MODELING: A Novel Dataset for Testing Linguistic Reasoning in Language Models

Nathan A. Chi[1], **Teodor Malchev**[2], **Riley Kong**[3], **Ryan A. Chi**[1], **Lucas Huang**[4], **Ethan A. Chi**[1,5], **R. Thomas McCoy**[4,6], **Dragomir Radev**[4]
[1]Stanford University [2]Harvard University [3]MIT [4]Yale University
[5]Hudson River Trading [6]Princeton University
nathanchi@cs.stanford.edu

1 Introduction

Large language models (LLMs) perform well on (at least some) evaluations of both few-shot multilingual adaptation (Lin et al., 2022) and reasoning (Bubeck et al., 2023). However, evaluating the intersection of these two skills—**multilingual few-shot reasoning**—is difficult: even relatively low-resource languages can be found in large training corpora, raising the concern that when we intend to evaluate a model's ability to generalize to a new language, that language may have in fact been present during the model's training. If such **language contamination** (Blevins and Zettlemoyer, 2022) has occurred, apparent cases of few-shot reasoning could actually be due to memorization.

Towards understanding the capability of models to perform multilingual few-shot reasoning, we propose MODELING, a benchmark of *Rosetta stone puzzles* (Bozhanov and Derzhanski, 2013). This type of puzzle, originating from competitions called Linguistics Olympiads, contain a small number of sentences in a target language not previously known to the solver. Each sentence is translated to the solver's language such that the provided sentence pairs uniquely specify a single most reasonable underlying set of rules; solving requires applying these rules to translate new expressions (Figure 1). MODELING's languages are chosen to be extremely low-resource such that the risk of training data contamination is low, and unlike prior datasets (Şahin et al., 2020), it consists entirely of problems written specifically for this work, as a further measure against data leakage. Empirically, we find evidence that popular LLMs do not have data leakage on our benchmark (Section 2.1).

Here are some phrases in Ayutla Mixe:
Ëjts nexp. → *I see.*
Mejts mtunp. → *You work.*
Juan yë'ë yexyejtpy. → *Juan watches him.*
Yë'ë yë' uk yexpy. → *He sees the dog.*
Ëjts yë' maxu'unk nexyejtpy. → *I watch the baby.*

Now, translate the following phrases.
Yë' maxu'unk yexp. → ***The baby sees.***
The baby watches the dog. → **Yë' maxu'unk yë' uk yexyejtpy**.

Figure 1: A representative sample puzzle (based on Ayutla Mixe, which is spoken in Oaxaca, Mexico). Providing the answers (in **bolded red**) requires using the labeled pairs to reason about word meanings, morphology (the *-y* suffix), and word order—all in an extremely low-resource environment (there appear to be fewer than 3 pages in Ayutla Mixe on the Internet, so models are unlikely to have had substantial experience with the language beyond the examples shown here).

2 Dataset

MODELING comprises 48 Rosetta Stone puzzles based on 19 extremely low-resource languages from diverse regions. All problems were written by authors familiar with linguistics problems and were test-solved and rated for difficulty by two International Linguistics Olympiad medalists (Table 2). It includes 272 questions falling into four types, each testing a model's ability to handle a distinct element of linguistic typology:

1. **noun-adjective order** problems, which require determining the relative ordering of nouns and adjectives;
2. **word order** problems, which require determining the relative ordering of subject (S), verb (V), and object (O);
3. **possession** problems, which require reasoning about possessive morphology;
4. **semantics** problems, which require aligning a set of non-English semantic compounds to their English translations (e.g. En. "alcohol" = Wik-Mungkan *ngak way*, lit. "bad water").

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 113–119

2.1 Data leakage

Because all the problems that we designed were newly written, models could not have encountered these puzzles in their training data. Nonetheless, it is possible that they may have encountered the specific words and phrases that we evaluate on.[1] To address this concern, we ran a baseline in which we evaluated all models without any target/reference pairs, prompting them to use "`existing knowledge of the language`" to translate the statements. Answering such questions is impossible without prior knowledge of the target language, so nonzero accuracy would suggest the presence of data leakage (Huang et al., 2022). The performance of all models in this setting is 0%, suggesting that the use of very low-resource languages successfully avoids data leakage.

