BLM-It — Blackbird Language Matrices for Italian: A CALAMITA Challenge

Chunyang Jiang[1,2,*], Giuseppe Samo[1], Vivi Nastase[1] and Paola Merlo[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Geneva, Switzerland

Abstract

In this challenge, we propose Blackbird Language Matrices (BLMs), linguistic puzzles to learn language-related problems and investigate deeper formal and semantic properties of language, through a process of paradigm understanding. A BLM matrix consists of a context set and an answer set. The context is a sequence of sentences that encode implicitly an underlying generative linguistic rule. The contrastive multiple-choice answer set includes negative examples produced following corrupted generating rules. We propose three subtasks —agreement concord (*Agr*), causative (*Caus*) and object-drop (*Od*) alternation detection— each in two variants of increasing lexical complexity. The datasets comprise a few prompts for few-shot learning and a large test set.

Keywords

Blackbird Language Matrices, Causative/inchoative alternation, Object-drop alternation, subject-verb number agreement, rule-based abstraction, disentanglement

1. Introduction and Motivation

Current generative large language models (LLMs) translate across close languages, produce fluent and informative summaries, and answer questions promptly. And yet, they still fail in very non-human ways. As proven by their prohibitive needs in size of training data and expensive computational resources, large language models do not generalise nor abstract systematically. Humans, instead, are good at abstraction and generalisation.

To reach systematic abilities in abstraction and generalisation in neural networks, we need to develop tasks and data that help us understand their current generalisation abilities —what exactly do LLMs understand of the language they produce and process so well?— and help us train them to more complex skills.

In the CALAMITA challenge[1], we propose to find the solution to Blackbird Language Matrices (BLMs), linguistic puzzles developed in analogy to the visual Raven Progressive Matrices tests [2]. Raven's Progressive Matrices (RPMs) consist of a sequence of images, called the *context*, connected in a logical sequence by underlying generative rules [3]. The task is to determine the missing element in this visual sequence, the *answer*, chosen among a set of closely or loosely similar alternatives, as illustrated in Figure 1.

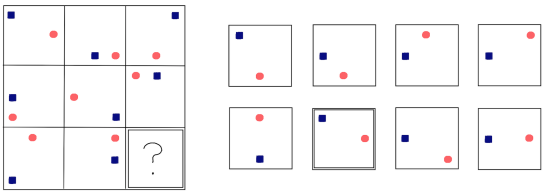


Figure 1: Example of a Raven's Progressive Matrix (RPM) from visual intelligence tests. This instance is generated with two generative rules: (i) the red dot moves one place clockwise when traversing the matrix left to right; (ii) the blue square moves one place anticlockwise when traversing the matrix top to bottom. The task consists in finding the tile in the answer set that correctly completes the sequence (indicated with a double border).

Unlike other attempts to create textual versions of RPMs, BLMs are not simplistic transcriptions of visual stimuli [4]—a technique that, in practice, might give away parts of the solution to the problem—, nor are they auxiliary abstractions of stimuli in the visual domain [5]. Instead, BLMs are matrices developed specifically to learn language-related problems and delve into deeper formal and semantic properties of language, through a process of linguistic paradigm understanding.

Like RPMs, a BLM instance consists of a context set and an answer set. The context is a sequence of sentences that encode a linguistic rule. They encode, for example, the rule of grammatical number concord: subject and verb agree in their grammatical number, and they do so independently of how many noun phrases intervene between them. BLMs are presented as linguistic puzzles requiring the selection of the missing sentence. In order

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

*Corresponding author.
chunyang.jiang@unige.ch (C. Jiang); giuseppe.samo@idiap.ch (G. Samo); vivi.a.nastase@gmail.com (V. Nastase); Paola.Merlo@unige.ch (P. Merlo)
https://www.idiap.ch/en/scientific-research/researchers (P. Merlo)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3878/125_calamita_long.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

to examine the representations underlying the response, the answer sets include not only the correct answer, but also erroneous candidates constructed by corrupting the generating rules. An example template is illustrated in Figure 2.

BLM datasets are richly structured and support many different types of investigations, at both the sentence and matrix levels. The context-answer set up support counterfactual investigations of possible types of errors: language errors, reasoning errors, and their interactions [6, 7, 8]. The regular syntactic forms and the systematic semantic properties support investigations on systematicity and compositionality in neural networks. The predictable syntactic structure of individual sentences, and the structure within the sequence of a BLM context, also support investigations on sentence embeddings [9, 10]. BLMs exists for several tasks and different languages, enabling multi-tasks and multi-language comparative studies [11, 12]. Finally, each BLM problem is a linguistic paradigm and can be seen as a tool for linguistic investigation of specific phenomena.

