TRACE-it: Testing Relative clAuses Comprehension through Entailment in ITalian: A CALAMITA Challenge

Dominique Brunato[1]

[1]Istituto di Linguistica Computazionale "A. Zampolli", CNR-ILC, ItaliaNLP Lab

Abstract

Introduced in the context of CALAMITA 2024 [1], TRACE-it (Testing Relative clAuses Comprehension through Entailment in ITalian) is a benchmark designed to evaluate the ability of Large Language Models (LLMs) to comprehend a specific type of complex syntactic construction in Italian: object relative clauses. In this report, we outline the theoretical framework that informed the creation of the dataset and provide a comprehensive overview of the linguistic materials used.

Keywords

Object Relative Clauses, Italian language, benchmark, syntactic assessment, entailment

1. Introduction and Motivation

TRACE-it (Testing Relative clAuses Comprehension through Entailment in Italian) is a benchmark designed to assess the ability of Large Language Models (LLMs) to comprehend complex sentences in Italian. Complex sentences, in this context, are defined as those containing a type of unbounded dependency, whose correct understanding requires the computation of a grammatical relationship between phrases that are pronounced in a position different from the one where they are interpreted.

These structures, also known as "filler-gap" constructions in psycholinguistics, pose significant challenges for human sentence processing, particularly pronounced when the "filler" (the pronounced element) is distant from the "gap" (the position where it is interpreted) [2, 3, 4, 5]. Examples of this include object-gap relationships, which occur in constructions such as relative clauses (1), cleft sentences (2), or wh-questions (3), like the following[1]:

1. Il giornalista che il senatore contestò ammise l'errore. [*The reporter who the senator attacked admitted the error.*]
2. E' il giornalista che il senatore contestò. [*It is the reporter that the senator attacked.*]
3. Quale giornalista il senatore contestò? [*Which reporter did the senator attack?*]

The higher complexity of these constructions compared to their subject counterparts –typically measured in terms of reading times and often accompanied by error rates in comprehension questions after reading– has been extensively studied and explained by formal linguistic theories and processing models [7, 4, 8, 6], including child language acquisition data [9, 10, 11]. This benchmark aims to determine whether LLMs encounter similar difficulties and to explore various factors that were shown to modulate this complexity for humans, such as altering the nature of the elements involved in the dependency in terms of grammatical and/or semantic features, as well as varying the distance between the filler and the gap.

In this respect, the proposed benchmark is part of a growing set of resources specifically designed for syntactic evaluation of neural language models, which are typically composed by minimal pairs of grammatical and non-grammatical sentences addressing a specific linguistic phenomenon that differs in the sentence (see [12, 13, 14, 15, 16], *i.a.*). To succeed, a model must score the grammatical sentence higher than its ungrammatical counterpart, either assigning a binary value or in terms of model perplexity. Two main resources in this respect are Corpus of Linguistic Acceptability (CoLA) [17] and BLiMP (Benchmark of Linguistic Minimal Pairs) [18], which include minimal pairs for various grammatical phenomena in English. Adaptations of these resources have been recently released also in other languages, Italian included. Notable examples include ITaCoLA [19], which is directly inspired by CoLA, and the dataset developed for the AcCompl-It task (Acceptability & Complexity Evaluation for Italian) held in the context of Evalita 2020 campaign [20].

While similar for purposes, the novelty of TRACE-it lies in its approach. Unlike previous benchmarks that have focused on testing LLMs' ability to distinguish between grammatical and ungrammatical sentences through minimal pairs or assigning a complexity score to such sentences, this benchmark introduces

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

dominique.brunato@ilc.cnr.it (D. Brunato)

0000-0003-3256-4794 (D. Brunato)

[1]Examples are taken from [6].

a more advanced task based on entailment. Instead of simply assessing grammaticality, the model is tasked with determining whether a given complex sentence logically entails a simpler yes/no implication. This approach would thus provide a more nuanced evaluation of the model's ability to understand deep syntactic structures, going beyond surface-level grammaticality to probe its comprehension of meaning.

