Can a Remedy Find a Researcher? Exploring Sensitivity to Semantic Violations in Italian BabyLMs

Alice Suozzi[1], Luca Capone[2], Gianluca E. Lebani[1], Alessandro Lenci[2]

[1] Ca' Foscari University of Venice, [2] University of Pisa

Correspondence: alice.suozzi@unive.it

Abstract

A large body of research has examined the linguistic abilities of language models (LMs) across various languages. However, conclusive evidence regarding their semantic competence and world knowledge remains limited, especially for low-resource languages. In this study, we explore the semantic competence of Italian BabyLMs, focusing on their sensitivity to semantic violations. To this end, we adapt a minimal pair benchmark targeting semantic violations to evaluate the semantic abilities of BAMBI, a family of small-scale models trained on progressively larger and more complex datasets. We further compare their performance, assessed through accuracy, mean log-likelihood offset, and expected calibration error, with that of three larger Italian LMs. Our findings shed light on this aspect of semantic competence in small-scale models and how this is affected by data scale and training strategies.

1 Introduction

This work extends research on language models' (LMs) sensitivity to semantic violations. We adapt an existing minimal pair benchmark targeting prototypical, unlikely, and impossible events (Kauf et al., 2023) to evaluate Italian Large LMs and small-scale BabyLMs trained on datasets of progressively increasing size and complexity.

Minimal pair benchmarks are widely used in linguistic evaluation and have recently been adopted for zero-shot LM assessment (Marvin and Linzen, 2018; Warstadt et al., 2020; Liu et al., 2024; Başar et al., 2025; Barbini et al., 2025; Jumelet et al., 2025). These datasets typically test syntactic competence through sentence pairs differing by a single element, offering a simple yet effective way to probe specific phenomena. Overall, LMs tend to assign higher probabilities to grammatical than to ungrammatical sentences across many constructions, suggesting that syntactic knowledge is acquired relatively robustly in early pretraining (Liu et al., 2021; Zhang et al., 2021). Even small models trained on limited data can reliably distinguish grammatical from ungrammatical sentences (Huebner et al., 2021). However, although performance generally improves with training, accuracy is not always monotonic and may vary with item frequency and lexical category (Wei et al., 2021; Chaves and Richter, 2021).

Similar patterns emerge for semantic and world/commonsense knowledge. LMs perform well on certain world-knowledge tasks (e.g., Levesque et al. 2012; Gordon et al. 2012; Zellers et al. 2018) and assign higher probabilities to verbs in typical semantic contexts (Cho et al., 2021; Kauf et al., 2023), yet results remain mixed. World/commonsense knowledge in contemporary LLMs is often brittle and highly sensitive to task formulation (Kauf et al., 2023, p. 4) (see also Ettinger 2020; Ribeiro et al. 2020; Ravichander et al. 2020; Elazar et al. 2021; Pedinotti et al. 2021). Moreover, semantic and commonsense knowledge appear subject to frequency effects, or more precisely, to reporting bias in pretraining corpora. Web-scraped data for adult audiences tend to underrepresent implicit commonsense knowledge while over-reporting rare or noteworthy events. Romero and Razniewski (2022) suggest that this bias can be mitigated by training on simplified child-directed corpora, which provide more explicit representations of everyday situations. The present study extends this line of research to Italian-trained language models along three main directions:

1. Evaluating Italian LMs on semantic violations, to tap into their semantic knowledge;

2. Adopting a comparative and developmental perspective, by assessing both Large LMs (2B, 3B, and 7B parameters) and BabyLM-style models trained on ecologically realistic corpora that simulate the linguistic input avail-

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 366–377

able across language acquisition stages (from roughly six to over twenty years of exposure);

3. Examining how a model semantic sensitivity evolves during training, by analyzing how the mean log-likelihood assigned to sentences varies as a function of noun frequency and syntactic structure (active vs. passive voice).

Together, these objectives aim to clarify how model scale, training data, and linguistic structure interact in shaping event knowledge in Italian LMs, building on recent work on LM development and acquisition. Further, as research on BabyLMs shows that reduced-scale models acquire syntactic competence relatively early but often struggle with semantic phenomena under a 100-million-word training budget (Warstadt et al., 2023; Hu et al., 2024), a benchmark of semantic minimal pairs directly probes this issue: i) **Does semantic event knowledge require substantially more exposure than syntactic knowledge, or do models begin to structure an appropriate semantic space for events and their prototypical participants from the earliest stages of pretraining?** ii) **how does this ability compare to that of Large LMs?** While larger architectures and more diverse data clearly advantage Large LMs, findings on reporting bias may conversely favor BabyLM-style models. Child-directed corpora are expected to encode commonsense information often absent from web-crawled data, potentially offering an advantage in modeling everyday semantic knowledge.