Context			
NP-sg	PP1-sg		VP-sg
NP-pl	PP1-sg		VP-pl
NP-sg	PP1-pl		VP-sg
NP-pl	PP1-pl		VP-pl
NP-sg	PP1-sg	PP2-sg	VP-sg
NP-pl	PP1-sg	PP2-sg	VP-pl
NP-sg	PP1-pl	PP2-sg	VP-sg

Answer set				
NP-pl	PP1-pl	PP2-sg	VP-pl	Correct
NP-pl	PP1-pl	et PP2-sg	VP-pl	Coord
NP-pl	PP1-pl		VP-pl	WNA
NP-pl	PP1-sg	PP1-sg	VP-pl	WN1
NP-pl	PP1-pl	PP2-pl	VP-pl	WN2
NP-pl	PP1-pl	PP2-pl	VP-sg	AEV
NP-pl	PP1-sg	PP2-pl	VP-sg	AEN1
NP-pl	PP1-pl	PP2-sg	VP-sg	AEN2

Figure 2: BLM-AgrI template for verb-subject agreement, with one-two intervening phrases. Three generative rules: (i) Subject matches in number with verb (singular or plural); (ii) material can intervene and is of unbounded length; (iii) singular and plural alternate in regular patterns. NP=Noun Phrase, PP=Prepositional Phrase, VP=Verb Phrase. Answers: WNA= wrong number of attractors; WN1= wrong nr. for 1st attractor noun (N1); WN2= wrong nr. for 2nd attractor noun (N2); AEV=agreement error on the verb; AEN1=agreement error on N1; AEN2=agreement error on N2.

2. The BLM-It Challenge

The BLM-It challenge consists of six sub-tasks.[1] All subtasks are instances of the general BLM task, but they differ along two dimensions: the linguistic problem defined (*Agr*, *Caus*, *Od*) and the lexical complexity of the data (II, III).[2] While the agreement (*Agr*) task focuses on information about the formal grammatical property of agreement, the causative (*Caus*) and object-drop (*Od*) alternation tasks focus on lexical semantic properties of verbs, their ability to enter or not in a causative alternation and their systematic alternation in the syntactic-semantic mapping of grammatical functions and semantic roles.

BLM-AgrI The BLM problem for subject-verb agreement [6] consists of a context set of seven sentences that share the subject-verb agreement phenomenon, but differ in other aspects – e.g. number of intervening attractors between the subject and the verb, different grammatical numbers for these attractors, and different clause structures. The answer set comprises contrastive sentences that violate some of the generative rules. The BLM-AgrI Template can be seen in Figure 2.

BLM-CausI The BLM-CausI matrix represents the causative/inchoative alternation, where the object of the transitive verb bears the same semantic role (Patient) as the subject of the intransitive verb (*L'artista ha aperto la finestra/La finestra si è aperta* 'The artist opened the window'/'The window opened'). The transitive form of the verb has a causative meaning [13].

The BLM-CausI template is shown in Figure 4. The context set of the causative alternation varies depending on the presence of one or two arguments and their attributes (agents, Ag; patients, Pat) and the active (Akt) and passive (Pass) or passive voice of the verb. The sentences are organised in a structured sequence: an alternation every two items between a prepositional phrase introduced by multifarious prepositions (e.g., *in pochi secondi*, P-NP) and a PP introduced by the agentive da-NP (e.g., *dall'artista*, da-Ag/da-Pat).

The answer set is composed of one correct answer and contrastive erroneous answers, all formed by the same four elements: a verb, two nominal constituents and the presence (or absence) of a prepositional phrase.

BLM-OdI The BLM-OdI template is minimally different from BLM-CausI. They also act as each other's controls. In contrast to *Caus*, the subject in *Od* bears the same semantic role (Agent) in both the transitive and intransitive forms (*L'artista dipingeva la finestra/L'artista dipingeva* 'the artist painted the window'/'the artist

[2]We choose names of tasks and lexical complexity levels that make it easier to cross-reference and compare the data described here with other papers published on BLMs.