3 Experiments

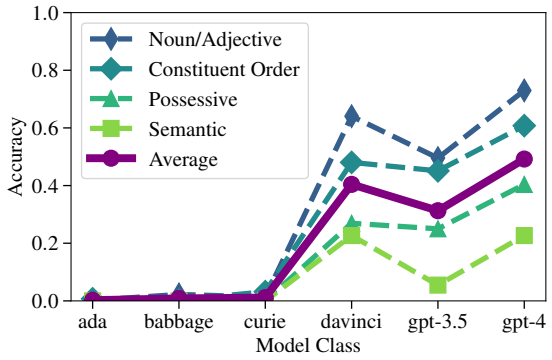
We evaluated six GPT models (GPT-3 {Ada, Babbage, Curie, Davinci}; GPT-3.5; and GPT-4) on our dataset on August 13, 2023 (Brown et al., 2020; OpenAI, 2023). We evaluated under the following conditions: **minimal prompt** (a brief, basic prompt specifying the task); **hand-tuned prompt** (a prompt fine-tuned by an International Linguistics Olympiad medalist); **basic chain-of-thought** (Kojima et al., 2022) (which encourages models to think step-by-step); and **full chain-of-thought** (Wei et al., 2022) (which provides an example of reasoning step-by-step). We report exact-match accuracies taken over all individual questions.

We observe strong performance from Davinci, GPT-3.5, and GPT-4 (Table 1). Across prompting approaches, we observe roughly similar accuracies. However, smaller models (Ada, Babbage, Curie) perform much worse, with accuracies near 0. All three of the large, accurate models (GPT-3-Davinci, GPT-3.5, GPT-4) struggle with particular problem categories, with possessive and semantic problems being harder than noun/adjective ordering and basic word order (Figure 2a). Finally, model performance closely follows human difficulty ratings (Figure 3a), suggesting that as large models continue to improve, we can scale our benchmark by producing more challenging problems (even the hardest problems in our benchmark are relatively easy by Linguistics Olympiad standards).

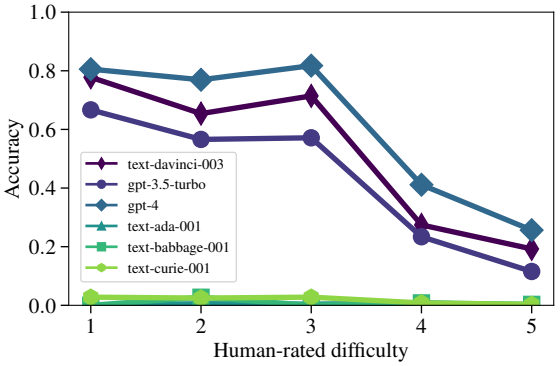
[1] e.g., perhaps their training data included the Ayutla Mixe sentence *Yë' maxu'unk yexp* shown in Figure 1.

Model	Minimal prompt	Hand-tuned prompt	Basic CoT	Full CoT
Ada	.000	.004	.011	.000
Babbage	.011	.011	.004	.018
Curie	.015	.018	.015	.022
Davinci	.496	.485	.490	.514
GPT-3.5	.404	.412	.401	.397
GPT-4	**.588**	**.591**	**.589**	**.607**

Table 1: Accuracy (exact match) of several large language models (LLMs) on MODELING. *CoT* stands for *chain of thought*.



(a) Accuracy across different language models on our dataset, reporting average score across all prompts.



(a) LLM accuracy on our dataset, bucketed by difficulty. The 3 larger models (Davinci, GPT-3.5, GPT-4) display relatively high accuracy, while the smaller models are close to zero.

4 Conclusion

We have introduced MODELING, a dataset designed to evaluate LLMs' capacity to reason analytically in unseen languages. We believe that the approach used to develop MODELING—given its use of languages that occur very rarely on the Internet and its capacity to be extended to more challenging cases—has a strong potential to serve as a durable approach for evaluating reasoning.

References

Keith Berry, Christine Berry, et al. 1999. *A description of Abun: a West Papuan language of Irian Jaya*. Pacific Linguistics.