2 Semantic Knowledge and Semantic Violations

Semantic knowledge is central to human language. Understanding any linguistic message requires integrating the meanings of morphemes and words within broader linguistic (e.g., phrases, sentences) and extra-linguistic contexts. During comprehension, listeners and readers map incoming signals onto stored mental concepts and combine them to reconstruct intended meaning. This knowledge includes word meanings, syntactic constructions, and the probabilistic constraints governing how elements combine into larger units. Crucially, comprehenders continuously generate expectations about upcoming material based on semantic knowledge (Boland et al., 1990, 1995; Ferretti et al., 2001; McRae et al., 1998, 2005). For example, in *John is drinking* _, the object of *drink* is expected to be a consumable liquid, reflecting the verb's selectional preferences. Likewise, in _ *is drinking soda*, the subject is expected to be an animate, likely human, entity due to the semantic interaction between *drink* and *soda*. When such expectations are violated (e.g., *#John is drinking a rock* or *#A rock is drinking soda*), comprehenders must reanalyze the input. These violations yield measurable behavioral and neurophysiological effects. Behaviorally, readers show longer reading times or atypical eye movements (Coco et al., 2020). Neurophysiologically, semantic expectancy violations elicit the N400 ERP component, a negative deflection peaking around 400 ms, widely regarded as a marker of semantic integration difficulty (Kutas and Hillyard, 1980; Jouen et al., 2019; Coco et al., 2020; Kutas and Federmeier, 2011). Because such violations provide a privileged window into meaning construction, they are widely used to study semantic processing. They have informed debates on the relationship between semantic and syntactic processing, the temporal dynamics and cortical localization of integration, and semantic comprehension in clinical populations (e.g., aphasia, autism spectrum disorder; Fedorenko et al. 2020; Ivanova et al. 2021; see Kutas and Federmeier 2011).

Building on Kauf et al. (2023), whose dataset we adapt to Italian (cf. Section 3.1), we use semantic violations to test whether pretrained LLMs encode human-like generalized world knowledge about events (Kauf et al., 2023, p. 5). An event is defined as the action denoted by a verb together with its participants (e.g., for *drink*, a drinking event involving an Agent and a Patient). Following Kauf et al. (2023), we probe models' implicit event knowledge using two violation types: (a) **strong violations**, which breach verb selectional restrictions and yield impossible events (e.g., *#The table tidied the receptionist*); and (b) **weak violations**, which produce implausible but possible events given world knowledge (e.g., ?*The child drove the car*).

The detection of strong (a) and weak (b) violations relies on different, albeit intertwined, aspects of linguistic knowledge. On the one hand, the detection of strong violations depends primarily on "purely" semantic knowledge (e.g., selectional restrictions). On the other hand, the detection of weak violations relies more heavily on world knowledge. For this reason, in the present study we further investigate whether sensitivity varies across violation types (strong vs. weak).

3 Measuring LMs' sensitivity to semantic violations

3.1 Materials

Our dataset is the Italian adaptation of two datasets used by Kauf et al. (2023), in turn derived from cognitive and neurolinguistic studies by Fedorenko et al. (2020) and Ivanova et al. (2021). The datasets consist of minimal sentence pairs designed to assess semantic knowledge by manipulating plausibility. The Italian adaptation preserved the original structure and involved only linguistic translation, carried out semi-automatically. Sentences from both datasets were translated using ChatGPT (OpenAI, 2023) and then manually checked and revised by the authors. Specifically, one author reviewed each sentence to ensure that it was grammatical, meaningful, and semantically consistent with the corresponding English item. Overall, the automatic translation produced grammatical and meaningful sentences, while the main manual corrections involved replacing inappropriate lexical items. For the second dataset, passive counterparts of active transitive sentences were manually created, and verb tense was shifted to the past to ensure consistency with the first dataset. Example items are shown in Table 1.