[2]Our datasets are available here:
https://www.idiap.ch/en/scientific-research/data/blm-agri-gen,
https://www.idiap.ch/en/scientific-research/data/blm-causi-gen,
https://www.idiap.ch/en/scientific-research/data/blm-odi-gen.

type II	
	Context
1	La zia mangia una bistecca nella sala grande
2	La presidente può mangiare una bistecca da programma
3	La specialità della casa deve essere mangiata dalla turista nella sala grande
4	Una bistecca fu mangiata dalla presidente da sola
5	La specialità della casa deve essere mangiata in un secondo
6	Una bistecca deve poter essere mangiata da sola
7	La turista deve mangiare con fame
8	???
	Answer set
1	La specialità della casa può mangiare da sola
2	**La squadra di calcio deve mangiare da mezz'ora**
3	Una bistecca è mangiata dalla turista
4	La squadra di calcio può essere mangiata da una carbonara
5	La pasta col pomodoro può mangiare la squadra di calcio
6	La squadra di calcio mangia una bistecca
7	La specialità della casa deve poter mangiare dalla turista
8	La presidente mangia da una bistecca

type III	
	Context
1	L'attore deve canticchiare un motivetto dopo il festival
2	L'amica di mia mamma deve cucire la tasca da qualche giorno
3	L'inno nazionale può essere cantato dal vincitore del festival con solo pianoforte
4	Una bistecca deve essere mangiata dalla turista da sola
5	Il manuale è insegnato nell'aula magna
6	Questi attrezzi devono essere intagliati da manuale
7	I due fratelli studiano con molta attenzione
8	???
	Answer set
1	La pasta frolla deve impastare da sola
2	**L'autrice deve poter scrivere da qualche giorno**
3	I libri di testo devono poter essere studiati dai candidati
4	Questi stilisti devono poter essere tessuti dai vestiti per la parata
5	Questi motivi greci possono tessere questi stilisti
6	L'idraulico saldò i cavi del lampadario
7	La stanza pulisce da una delle propretarie dell'albergo
8	Le sommozzatrici pescarono da delle trote

Figure 3: Two instances of BLM-OdI data: with little (type II) and maximal (type III) lexical variation.

Context				Answer set	
1	Ag	Akt	Pat	p-NP	1 Pat Akt by-NP **Correct**
2	Ag	Akt	Pat	by-NP	2 Ag Akt by-NP I-Int
3	Pat	Pass	by-Ag	p-NP	3 Pat Pass by-Ag ER-Pass
4	Pat	Pass	by-Ag	by-NP	4 Ag Pass by-Pat IER-Pass
5	Pat	Pass		p-NP	5 Pat Akt Ag R-Trans
6	Pat	Pass		by-NP	6 Ag Akt Pat R-Trans
7	Pat	Akt		p-NP	7 Pat Akt by-Ag E-WrBy
8	???				8 Ag Akt by-Pat IE-WrBy

Figure 4: BLM-CausI Template. Three generative rules: (i) the presence of either one or two arguments and their attributes (agents, Ag; patients, Pat); (ii) the active (Akt) and passive (Pass) voice of the verb; the number and quality of nominal phrases (NP) following the verb. Answers: I-Int=wrong subject semantic role; ER-Pass=wrong verb mood; IER-Pass=wrong mood and wrong subject semantic role; R-trans=wrong sequence reasoning (transitive sentence with the second NP not preceded by a preposition); IE-WrBy=ungrammatical sentence (NP following the preposition *da*).

Context				Answer set	
1	Ag	Akt	Pat	p-NP	1 Pat Akt by-NP I-Int
2	Ag	Akt	Pat	by-NP	2 Ag Akt by-NP **Correct**
3	Pat	Pass	by-Ag	p-NP	3 Pat Pass by-Ag IER-Pass
4	Pat	Pass	by-Ag	by-NP	4 Ag Pass by-Pat ER-Pass
5	Pat	Pass		p-NP	5 Pat Akt Ag IR-Trans
6	Pat	Pass		by-NP	6 Ag Akt Pat R-Trans
7	Ag	Akt		p-NP	7 Pat Akt by-Ag IE-WrBy
8	???				8 Ag Akt by-Pat E-WrBy

Figure 5: BLM-OdI Template. Same generative rules as BLM-CausI, with the difference that here the passive/active voice is confounding, and the correct answer is an erroneous answer for BLM-CausI.

painted') and the verb does not have a causative meaning [13].

The BLM template for *Od* is the same as for *Caus*, but here the passive voice serves as a confounding element and one of the contrastive answers for *Caus* is, in fact, the correct answer here.

The template for BLM-OdI is in Figure 5. Due to the asymmetry between the *Caus* and *Od* BLM templates, the contexts of the BLMs minimally differ in the intransitive followed by P-NP (sentence 7). The correct answer also varies across the two groups, although in both cases it is an intransitive form with a da-NP.