Roger Blench and Mallam Dendo. Baŋgi me, a language of unknown affiliation in Northern Mali.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4.

Dmitry V. Bubrikh. 1949. Grammatika literaturnogo komi yazyka (grammar of the Komi literary language).

Anna Bugaeva. 2022. *Handbook of the Ainu Language*, volume 12. Walter de Gruyter GmbH & Co KG.

Matthew S Dryer et al. 1994. The discourse function of the Kutenai inverse. *Voice and Inversion*, pages 65–99.

John B Haviland. 1998. Guugu Yimithirr cardinal directions. *Ethos*, 26(1):25–47.

Jeffrey Heath. 2015. A grammar of Toro Tegu (Dogon), Tabi mountain dialect.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eugene S. Hunn, Akesha Baron, Roger Reeck, Meinardo Hernández Pérez, and Hermilo Silva Cruz. A sketch of Mixtepec Zapotec grammar.

Paulus Kievit. 2017. *A grammar of Rapa Nui*. Language Science Press.

Lyle M Knudson. 1975. A natural phonology and morphophonemics of Chimalapa Zoque. *Research on Language & Social Interaction*, 8(3-4):283–346.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Jonathan Lane. 2007. *Kalam serial verb constructions*. Pacific Linguistics.

Stephen C Levinson. 1997. Language and cognition: The cognitive consequences of spatial description in Guugu Yimithirr. *Journal of linguistic anthropology*, 7(1):98–131.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Carolyn J MacKay. 1994. A sketch of Misantla Totonac phonology. *International Journal of American Linguistics*, 60(4):369–419.

Teresa Ann McFarland. 2009. *The phonology and morphology of Filomeno Mata Totonac*. University of California, Berkeley.

Mary Beck Moser and Stephen Alan Marlett. 2005. *Comcáac quih yaza quih hant ihíip hac: cmiique iitom, cocsar iitom, maricáana iitom*. Plaza y Valdes.

OpenAI. 2023. Gpt-4 technical report.

Andrew Pawley. 2006. Where have all the verbs gone? Remarks on the organisation of languages with small, closed verb classes. In *11th Binnenial Rice Univerity Linguistics Symposium*, pages 16–18.

Rodrigo Romero-Méndez. 2009. *A reference grammar of Ayutla Mixe (Tukyo'm ayuujk)*. Ph.D. thesis, State University of New York at Buffalo.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Lyle Scholtz. 1967. Kalam verb phrase. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, 11(1):10.

Ineke Smeets. 2008. *A Grammar of Mapuche*. De Gruyter Mouton, Berlin, Boston.

Elaine Thomas. 1969. *A grammatical description of the Engenni language*. University of London, School of Oriental and African Studies (United Kingdom).

Edward Tregear and Stephenson Percy Smith. 1907. *Vocabulary and grammar of the Niue dialect of the Polynesian language*. J. Mackay, Government Printer.

Darrell T Tryon. 1995. *Comparative Austronesian dictionary: An introduction to Austronesian studies*. De Gruyter Mouton.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

A Dataset

A.1 Overview

Category	# Problems	# Questions	% Questions
Noun/Adj.	19	112	41%
Order	19	102	37%
Possessive	5	26	10%
Semantics	5	32	12%
Total	**48**	**272**	**100%**

Table 2: Dataset split by problem type (Section 2). We have 48 problems and a total of 272 questions.

A.2 Difficulty

Difficulty	# Problems	# Questions	% Questions
1	2	9	3%
2	16	91	34%
3	12	63	23%
4	6	31	11%
5	12	78	29%
Total	**48**	**272**	**100%**

Table 3: Distribution of difficulty levels over the dataset, as jointly evaluated on a Likert scale by two expert evaluators who have received medals at the International Linguistics Olympiad.

A.3 Orthography

B Prompts

Our four different prompting styles are illustrated inFigures 4 through 7.

C Data sources

Minimal-prompt

Here are some expressions in `Language` (a never-seen-before foreign language) and their translations in English:

`Language`: ...
English: ...

Given the above examples, please translate the following statements.

Figure 4: Minimal prompt.

Hand-tuned prompt

This is a translation puzzle. Below are example phrases in `Language` (a never-seen-before foreign language) as well as their English translations. Some test phrases follow them. Your task is to look closely at the example phrases and use only the information from them to translate the test phrases.