The first dataset includes 1,648 sentences grouped into 824 items (i.e., minimal pairs of sentences). Each item consists of i) a sentence describing a transitive event and ii) a semantically anomalous version of the same sentence obtained by swapping the NP-subject and NP-object. Of the 824 items, 412 are active sentences, and the remaining 412 are their passive-voice counterparts (cf. Table 1). In addition, 522 out of 824 items form near-synonymous pairs differing only in word frequency: half of these items contain high-frequency words and the other half contain low-frequency (near-)synonyms. For instance, the low-frequency synomymous version of *La maestra ha comprato il computer* 'The teacher bought the laptop' is *L'insegnante ha acquistato il portatile* 'The instructor purchased the computer'.[1]

Finally, the items are grouped into three conditions according to the animateness of the NP-object: a) Animate-Inanimate (AI) items (n = 272, 160 with a synonymous version), in which the NP-object is inanimate. In these items, the role-reversal manipulation results in an impossible sentence that violates the verb's selectional restrictions on animateness (**strong semantic violation**); b) Animate-Animate (AA) items (n = 270, 176 with a synonymous version), in which both NPs are animate. Here, the role-reversal manipulation yields an implausible sentence whose anomaly is grounded in world knowledge rather than strict semantic constraints (**weak semantic violation**); c) Animate-Animate Reversible (AA-rev) items (n = 282, 186 with a synonymous version), which serve as the control condition. In this case, both NPs are animate and the sentences are reversible, resulting in two equally plausible sentences.

The second dataset now includes 62 items (minimal pairs), all falling under the AA condition introduced above in b). Of these, 22 items now have a passive-voice corresponding version, manually built by the authors (ITEM QR: *Il bagnino/la nonna ha salvato la nonna/il bagnino* 'The lifeguard/grandmother saved the grandmother/lifeguard' + passive versions: *La nonna/il bagnino è stata/o salvata/o dal/la bagnino/nonna* 'The grandmother/lifeguard was saved by the lifeguard/grandmother'). The remaining pairs contain intransitive verbs, not allowing a passive version (ITEM QL: *Il capo/lavoratore ha urlato contro al lavoratore/capo* 'The boss/worker yelled at the worker/boss'). No frequency-based synonymous versions are included in this dataset.

The Italian dataset includes 272 items for the AI condition (80-80 being high/low frequency synonyms, 112 without synonyms); 332 items for the AA condition (88-88 being high/low frequency synonyms, 156 without synonyms), and 282 items for the AA-rev condition (92-92 being high/low frequency synonyms, 282 without synonyms).

To ensure the efficacy of the adaptation process, all authors reviewed and checked the adapted sentences, which were subsequently validated by a sample of 232 Italian-speaking participants. Specifically, for each sentence pair (cf. Table 1), participants were asked to rate the plausibility of the event described by each sentence on a seven-point scale. For example, for the pair *La maestra ha comprato il computer / Il computer ha comprato la maestra* 'The teacher bought the laptop / The laptop bought the teacher', participants answered the fol-

[1] As for the low-frequency version, the pair also contains the active impossible sentence *Il portatile ha acquistato l'insegnante* 'The computer purchased the instructor'. The passive minimal pair is included as well: possible sentence *Il portatile è stato acquistato dall'insegnante* 'The computer was purchased by the instructor'; impossible sentence: *L'insegnante è stata acquistata dal portatile* 'The instructor was purchased by the computer'

Condition	Judgment	Voice	Sentence
Animate-Inanimate	Possible	Active	La maestra ha comprato il computer 'The teacher bought the laptop'
	Impossible	Active	Il computer ha comprato la maestra 'The laptop bought the teacher'
	Possible	Passive	Il computer è stato comprato dalla maestra 'The laptop was bought by the teacher'
	Impossible	Passive	La maestra è stata comprata dal computer 'The teacher was bought by the laptop'
Animate-Animate	Plausible	Active	Il paramedico ha rianimato il giovane 'The paramedic revived the youth'
	Implausible	Active	Il giovane ha rianimato il paramedico 'The youth revived the paramedic'
	Plausible	Passive	Il giovane è stato rianimato dal paramedico 'The youth was revived by the paramedic'
	Implausible	Passive	Il paramedico è stato rianimato dal giovane 'The paramedic was revived by the youth'
Animate-Animate Reversible	Reversible	Active	Il prete ha abbracciato il fedele 'The preacher hugged the churchgoer'
	Reversible	Active	Il fedele ha abbracciato il prete 'The churchgoer hugged the preacher'
	Reversible	Passive	Il fedele è stato abbracciato dal prete 'The churchgoer was hugged by the preacher'
	Reversible	Passive	Il prete è stato abbracciato dal fedele 'The preacher was hugged by the churchgoer'