Lexical variants Each of the three BLM templates described above is developed in two lexical variants, with less (II) or more (III) lexical variation. In type II BLMs, only one word in each sentence changes for each matrix, compared to the other sentences, while in type III data, all words can change. Instances of the two variations are shown in Figure 3.

3. Data description

The data is generated by the process described in Figure 6: (i) start from identifying a linguistic phenomenon of interest, its forms of expression and factors influencing it within a context, (ii) produce a set of seed examples from

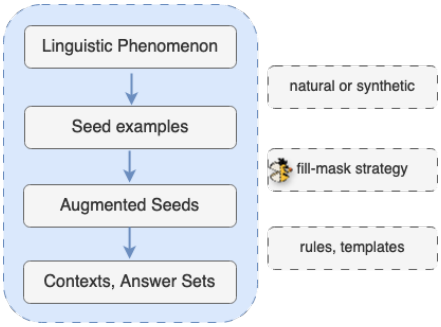


Figure 6: BLM data generation process, from seed examples of a linguistic problem to the complete dataset

natural or synthetic data, (iii) automatically augment the seeds using a fill-mask strategy, (iv) produce BLM instances following the designed templates and generative rules. Two instances of *Od* verb alternations are shown in Figure 3.

3.1. Origin of data

BLM-AgrI To instantiate the templates, our starting point are the examples in Franck et al. [14, appendix1]. They provide a set of subject NPs of various complexity – including prepositional phrases, themselves of various complexity. The sentences were produced based on these subject NPs by manually adding verb phrases, and by making the NPs more complex to increase the distance between the subject and the verb in the sentence [6]. Each of these sentences is used to produce a seed.

BLM-CausI and BLM-OdI Thirty verbs from each of the causative and object-drop classes in English in Levin [13] were selected and translated by a native speaker into Italian, where translations maintain the same alternation structure.

The seeds were augmented using masked modeling on BERT-BASE-UNCASED [15]. The Italian data are built as native-speaker translations of the English data, with manual corrections to guarantee the acceptability and semantic plausibility of the sentences, and assure variability in gender and number.

3.2. Data format

The structured BLM data is provided in a json file, each instance as one element with specific fields described in Figure 7. A data instance is shown in Figure 10 in the appendix.

dataset	(few-shot) train	test
BLM-AgrI (II/III)	10	2000
BLM-CausI (II/III)	80	2080
BLM-OdI (II/III)	80	2080

Table 1
Data statistics for the three datasets, in terms of few-shot training and testing. There are the same number of examples in the type II (small lexical variation within an instance) and type III (maximal lexical variation within an instance) variations of the three datasets.

3.3. Detailed data statistics

For the BLM-AgrI datasets, for each of types II and III, we randomly sample 10 instances for few-shot learning from a dataset of 2010 instances. The rest will be used for testing. For the BLM-CausI and BLM-OdI datasets, which are focused on specific verbs, we extract all instances for one verb (based on the correct answer in each instance) for few-shot training. From an initial dataset of 2160 instances for 27 verbs (80 instances per verb), we select the 80 instances for one verb for few-shot training, and the rest are left for testing.

3.4. Example of prompts

We design prompts in English and Italian in zero-shot and few-shot prediction settings, to test the impact of the language of the prompt on the task. These prompts test LLMs' ability to perform complex linguistic tasks with varying levels of context. Both types of prompts are structured to minimize ambiguity and focus on the core task of selecting the best sentence to follow the given context.

Zero-Shot Prompt Example in English The prompt in Figure 8 is designed to create a clear zero-shot baseline for challenging linguistic tasks. We avoid complex prompting techniques, like chain-of-thought or step-by-step reasoning [16, 17]. This ensures that the model's performance reflects its intrinsic capabilities for linguistic understanding and reasoning without prior in-context learning or guided reasoning steps.

We format the prompt in `Markdown` format and explicit label sections for Context and Answer Set. The task is framed as a simple "puzzle" with the instruction to "choose [...] the sentence that could [...] follow the context". This abstract formulation guides the model to focus on identifying the best sequential fit without introducing ambiguity. The prompt also aims to reduce noise and simplify the evaluation by fixing its output format.