`Language`: ...
English: ...

Given the above examples, please translate the following statements.

Figure 5: Hand-tuned prompt.

Basic chain-of-thought

This is a translation puzzle. Below are example phrases in `Language` (a never-seen-before foreign language) as well as their English translations. Some test phrases follow them. Your task is to look closely at the example phrases and use only the information from them to translate the test phrases.

`Language`: ...
English: ...

Given the above examples, please translate the following statements. Let's think step by step in a logical way, using careful analytical reasoning to get the correct result.

Figure 6: Basic chain-of-thought prompt.

Full chain-of-thought

This is a translation puzzle. In a moment, you will use logic and analytical reasoning to translate from a never-seen-before language (Language) to English. As a training example, here are some expressions in Spanish and their translations in English.

1. Spanish: ventana roja
English: red window

2. Spanish: ventana azul
English: blue window

3. Spanish: manzana azul
English: blue apple

Using the above examples, translate the following.
Spanish: manzana roja

ANSWER: English: red apple

EXPLANATION: The first step we notice is that the word "ventana" must mean window because (1) the word "ventana" appears twice between sentences 1 and 2, and (2) the only word that appears twice in the English translation is "window." Next, we infer that "roja" must be "red" and "azul" must be "blue" by process of elimination. Next, we guess that in Spanish, the noun precedes the adjective because "ventana" comes before "roja" and "azul." Therefore, the noun in sentence 3 ("apple") must correspond to the word preceding the adjective ("manzana") in the Spanish translations. Putting this together, "manzana roja" must mean "red apple" in English.

Do you see how we're using logical and analytical reasoning to understand the grammar of the foreign languages step by step?

Figure 7: Full chain-of-thought prompt.

Language	Original	New
Ayutla Mixe	ë	eu
Bangime	ç	ch
Seri	ö	w
Rapa Nui	ā	aa

Table 4: Sample orthographic conversions.

Language	Family	ISO	#	Type	Source
Abun	West Papuan	kgr	1	POSS	Berry et al. (1999)
Ainu	Ainuic	ain	1	ORDER	Bugaeva (2022)
Ayutla Mixe	Mixe-Zoque	mxp	1	ORDER	Romero-Méndez (2009)
Bangime	Isolate	dba	7	NOUN-ADJ, ORDER	Blench and Dendo
Chimalapa Zoque	Mixe-Zoque	zoh	1	ORDER	Knudson (1975)
Toro-tegu Dogon	Niger–Congo	dtt	2	POSS	Heath (2015)
Engenni	Niger–Congo	enn	5	ORDER	Thomas (1969)
Guugu Yimithirr	Pama-Nyungan	kky	1	SEM	Haviland (1998); Levinson (1997)
Kalam	Kalam	kmh	1	SEM	Pawley (2006); Lane (2007), Scholtz (1967)
Komi-Zyrian	Permic	kpv	1	SEM	Bubrikh (1949)
Kutenai	Isolate	kut	1	SEM	Dryer et al. (1994)
Mapudungan	Araucanian	arn	4	NOUN-ADJ	Smeets (2008)
Misantla Totonac	Totonacan	tlc	1	NOUN-ADJ	MacKay (1994)
Mixtepec Zapotec	Oto-Manguean	zpm	4	NOUN-ADJ	Hunn et al.
Ngadha	Malayo-Polynesian	nxg	2	NOUN-ADJ	Tryon (1995)
Niuean	Malayo-Polynesian	niu	3	NOUN-ADJ	Tregear and Smith (1907)
Rapa Nui	Malayo-Polynesian	rap	7	NOUN-ADJ, ORDER	Kievit (2017)
Seri	Isolate	sei	4	NOUN-ADJ, ORDER, POSS, SEM	Moser and Marlett (2005)
Filomeno Mata Totonac	Totonacan	tlp	1	POSS	McFarland (2009)

Table 5: Problem Data Sources: sentences in MODELING were either taken directly from or written according to rules contained within the sources.

Figure 8: The 19 distinct languages included in the **MODELING** benchmark. Note that some languages have more than one problem.