Large Language Models Assessment through Linguistically Motivated Contrasts: a benchmark for Italian (BLiMP-IT)

Veronica Bressan (IUSS Pavia, Ca' Foscari University of Venice), Matilde Barbini (IUSS Pavia), Achille Fusco (IUSS Pavia, University of Florence), Sofia Neri (IUSS Pavia), Maria Letizia Piccini Bianchessi (IUSS Pavia), Sarah Rossi (IUSS Pavia), Cristiano Chesi (IUSS Pavia)

Background. Large Language Models (LLMs) are advanced natural language processing systems trained to comprehend and generate human language. Their impressive and versatile performance has led some to suggest that they express genuine theories of language (Piantadosi 2023). A lively, ongoing debate centers on whether LLMs are able to draw linguistically relevant abstractions (Wilcox et al., 2024), or whether they are only fueled by spurious statistical generalizations (Bender et al. 2021). Crucially, these systems are largely opaque in the way they process and represent language patterns, ultimately raising questions about their interpretability and utility in theoretical terms. First, LLMs tend to conflate world knowledge with morphosyntactic competence, with factual recall not necessarily supported by linguistic generalizations (Bender and Koller 2020). Second, their relatively good performance with complex grammatical configurations, interpreted as evidence against the Poverty of Stimulus hypothesis (Piantadosi 2023), leans on dramatically oversized training data, compared to what children are exposed to (Katzir 2023). Third, higher performance of LLMs in increasingly specific tasks is not always matched by genuine gains in linguistic understanding (Ethayarajh & Jurafsky, 2020), suggesting that current performance metrics alone may not adequately capture linguistic competence (Chomsky, 1965).

The issue. Within this context, a fundamental line of research consists in developing linguistically informed benchmarks to evaluate model performance and the nature of their competence, offering a perspective that standard performance metrics often obscure (Coda-Forno et al., 2024). Recent shared tasks brought to the attention of the scientific community the effects of a small-sized training diet (10-100M tokens) on these models: relatively good results are achieved on various linguistic benchmarks (GLUE, Wang et al., 2019a, including BLiMP, and CoLA, Warstadt et al., 2020; Warstadt et al., 2019). However, most performant architectures (e.g., optimized transformers, Charpentier & Samuel, 2023) that show significant improvement with additional training sweeps, also yield diminishing returns in psycholinguistic terms (worse correlations with reading time, Steuer et al., 2023). In this work we capitalize on the results of the last BabyLM Challenge in English (Chesi et al., 2024) and a similar task on Italian (Fusco et al., 2024) to stress the importance of the linguistic benchmark adopted to truly challenge the Poverty of Stimulus hypothesis in a consistent way.

Our contribution. We developed a linguistically-motivated benchmark for Italian (BLiMP-IT) inspired by the English Benchmark for (Warstadt *et al.*, 2020), comprising a varied set of 78 phenomena, in turn collected into four broad linguistic macrophenomena: i. Agreement and Inflection (e.g.: noun-determiner agreement, subject-verb agreement, past participle agreement); ii. Verb class and argument structure (e.g.: θ-role assignment, auxiliary selection); iii. Pronouns (e.g.: clitics, reflexive pronouns); iv. Non-local dependencies (e.g.: island constraints, A'-dependencies). The main sources for phenomena selection and adaptation were not only previous benchmarks (Warstadt et al., 2020; Brunato et al., 2020; Trotta et al., 2021), but also standardized psycholinguistic tests assessing child perception of grammaticality across varying degrees of complexity (Chesi et al., 2024), in the aim of allowing for a detailed assessment of LLMs competence and its adherence to actual human linguistic generalizations. Crucially, BLiMP-IT evaluates LLMs performance using a forced-choice task based on minimal pair contrasts. This warrants solid generalizations (Sprouse & Almeida, 2017) and avoids discussing the intricate relationship between probability (the gradual, word-by-word output provided by each LLM) and grammaticality (a binary judgment), (Lau *et al.*, 2017). Each minimal pair expresses a precise phenomenon and it includes a grammatical sentence and an ungrammatical counterpart, obtained by violating one single structural aspect crucial in the expression of that phenomenon. One paradigmatic difference between English and Italian pairs is illustrated below ("wh- object gap" in original BLiMP):

(1) Teresa knew that man that April remembered. Vs
*Teresa knew who April remembered that man.