Few-Shot (One-Shot) Prompt Example in Italian For the one-shot prediction setup (as is shown in Figure 9), we provide an example of the task in Italian before presenting the new instance to the model. The prompt serves to test the model's ability to use prior examples

```
{
    "ID": <ID NUMBER>,
    "Context": [<List of comma-separated, double-quoted sentences>],
    "Context_concatenated": <Double-quoted concatenation of context sentences,
        each prefixed by a numeral (1 to 7) followed by a tab, separated by newlines>,
    "Answer_set": [<List of comma-separated, double-quoted sentences>],
    "Answer_concatenated": <Double-quoted concatenation of answer sentences,
        each prefixed by a letter (A, B, C, ...) followed by a tab, separated by newlines>,
    "Correct_option": <Double-quoted single letter label>,
    "Correct_answer": <Double-quoted single correct answer sentence>,
    "Answer_set_annotation": [<List of comma-separated triplets
    {"label":<error-type>,"value":<truth value>,"option":<single letter label>}>],
    "Verb": <Double-quoted single verb>
},
```

Figure 7: Data format

```
# TASK: I'm asking you to solve a puzzle. The language of the puzzle is Italian.
I will give you a list of sentences (numbered from 1 to 7) called the **Context**, and a set of sentences (identified by capital letters) called the **Answer Set**.
Your task is to choose among the **Answer Set** the sentence that could be the next sentence following the **Context**.

# FORMAT:  You should **ONLY** output the letter corresponding to the best answer. Do not output other text before or after.


# QUESTION
**Context**
{{Context_concatenated}}

**Answer Set**
{{Answer_concatenated}}

**Your Choice**
```

Figure 8: Zero-Shot Prompt in English.

and adapt to a new linguistic context.

4. Metrics

We perform zero-shot and one-shot evaluation on BLM-AgrI, BLM-CausI and BLM-OdI tasks, using English and Italian prompts, with 100 samples each (batch size of one, evaluated instance by instance, over three independent runs) with `Meta-Llama-3-8B-Instruct` (ML-8), `Meta-Llama-3-70B-Instruct` (ML-70), `Mistral-7B-Instruct-v0.3` (M-7), and `Gemma-2-9b-It` (G-2). We report averaged F1 scores over 3 runs in Table 2.

```
# COMPITO: Ti chiedo di risolvere un quesito. La lingua di questo quesito e' l'italiano.
Ti daro' una lista di frasi (numerate da 1 a 7) che chiameremo **Contesto**, e un insieme di frasi (identificate da una lettera) che chiameremo **Risposte**.
Il tuo compito e' di scegliere fra le **Risposte** la frase che potrebbe essere la frase seguente del **Contesto**.

# FORMATO: Devi mettere **SOLO** la lettera che corrisponde alla risposta migliore. Non inserire altro testo, ne' prima ne' dopo.


# ESEMPIO 1
**Contesto**
{{Context_concatenated}}

**Risposte**
{{Answer_concatenated}}

**Scelta corretta**
{Correct_option}

# DOMANDA
**Contesto**
{{Context_concatenated}}

**Risposte**
{{Answer_concatenated}}

**La tua scelta**
```

Figure 9: Few (One)-Shot Prompt in Italian.

BLM-AgrI tasks `Meta-Llama-3-70B-Instruct` consistently outperforms the other models, particularly in zero-shot English prompts, while also competitive in