The ability to grasp complex syntactic relationships, such as those present in filler-gap constructions, is fundamental to higher-order language tasks. For instance, summarization, information extraction, and question answering all depend on the model's capacity to correctly interpret sentence structure and meaning. By requiring the model to process complex syntactic dependencies, this benchmark aims to provide a further step towards more rigorous and meaningful evaluation of syntactic comprehension, with a specific focus on Italian. Moreover, TRACE-it contributes to the growing field of linguistically informed resources that enhance interpretability in NLP [21]. These benchmarks are essential for unraveling the linguistic competence implicitly encoded in neural network representations, and they can shed light on the similarities and differences between how humans and LLMs acquire, represent, and process linguistic knowledge [22, 23].

2. Challenge: Description

The proposed challenge focuses on evaluating LLMs' understanding of a precise linguistic structure in the Italian language: **restrictive object-extracted relative clauses** (ORCs). We specifically examine centre-embedded ORCs where both the relative head and the embedded subject are expressed as lexical noun phrases.

The assessment involves a **yes/no entailment task** in which the model is given two paired sentences. The first contains the target structure, and the second is a simple declarative sentence whose meaning may or may not be logically inferred from the first based on the syntactic relationship between the elements in the ORC. Specifically, the second sentence focuses either on the relative head (NP1) or the embedded subject (NP2) and has been designed according to the following criteria: When the focus is on NP1, the entailment is true if the second sentence presents NP1 as the active subject of the matrix verb of the main clause or as the passive subject of the embedded verb (see examples 1 and 2 in Table 1, respectively). The entailment is false if NP1 is shown as the active subject of the embedded verb or if the verb of the main clause is negated (see examples 3 and 4, respectively).

When the focus is on NP2, the entailment is true if the second sentence presents NP2 as the subject of the embedded verb (example 5). It is false if NP2 is the passive subject of the embedded verb or is presented as the subject of the main clause's verb (examples 6 and 7, respectively). In the majority of cases, the second sentence closely mirrors the lexical structure of the first, as the dataset is firstly designed to investigate syntactic entailment. However, in some instances, a paraphrase is used (e.g.. 8).

These criteria were almost equally balanced across the distinct portions of the whole dataset, which are detailed in the following section.

3. Data description

The benchmark consists of 566 sentence pairs, all structured to evaluate the comprehension of Object Relative Clauses (ORCs). While the task's main objective and the criteria for determining entailment between the two sentences in each pair remain constant, the dataset is divided into four main sections. Each section corresponds to a distinct type of ORC in the first sentence, differentiated by specific conditions that characterize the two lexical noun phrases (NPs) involved in the relative clause:

These conditions are inspired by findings from psycholinguistic literature, which reveal that the processing difficulty humans encounter with ORCs - particularly in online comprehension - can be reduced when there is a mismatch between the two NPs in certain grammatical and semantic features [24, 10, 25, 26, 27]. Specifically, we focus on three key features that were shown to have this effect: **gender, number**, and **animacy**. To ensure a balanced dataset, we consulted existing resources and literature that have carefully controlled for these conditions.

For gender and number, we utilized the Italian experimental stimuli set described by [24], focusing exclusively on the center-embedded ORCs portion. This dataset, referred to as `Biondo-et-al-2023`, contains 306 ORCs equally divided into three subsets:

- The first subset (*gen-num-match* condition) contains ORCs where both NPs match in gender and number (i.e., both singular and masculine);
- The second subset (*gen-mismatch* condition) introduces a gender mismatch, where NP2 remains singular but is feminine;
- The third subset (*num-mismatch* condition) introduces a number mismatch, where NP2 is masculine but plural.

