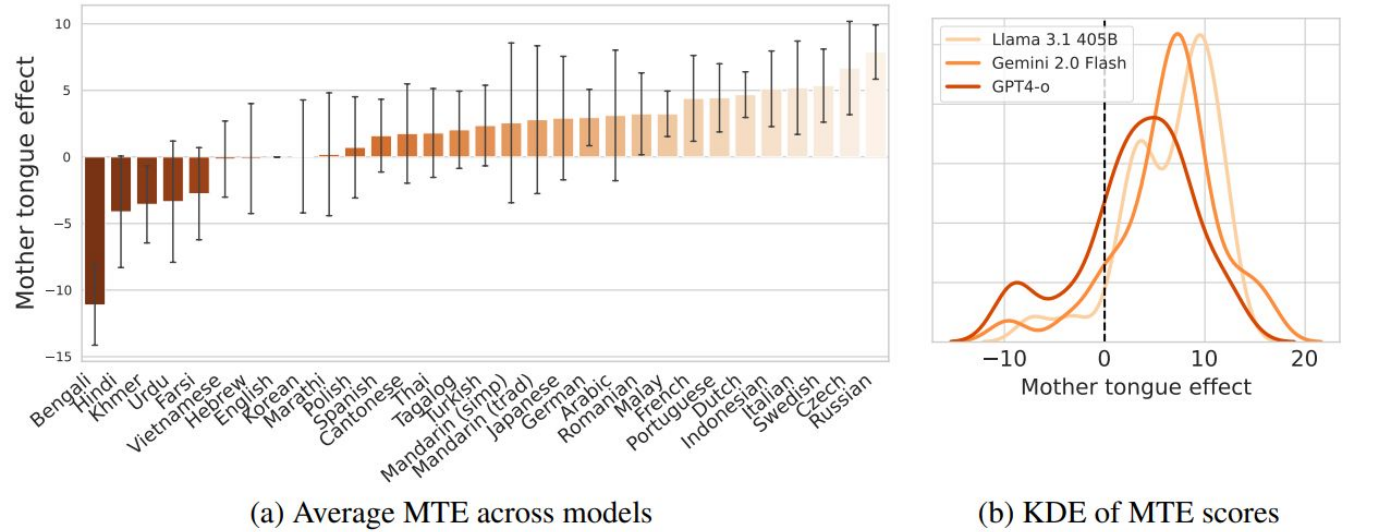
- **31 languages**
- **500 questions per language**, spread out over a validation and test set
- Sourced **independently for each language**, thus pertaining to **locally relevant knowledge**
- Includes **human translations** as well as **machine translations** (GT) for all non English language back to English, and for the English data to all other languages

https://huggingface.co/datasets/facebook/multiloko

Models don’t perform well, and the split is OOD



Poor disentanglement between form and meaning



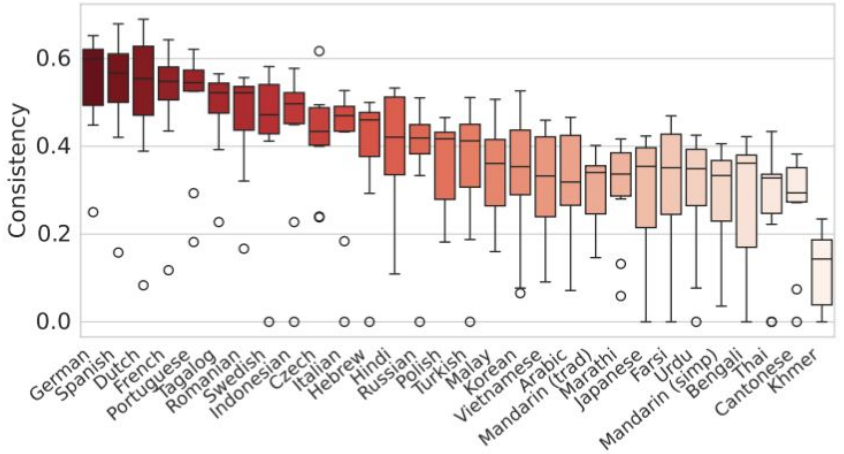
(a) Average MTE across models

(b) KDE of MTE scores

Figure 2: **Mother tongue effect dev.** (a) Per language MTE for MultiLoKo dev, indicating the difference between questions asked in the mother tongue (locally relevant) and in English. Error bars indicate 2 times standard error across all models, excluding Claude 3.5 Sonnet. (b) KDE plot of the distribution of MTE scores for the top-3 performing models.

Model	Consistency
Gemini 2.0 Flash	0.46± 0.04
Llama 3.1 405B	0.46± 0.04
Llama 3.1 70B	0.45± 0.03
GPT4-o	0.45± 0.05
Llama 3.1 405B Chat	0.42± 0.04
Qwen2.5 72B	0.40± 0.04
Llama 3.1 70B Chat	0.40± 0.04
Mixtral 8x22B	0.36± 0.05
Mixtral 8x22B-it	0.21± 0.05
Qwen2.5 72B instruct	0.08± 0.03

(a) Consistency scores per model



(b) Consistency scores per language

Figure 3: **Consistency results dev.** (a) Average per-model consistency scores, ± 2 times the standard error across languages. (b) Boxplot of model consistency scores per language, indicating the relative overlap of correctly answered questions when asked in the mother tongue vs in English.

Conclusion

- Using “traditional” compositionality tests is sheer impossible for LLMs
- With multilinguality we can assess if there is disentanglement between form and meaning, and rely on natural distribution shifts, for now…
- Tests utilising this suggest that there is still some improvements to be made
- What is next?

Thanks!

Dieuwke Hupkes
Meta

www.dieuwkehupkes.nl
dieuwkehupkes@meta.com