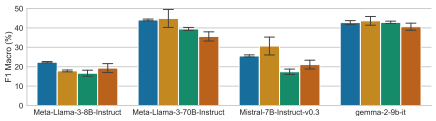
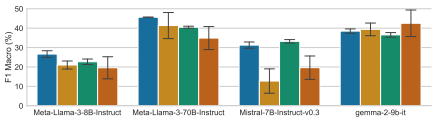
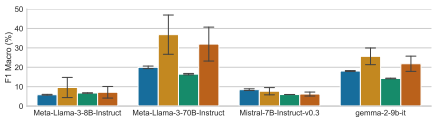
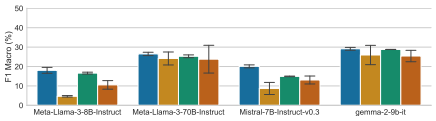
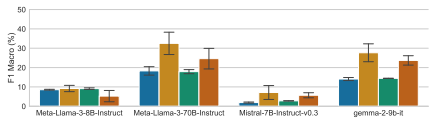
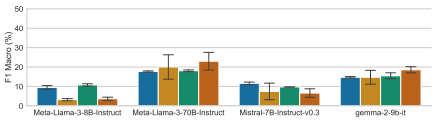
Model	English Prompt		Italian Prompt		Results
	Zero-Shot	One-Shot	Zero-Shot	One-Shot	
BLM-AgrI type II					
ML-70	**44.1 ± 0.46**	**44.88 ± 4.63**	39.46 ± 0.79	35.62 ± 2.36	
ML-8	22.34 ± 0.33	17.84 ± 0.48	16.66 ± 1.56	19.30 ± 2.30	
M-7	25.54 ± 0.58	30.66 ± 4.60	17.41 ± 1.37	21.1 ± 2.26	
G-2	*42.75 ± 1.01*	*43.64 ± 2.25*	**42.87 ± 0.62**	**40.62 ± 1.83**	
BLM-AgrI type III					
ML-70	**45.64 ± 0.05**	**41.35 ± 6.71**	**40.48 ± 0.52**	*34.89 ± 5.93*	
ML-8	26.65 ± 1.71	21.00 ± 2.07	22.68 ± 1.41	19.58 ± 5.68	
M-7	31.26 ± 1.60	12.75 ± 6.28	33.21 ± 0.91	19.64 ± 6.02	
G-2	38.48 ± 1.12	*39.36 ± 3.27*	36.54 ± 1.18	**42.52 ± 6.83**	
BLM-CausI type II					
ML-70	**19.97 ± 0.65**	**36.81 ± 10.11**	**16.46 ± 0.36**	**31.95 ± 8.75**	
ML-8	5.85 ± 0.20	9.57 ± 5.20	6.72 ± 0.09	7.12 ± 3.00	
M-7	8.45 ± 0.44	7.66 ± 1.87	5.94 ± 0.04	6.21 ± 1.02	
G-2	18.06 ± 0.25	25.64 ± 4.30	14.23 ± 0.16	*21.81 ± 3.93*	
BLM-CausI type III					
ML-70	26.49 ± 0.85	*24.14 ± 3.34*	25.27 ± 0.72	*23.78 ± 7.16*	
ML-8	18.03 ± 1.52	4.65 ± 0.38	16.59 ± 0.49	10.52 ± 2.21	
M-7	20.08 ± 0.76	8.69 ± 3.12	14.91 ± 0.15	13.05 ± 2.05	
G-2	**29.12 ± 0.73**	**25.93 ± 4.98**	**28.8 ± 0.04**	**25.41 ± 2.94**	
BLM-OdI type II					
ML-70	**18.28 ± 2.18**	**32.51 ± 5.77**	**17.89 ± 1.06**	**24.61 ± 5.31**	
ML-8	8.55 ± 0.21	9.18 ± 1.62	9.1 ± 0.41	5.25 ± 2.92	
M-7	1.92 ± 0.27	7.11 ± 3.59	2.79 ± 0.07	5.69 ± 1.31	
G-2	14.07 ± 0.78	*27.64 ± 4.63*	14.43 ± 0.08	*23.70 ± 2.42*	
BLM-OdI type III					
ML-70	**17.70 ± 0.32**	**20.05 ± 6.28**	**18.10 ± 0.44**	**23.01 ± 4.56**	
ML-8	9.50 ± 0.95	3.20 ± 0.57	10.78 ± 0.61	3.64 ± 0.85	
M-7	11.60 ± 0.64	7.45 ± 4.27	9.74 ± 0.01	6.6 ± 2.19	
G-2	14.74 ± 0.40	*14.75 ± 3.55*	15.49 ± 1.54	*18.58 ± 1.60*	

Table 2
Evaluation results on BLM-It tasks (AgrI, CausI, and OdI) using macro averaged F1 score (over 3 runs) and standard deviations (±std). Each run was evaluated with 100 samples, one instance at a time, for Meta-Llama-3-70B-Instruct (ML-70), Meta-Llama-3-8B-Instruct (ML-8), Mistral-7B-Instruct-v0.3 (M-7), Gemma-2-9b-It (G-2). Best performance is in bold, second best, if overlapping intervals, in italics.

one-shot settings. `Gemma-2-9b-it` shows robust performance, especially with Italian prompts, performing similarly to the larger Meta-Llama model. In contrast, smaller models, such as `Meta-Llama-3-8B-Instruct` and `Mistral-7B-Instruct-v0.3`, perform more weakly, especially with Italian prompts.

BLM-CausI tasks `Meta-Llama-3-70B-Instruct` leads across both English and Italian prompts, with improvement in one-shot English for type II. `Gemma-2-9b-it` shows comparable performance across both languages, in both zero-shot and one-shot settings. Smaller models perform worse for this task, especially in one-shot Italian prompts.

dataset	train:test	avg F1	
		E-M	E-It
BLM-AgrI type II	2400:4121	0.881 (0.003)	0.784 (0.007)
BLM-AgrI type III	2400:4121	0.874 (0.006)	0.336 (0.005)
BLM-CausI type II	2160:240	0.486 (0.005)	0.903 (0.010)
BLM-CausI type III	2160:240	0.475 (0.010)	0.918 (0.010)
BLM-OdI type II	2160:240	0.596 (0.010)	0.983 (0.003)
BLM-OdI type III	2160:240	0.592 (0.024)	0.994 (0.004)

Table 3
Dataset statistics and evaluation results on a two-level variational encoder-decoder architecture using an Italian Electra (E-It) and a multilingual Electra (E-M) pretrained model to provide sentence embeddings.