For animacy, we incorporated 56 examples drawn from a larger set of experimental stimuli described in the paper by Gennari and McDonald, 2008 [25]. These sentences were originally in English and were translated into Italian, ensuring that the object relative clause construction

PAIR	SENTENCE1	SENTENCE2	NP target	GOLD
1	Il professore che lo studente chiama apre la porta dell'aula.	Il professore sta aprendo una porta.	NP1	YES
2	Il pittore che il fotografo coinvolge inaugura una mostra d'avanguardia.	Il pittore è stato coinvolto dal fotografo.	NP1	YES
3	L'attore che il ballerino ringrazia rompe il microfono nuovo.	L'attore sta ringraziando il ballerino.	NP1	NO
4	L'infermiere che il dottore critica aggiorna i turni della settimana.	L'infermiere non ha aggiornato i turni settimanali.	NP1	NO
5	L'allenatore che il nuotatore accusa commette un'infrazione del regolamento.	Il nuotatore sta accusando l'allenatore.	NP2	YES
6	Il cuoco che il cameriere consulta introduce un menù per vegetariani.	Il cameriere è stato consultato dal cuoco.	NP2	NO
7	Il nonno che il bambino insegue calpesta un sasso appuntito.	Il bambino ha calpestato un sasso.	NP2	NO
8	Il pagliaccio che la ragazza deride attira l'attenzione di tutti.	La ragazza sta prendendo in giro il pagliaccio.	NP2	YES

Table 1
Extract of the dataset with the main criteria for yes/no entailment exemplified.

remained syntactically correct and semantically natural in the target language. All of these sentences exhibit an animacy mismatch: in half of the examples, NP1 is animate and NP2 is inanimate, while in the other half, the reverse configuration is applied.

Additionally, we introduced a fourth condition, also inspired by psycholinguistic research, which focuses on manipulating the **distance** between the two NPs. This manipulation aims to increase sentence complexity due to a longer subject-verb agreement dependency in the main clause[4, 28], which might result in agreement attraction effects [29, 30]. This condition was obtained by adding one or more prepositional phrases (PP) to either NP1 or NP2, thereby extending the distance between the noun phrases and increasing the subject-verb agreement dependency in the main clause. This fourth condition was applied to 156 sentences, which were sourced from the two aforementioned datasets. Specifically, 100 sentences were selected from the `Biondo-et-al-2023` dataset, distributed evenly across the three subsets (match, gender mismatch, and number mismatch), and the entire set from [25] was used.

Finally, we included a small set of **'mix-category'** ORCs, with sentences sourced from 'sister challenge' benchmarks such as CoLA [17], ITaCoLA [19], and ACCOMPL-it [20], specifically selecting only those marked as grammatical in the original datasets. While these sentences all contain ORC constructions, the two NPs were not controlled for specific features. Furthermore, except for the CoLA sentences[2], these examples feature right-branching rather than center-embedded structures. Given the novel formulation of our task (to our knowledge), it will be interesting to determine whether these models have acquired the ability to reason about complex constructions they might have already encountered and been tested on, beyond simply recognizing their grammaticality.

Table 2 summarizes the types of ORCs included in the dataset, along with an example for each condition.

3.1. Human Evaluation

Since the assignment of gold labels to sentence pairs in the benchmark was manually derived, though primarily informed by linguistic literature, we conducted a human evaluation with untrained native speakers to validate the examples and ensure they conveyed clear implications.

For this validation, we selected 240 sentence pairs, representing approximately 42% of the entire benchmark, with an equal distribution across all conditions. These pairs were annotated by Italian native speakers, recruited via the Prolific platform[3]. The annotation process was organized into eight questionnaires, each containing 30 sentence pairs. Each pair was labeled by five different workers, resulting in a total of 1,050 human judgments.

To maintain accuracy and reliability, each questionnaire included five control items where the first sentence was a simple declarative. Annotators were given very simple instructions, similar to the prompt used for the LLM, and were asked to carefully evaluate each pair and determine whether the first sentence implied the second.

The final label for each pair was determined through majority voting. This process yielded an accuracy rate of 94.2% (226 correct; 14 incorrect). Of the 226 correctly annotated pairs, 207 achieved agreement from at least

[2]Sentences included in TRACE-it were translated into Italian.