(2) a. Teresa conosceva chi Luisa dice di aver riconosciuto. Vs
b. *Teresa conosceva chi Luisa dice di averlo riconosciuto
T. knew who L. said to have-it.CL recognized
b'. *Teresa conosceva chi Luisa dice di aver riconosciuto quest'uomo
T. knew who L. said to have recognized that man

In BLIMP, the contrast in (1) compares wh-extraction with a complement clause (not exactly a structural minimal pairs), while in Italian we compared a long distance (to use the pre-verbal subject which is otherwise infelicitous in short wh-extractions, Rizzi, 1997) wh extraction, with exactly the same sentence in which the gap is filled either with a (resumptive) clitic pronoun, (2).a vs (2).b, or with a full DP (demonstrative D), (2).a vs (2).b'. Other contrasts, for instance assessing Agreement, adopts minimal variations of either number or gender just for the relevant agreeing items, controlling any other DP in the sentence. Presence absence of structural, or simply linear interveners are also considered (Franck *et al.*, 2006). Structural templates of each phenomenon were created and used to generate 100 structurally irrelevant lexical variations of the minimal pairs in a semi-automatic way using the lexicon extracted from the Italian 3M child-directed corpus (Fusco *et al.*, 2024).

Discussion. Preliminary results on the use of BLiMP-IT to evaluate models that simulate the qualitative and quantitative limitations faced during child linguistic development confirm that good performance on prototypical metrics (Loss/Accuracy in data training; model dimension) does not map into an equally good performance with minimal pair tasks, and miss child-like linguistic generalizations (Fusco et al., 2024). These results suggest that the Poverty of Stimulus Hypothesis cannot be convincingly refuted, and that the linguistically-motivated benchmarks may effectively address the issue of LMs cognitive plausibility by capitalizing on the intuitive sensitivity to deviant structures that defines (human) linguistic competence. Moreover, the underlying complexity metrics on BLiMP-IT allows for a strict control on linguistic coherence, highlighting discrepancies between human vs. LMs performance. In this spirit, the range of linguistic phenomena addressed by BLiMP-IT and the number of minimal pairs for each phenomenon is currently under expansion. A natural follow-up would also be to adapt this benchmark to other languages, minimally within the Romance family and possibly beyond it: the organization in minimal pairs isolating single linguistic properties and the semi-automatic generation process of these pairs makes it easy to operate on single points of parametric variation, and in principle to adapt the benchmark to any other language.

Selected references.

Bender, Gebru, McMillan-Major, Mitchell. 2021. doi: 10.1145/3442188.3445922. - Bender and Koller. 2020. doi: 10.18653/v1/2020.acl-main.463. - Brunato, Chesi, Dell'Orletta, Montemagni, Venturi, Zamparelli. 2020. In Proceedings of EVALITA 2020 - Charpentier and Samuel. 2023. doi: 10.18653/v1/2023.conll-babylm.20. - Chesi, Barbini, Bressan, Neri, Piccini Bianchessi, Rossi, Sgrizzi. 2024. In Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning. - Chesi, Ghersi, Musella, Musola. 2024. COnVERSA: Test Di Comprensione Delle Opposizioni Morfo-Sintattiche VERbali Attraverso La ScritturA. Firenze: Hogrefe. - Fusco, Barbini, Piccini Bianchessi, Bressan, Neri, Rossi, Sgrizzi, Chesi. 2024. In Proceedings of CLiC-It 2024 - Katzir. 2023. lingbuzz/007190. - Piantadosi. 2023. lingbuzz/007180. - Steuer, Mosbach, Klakow. 2023. doi: 10.18653/v1/2023.conll-babylm.12. - Trotta, Guarasci, Leonardelli, Tonelli. 2021. doi: 10.18653/v1/2021.findings-emnlp.250. - Warstadt, Parrish, Liu, Mohananey, Peng, Wang, Bowman. 2020. doi: 10.1162/tacl_a_00321. - Wilcox, Futrell, Levy. 2023. doi: 10.1162/ling_a_00491.