BLM-OdI tasks OdI tasks show the lowest overall performance across models. This indicates that the task is the most complex and challenging for the models. `Meta-Llama-3-70B-Instruct` performs best, particularly in one-shot English and Italian prompts. However, `Mistral-7B-Instruct-v0.3` struggles the most, particularly in zero-shot settings, which reflects that the model has limited generalisation capabilities in complex linguistic tasks.

Key Observations Larger models, such as `Meta-Llama-3-70B-Instruct` and `Gemma-2-9b-it`, consistently outperform smaller models, showing better generalisation and stability across tasks. English prompts generally result in higher F1 scores, though Italian prompts sometimes achieve comparable performance, particularly with `Gemma-2-9b-it`. One-shot prompting tends to improve performance, though the degree of improvement varies by model and task complexity. Smaller models, such as `Mistral-7B-Instruct` and `Meta-Llama-3-8B-Instruct`, show substantial variance, especially in one-shot scenarios, indicating instability in complex linguistic tasks.

Comparison with Multitask Learning Approaches We compare our LLM prompting results with the work of [12, 11], which explored the properties of Italian sentence embeddings – the embeddings of the [CLS] token from a pretrained Electra model[18][3] – through the agreement and the causative and object-drop BLM datasets, using a two-level Variational Encoder-Decoder architecture. This system learns to compress the sentence embeddings into representations relevant for the specific BLM tasks. The dataset statistics, and results on the individual BLM tasks as averaged F1 score over three runs and different amounts of lexical variation are shown in Table 3.

[3]Italian Electra (E-It) pretrained model: dbmdz/electra-base-italian-xxl-cased-discriminator, multi-lingual Electra (E-M) model: google/electra-base-discriminator

While not directly comparable due to the different training process and the different test data, using pretrained transformer encoder architectures, like Electra, significantly outperform the zero and one-shot prompting baseline. The performance gap suggests that while zero or one-shot prompting is flexible, it may not capture the complex syntactic and semantic features required for the BLM task in Italian.

5. Limitations

While the data is very rich and richly structured, it shares all the limitations of artificial and synthetic data: stilted sentence structure, limited variability, possibly sentences that are too short. This artificiality, though, might reduce, without eliminating, the risk of having sentences that were directly seen in the training data of the pretrained models that will be used, and that we use, for further experiments.

The initial seed sentences, although minimal, were crafted by experts. This approach is deliberate, like in the ARC dataset, to guarantee that the data are not algorithmically reproducible [19]. This expert-based approach, though, might not be easily scalable, especially given the complexity of the data. Exploring methods to leverage existing datasets for seed generation could mitigate this dependency.

The current dataset comprises three main tasks. More tasks and variants are needed to demonstrate the robustness and the wider appeal of the data.

6. Ethical issues

The data presented include an augmentation step that uses large language models (LLMs). LLMs are trained on extensive text data, which may unintentionally incorporate biases present in the training corpus.

7. Data license and copyright issues

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). For uses outside of these terms, please contact the authors.

Acknowledgments

We gratefully acknowledge the support of this work by the Swiss National Science Foundation, through grant SNF Advanced grant TMAG-1_209426 to PM.

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: Overview, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), 2024.

[2] P. Merlo, Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Motivations and formal specifications, ArXiv cs.CL 2306.11444 (2023). URL: https://doi.org/10.48550/arXiv.2306.11444. doi:10.48550/arXiv.2306.11444.

[3] J. C. Raven, Standardization of progressive matrices, British Journal of Medical Psychology 19 (1938) 137–150.

[4] T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7 (2023) 1526–1541. URL: https://doi.org/10.1038/s41562-023-01659-w. doi:10.1038/s41562-023-01659-w.

[5] X. Hu, S. Storks, R. Lewis, J. Chai, In-context analogical reasoning with pre-trained language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1953–1969. URL: https://aclanthology.org/2023.acl-long.109.

[6] A. An, C. Jiang, M. A. Rodriguez, V. Nastase, P. Merlo, BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1363–1374. URL: https://aclanthology.org/2023.eacl-main.99.

[7] V. Nastase, P. Merlo, Grammatical information in BERT sentence embeddings as two-dimensional arrays, in: Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), Toronto, Canada, 2023.