[3]https://www.prolific.com/

COND	FEAT	EXAMPLE	#	SOURCE
gen-num	all-match	Il professore che lo studente chiama apre la porta dell'aula.	102	[24]
	gen-mism	Il professore che la studentessa chiama apre la porta dell'aula.	102	
	num-mism	Il professore che gli studenti chiamano apre la porta dell'aula.	102	
animacy	mism [an-in]	Lo scienziato che il libro ha infastidito era rinomato per i suoi saggi sull'ecologia.	28	[25]
	mism [in-an]	Il libro che lo scienziato ha studiato era rinomato per i suoi argomenti sull'ecologia.	28	
distance	all-match_NP1+PP	Il professore di storia e filosofia di Marco che lo studente chiama apre la porta dell'aula.	50	[24]_m
	gen-mism_NP2+PP	Il primario che la specializzanda di oculistica rassicura lascia il reparto incustodito	50	
	anim-mism_NP1+PP	Lo scienziato dell'agenzia pubblica europea che il libro ha infastidito era rinomato per i suoi saggi sull'ecologia.	28	[25]_m
	anim-mism_NP2+PP	Il libro che lo scienziato dell'agenzia pubblica europea ha studiato era rinomato per i suoi argomenti sull'ecologia.	28	
sister-ch	mixed	Il cane che la macchina ferì aveva un collare giallo.	17	[17]
		Ho bevuto il vino che Tommaso mi ha portato.	10	[19]
		Carlo conosceva bene il compagno di classe che Anna voleva sempre incontrare.	21	[20]

Table 2
Types of ORCs included in the dataset, categorized into the four main conditions based on the type of manipulation applied and the number of examples for each. The suffix "_m" in the last column indicates that modifications have been made to the original stimuli described in the reference source.

four annotators, while the remaining 19 were decided by a majority vote of three out of five annotators.

3.2. Data format

The benchmark is provided as a tab-separated text file with the following information for each entry:

- UniqueID: a numerical identifier for the entry;
- Source: the original reference from which the sentence has been taken;
- ID-mapping: an identifier mapping for cross-referencing according to the condition;
- Condition: The type of ORC, based on the features (i.e. gender, number, animacy, distance, mixed) and specific configurations (match, mismatch) of the two NPs involved;
- Sentence1: the first sentence containing the ORC;
- Sentence2: the second sentence that may or may not be implied by sentence 1;
- NP target: indicates whether Sentence 2 targets the head of the relative clause (NP1) or the subject of the embedded clause (NP2) in sentence1.;
- Gold: the gold label assigned to the pair ("sì" if sentence 1 implied sentence 2, "no" otherwise).

4. Evaluation

4.1. Zero-shot Prompting

To evaluate knowledge that emerges from the model's training rather than through in-context learning, we chose to adopt a zero-shot evaluation paradigm.

We formulate a very simple prompt, which is nearly identical to the instruction presented to humans in the annotation task:

> "Data questa coppia di frasi, valuta se la prima frase implica la seconda. Rispondi sì o no."

Although we experimented with various prompt formulations, we ultimately decided to avoid any prompts that encouraged the model to explicitly analyze the linguistic structure of the sentence. Our aim was to evaluate the model's raw ability to infer entailment without any task-specific guidance.

Metrics Given the perfectly balanced data distribution across the two classes, the evaluation metrics will be based on the **Accuracy** and **F1_score**.

4.2. Preliminary Results

We conducted an initial evaluation of the TRACE-it challenge on `llama-3-8B Instruct` [31], achieving an accuracy of 0.71.

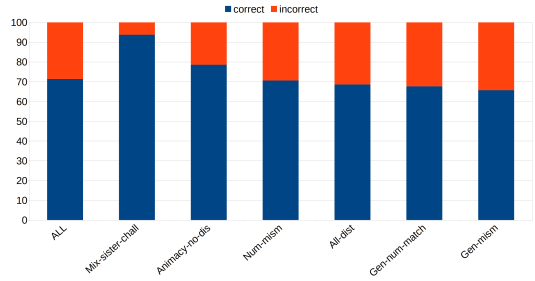

Figure 1: Percentage accuracy for the whole dataset (ALL) and across subsections.