[8] G. Samo, V. Nastase, C. Jiang, P. Merlo, BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023.

[9] V. Nastase, P. Merlo, Are there identifiable structural parts in the sentence embedding whole?, 2024. URL: https://aclanthology.org/2024.blackboxnlp-1.3. doi:10.18653/v1/2024.blackboxnlp-1.3.

[10] V. Nastase, P. Merlo, Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification, in: Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), Bangkok, Thailand, 2024, pp. 203–214. URL: https://aclanthology.org/2024.repl4nlp-1.15.

[11] V. Nastase, G. Samo, C. Jiang, P. Merlo, Exploring Italian sentence embeddings properties through multi-tasking, in: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-It 2024), Pisa, Italy, 2024.

[12] V. Nastase, C. Jiang, G. Samo, P. Merlo, Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement, in: Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-It 2024), Pisa, Italy, 2024.

[13] B. Levin, English verb classes and alternations: A preliminary investigation, University of Chicago Press, 1993.

[14] J. Franck, G. Vigliocco, J. Nicol, Subject-verb agreement errors in french and english: The role of syntactic hierarchy, Language and cognitive processes 17 (2002) 371–404.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[18] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre- training text encoders as discriminators rather than generators, in: ICLR, 2020, pp. 1–18.

[19] F. Chollet, On the measure of intelligence, 2019. URL: https://arxiv.org/abs/1911.01547. arXiv:1911.01547.

A. Example Data Format


```
[{
    "ID": 215,
    "Context": [
        "le pittrici possono disegnare delle forme in meno di due giorni",
        "le artiste possono disegnare delle rappresentazioni artistiche da un mese",
        "alcune coreografie sono disegnate dalle pittrici nel salone espositivo",
        "delle rappresentazioni artistiche devono poter essere disegnate da queste studentesse da un mese",
        "alcune coreografie devono essere disegnate con pochi mezzi economici",
        "le scenografie devono essere disegnate da pochi mesi",
        "le pittrici devono disegnare nel salone espositivo"],
    "Context_concatenated": "1\tle pittrici possono disegnare delle forme in meno di due giorni\n2\tle artiste possono
    disegnare delle rappresentazioni artistiche da un mese\n3\talcune coreografie sono disegnate dalle pittrici nel
    salone espositivo\n4\tdelle rappresentazioni artistiche devono poter essere disegnate da queste studentesse da
    un mese\n5\talcune coreografie devono essere disegnate con pochi mezzi economici\n6\tle scenografie devono essere
    disegnate da pochi mesi\n7\tle pittrici devono disegnare nel salone espositivo",
    "Answer_set": [
        "delle rappresentazioni artistiche devono poter disegnare le sue allieve",
        "le scenografie devono essere disegnate dalle sue allieve",
        "le sue allieve devono essere disegnate da delle rappresentazioni artistiche",
        "le pittrici possono disegnare le scenografie",
        "le pittrici possono disegnare da un anno circa",
        "delle forme devono poter disegnare da pochi mesi",
        "le artiste devono poter disegnare da alcune coreografie",
        "delle rappresentazioni artistiche devono disegnare dalle artiste"],
    "Answer_concatenated": "A\tdelle rappresentazioni artistiche devono poter disegnare le sue allieve\nB\tle scenografie
    devono essere disegnate dalle sue allieve\nC\tle sue allieve devono essere disegnate da delle rappresentazioni
    artistiche\nD\tle pittrici possono disegnare le scenografie\nE\tle pittrici possono disegnare da un anno circa\nF\tdelle
    forme devono poter disegnare da pochi mesi\nG\tle artiste devono poter disegnare da alcune coreografie\nE\tdelle
    rappresentazioni artistiche devono disegnare dalle artiste",
    "Correct_option": "E",
    "Correct_answer": "le pittrici possono disegnare da un anno circa",
    "Answer_set_annotation": [
        {   "label": "IR-trans",
            "value": false,
            "option": "A" },
        {   "label": "IER-pass",
            "value": false,
            "option": "B" },
        {   "label": "ER-pass",
            "value": false,
            "option": "C" },
        {   "label": "R-trans",
            "value": false,
            "option": "D" },
        {   "label": "Correct",
            "value": true,
            "option": "E" },
        {   "label": "I-Int",
            "value": false,
            "option": "F" },
        {   "label": "E-WrBy",
            "value": false,
            "option": "G" },
        {   "label": "IE-WrBy",
            "value": false,
            "option": "H" }
    ],
    "Verb": "disegnare"
},
....
]
```


Figure 10: Sample entry formatted for usage with the provided prompts.