Figure 1 reports accuracy results across the distinct subsections of the dataset. This preliminary analysis reveals that ORCs sourced from existing acceptability datasets were the easiest for the model to handle. In terms of ORCs with specific conditions applied to the two NPs, the model performed best on sentences where there was a mismatch in animacy, indicating that this condition is easier for the model to process. Conversely, when both NPs matched in animacy, the influence of grammatical features such as gender and number became more apparent. Specifically, a mismatch in number appeared to facilitate comprehension more effectively than either a full match or a gender mismatch, a finding that aligns with human data [24].

However, these observations are based on preliminary analysis and require further validation. Generalization capabilities should be verified across different models to obtain more robust conclusions.

5. Conclusion

In this report, we have described TRACE-it, a novel benchmark, with a corresponding task, presented for the CALAMITA challenge and designed to evaluate the ability of large language models (LLMs) to comprehend object relative clauses (ORCs) in Italian. By focusing on this specific type of complex syntactic construction, TRACE-it allows for a detailed examination of how models handle key grammatical and semantic features, such as gender, number, and animacy, which are known to influence human comprehension.

The results from our preliminary evaluation showed that while models are able to grasp ORC comprehension, challenges remain, and they are consistent with patterns observed in human language processing studies. Although the benchmark is small in scale and limited to a single syntactic structure, it serves as a crucial first step towards a deeper understanding of LLMs' syntactic capabilities in Italian. Future work should aim to expand both the dataset and the range of syntactic phenomena to create a more comprehensive evaluation framework.

6. Limitations

There are several limitations in the current benchmark. First, the dataset is small in scale and focuses exclusively on a single syntactic construction — object relative clauses. While this targeted approach enables a focused investigation into how language models process specific grammatical features, it restricts the generalizability of the results to other complex syntactic phenomena. Expanding the dataset to include a broader range of syntactic structures and increasing its size would provide a more comprehensive evaluation of language models' syntactic comprehension abilities.

Additionally, the binary-choice format required by the entailment task presents another limitation. By forcing models (and humans) to make a yes/no decision, this approach simplifies the evaluation and may not fully capture the complexity of syntactic understanding. Future work could explore alternative evaluation formats that allow for a more graded or probabilistic assessment of model performance.

References

[1] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[2] J. A. Hawkins, Processing complexity and filler-gap dependencies across grammars, Language 75 (1999) 244–285. URL: https://api.semanticscholar.org/CorpusID:89607408.

[3] L. Frazier, C. Clifton, Successive cyclicity in the grammar and the parser, Language and Cognitive Processes 4 (1989) 93–126. URL: https://api.semanticscholar.org/CorpusID:62152168.

[4] E. Gibson, Linguistic complexity: Locality of syntactic dependencies, Cognition 68 (1998) 1–76.

[5] L. A. Stowe, Parsing wh-constructions: Evidence for on-line gap location, Language and Cognitive Processes 1 (1986) 227–245. URL: https://api.semanticscholar.org/CorpusID:62596346.

[6] J. P. King, M. A. Just, Individual differences in syntactic processing: The role of working memory, Journal of Memory and Language 30 (1991) 580–602. URL: https://api.semanticscholar.org/CorpusID:144231849.

[7] A. Staub, Eye movements and processing difficulty in object relative clauses, Cognition 116 (2010) 71–86.

[8] M. De Vincenzi, Syntactic parsing strategies in Italian: The minimal chain principle, volume 12, Springer Science & Business Media, 1991.

[9] L. M. S. Corrêa, An alternative assessment of children's comprehension of relative clauses, Journal of psycholinguistic research 24 (1995) 183–203.

[10] N. Friedmann, A. Belletti, L. Rizzi, Relativized relatives: Types of intervention in the acquisition of a-bar dependencies, Lingua 119 (2009) 67–88.

[11] H. Diessel, M. Tomasello, A new look at the acquisition of relative clauses, Language (2005) 882–906.

[12] K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, M. Baroni, Colorless green recurrent networks dream hierarchically, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1195–1205. URL: https://aclanthology.org/N18-1108. doi:10.18653/v1/N18-1108.

[13] R. Marvin, T. Linzen, Targeted syntactic evaluation of language models, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1192–1202. URL: https://aclanthology.org/D18-1151. doi:10.18653/v1/D18-1151.

[14] S. A. Chowdhury, R. Zamparelli, Rnn simulations of grammaticality judgments on long-distance dependencies, in: Proceedings of the 27th international conference on computational linguistics, 2018, pp. 133–144.

[15] E. G. Wilcox, R. Levy, T. Morita, R. Futrell, What do rnn language models learn about filler–gap dependencies?, in: BlackboxNLP@EMNLP, 2018. URL: https://api.semanticscholar.org/CorpusID:52156878.

[16] J. Gauthier, J. Hu, E. Wilcox, P. Qian, R. Levy, SyntaxGym: An online platform for targeted evaluation of language models, in: A. Celikyilmaz, T.-H. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 70–76. URL: https://aclanthology.org/2020.acl-demos.10. doi:10.18653/v1/2020.acl-demos.10.

[17] A. Warstadt, A. Singh, S. R. Bowman, Neural network acceptability judgments, Transactions of the Association for Computational Linguistics 7 (2019) 625–641. URL: https://aclanthology.org/Q19-1040. doi:10.1162/tacl_a_00290.

[18] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, S. R. Bowman, Blimp: The benchmark of linguistic minimal pairs for english, Transactions of the Association for Computational Linguistics 8 (2020) 377–392.

[19] D. Trotta, R. Guarasci, E. Leonardelli, S. Tonelli, Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2929–2940. URL: https://aclanthology.org/2021.findings-emnlp.250. doi:10.18653/v1/2021.findings-emnlp.250.

[20] D. Brunato, C. Chesi, F. Dell'Orletta, S. Montemagni, G. Venturi, R. Zamparelli, Accompl-it @ evalita2020: Overview of the acceptability & complexity evaluation task for italian, EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020 (2020). URL: https://api.semanticscholar.org/CorpusID:229292651.

[21] J. Opitz, S. Wein, N. Schneider, Natural language processing relies on linguistics, arXiv preprint arXiv:2405.05966 (2024).

[22] A. Warstadt, S. R. Bowman, What artificial neural networks can tell us about human language acquisition, in: Algebraic structures in natural language, CRC Press, 2022, pp. 17–60.

[23] Y. Belinkov, J. Glass, Analysis Methods in Neural Language Processing: A Survey, Transactions of the Association for Computational Linguistics 7 (2019) 49–72. URL: https://doi.org/10.1162/tacl_a_00254. doi:10.1162/tacl_a_00254.

[24] N. Biondo, E. Pagliarini, V. Moscati, L. Rizzi, A. Belletti, Features matter: the role of number and gender features during the online processing of subject- and object- relative clauses in italian, Language, Cognition and Neuroscience 38 (2023) 802–820. URL: https://doi.org/10.1080/23273798.2022.2159989. doi:10.1080/23273798.2022.2159989.

[25] S. P. Gennari, M. C. MacDonald, Semantic indeterminacy in object relative clauses, Journal of memory and language 58 (2008) 161–187.

[26] M. W. Lowder, P. C. Gordon, Effects of animacy and noun-phrase relatedness on the processing of complex sentences, Memory & cognition 42 (2014) 794–805.

[27] W. M. Mak, W. Vonk, H. Schriefers, The influence of animacy on relative clause processing, Journal of Memory and Language 47 (2002) 50–68. URL: https://www.sciencedirect.com/science/article/pii/S0749596X01928372. doi:https://doi.org/10.1006/jmla.2001.2837.

[28] H. Liu, C. Xu, J. Liang, Dependency distance: A new perspective on syntactic patterns in natural languages, Physics of life reviews 21 (2017) 171–193.
[29] J. Franck, G. Lassi, U. H. Frauenfelder, L. Rizzi, Agreement and movement: A syntactic analysis of attraction, Cognition 101 (2006) 173–216.
[30] D. Parker, A. An, Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects, Frontiers in psychology 9 (2018) 1566.
[31] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).