Table 1: Example minimal pairs of sentences from our dataset, shown for each condition. Each sentence pair is labeled for the corresponding judgment.

lowing two questions: (1) How plausible is it that the teacher bought the computer? (1 = Impossible, 7 = Absolutely plausible); and (2) How plausible is it that the computer bought the teacher? (1 = Impossible, 7 = Absolutely plausible). Data were collected using an online survey administered through LimeSurve (LimeSurvey GmbH, 2026) and participants were recruited via Prolific (2026). Possible sentences received a mean plausibility rating of 6.73, whereas impossible sentences received a mean rating of 1.59 ($\Delta = 5.13$). Plausible sentences received a mean rating of 6.64, whereas implausible sentences received a mean rating of 3.34 ($\Delta = 3.29$). Finally, reversible sentences received mean ratings of 6.19 and 5.71, respectively ($\Delta = 0.48$). These results are consistent with those obtained for the English datasets (Fedorenko et al., 2020; Ivanova et al., 2021), thereby confirming the efficacy of our adaptation.

3.2 Models

The **BAMBI** model adopts a lightweight GPT-2–style decoder architecture, comprising approximately 130 million parameters (Table 2). It is progressively trained on a corpus consisting of transcripts of Child-Directed speech and multimedia content designed for children (Suozzi et al., 2025; Capone et al., 2025). The dataset is organized into four tiers of increasing linguistic complexity, corresponding to the age ranges 0-6, 6-12, 12-18, and 18-24. For comparison, in addition to the **BAMBI_CL** (Curriculum Learning) models, two additional variants were trained. **BAMBI_mc4** is trained on a random subset of the mC4 dataset (Xue et al., 2021), a large corpus derived from the public Common Crawl web scrape and commonly used for standard language model pretraining. This model shares the same architecture as BAMBI but does not employ CL. **BAMBI 0-18_noCL** is trained on the first three subsets of the BAMBI dataset (covering ages 0–18) in fully shuffled order, thus removing the curriculum structure while retaining exposure to child-directed data. The BabyLMs are evaluated against three Italian-oriented Large LMs (Tables 2 and 3): **Minerva_3B**, trained on a relatively small corpus (Orlando et al., 2024); **Velvet_2B**, trained on roughly 3 trillion tokens[2] and **Cerbero_7B**, whose training data size remains undisclosed (Galatolo and Cimino, 2023).

Architecture	Vocab. Size	L x H	Hidden Size	Trainable Params
BAMBI	30,000	12x12	768	135,856,128
Velvet_2B	126,976	28x32	2,048	2,223,097,856
Minerva_3B	32,768	32x32	2,560	2,894,236,160
Cerbero_7B	32,000	32x32	4,096	7,241,732,096

Table 2: Models hyperparameters.

[2] https://huggingface.co/Almawave/Velvet-2B

Model	Data size	Epochs	CL
BAMBI 0-6_CL	26M words	16	1 step
BAMBI0-12_CL	57M words	16	2 steps
BAMBI 0-18_CL	86M words	[13,18,10]	3 steps
BAMBI0-18_noCL	86M words	[13,18,10]	no
BAMBI_mc4	86M words	20	no
BAMBI 0-24_CL	117M words	[13,18,10,7]	4 steps
Velvet_2B	3T tokens	1	no
Minerva_3B	660B tokens	1	no
Cerbero_7B	UNK	1	no

Table 3: Training details of the BAMBI family models and the baseline models.

3.3 Methods and analyses

The primary goal of our analysis was to evaluate whether each model's probability distribution conforms to semantic constraints. More specifically, we investigated how models' sensitivity to semantic violations varies as a function of training and violation type.

To determine which sentence in each minimal pair was preferred by each model, we used **mean log-likelihood** (MLL), following Marvin and Linzen (2018).

To evaluate the models, we employed three metrics providing complementary insights into model performance. First, we used accuracy, a standard metric for evaluating LMs, which measures how often a model's preference distribution respects semantic constraints. In other words, accuracy captures how often a model correctly assigns a higher probability to the possible/plausible sentence than to the impossible/implausible one. Data concerning accuracy were analyzed using a generalized linear mixed-effects model fitted by maximum likelihood via Laplace approximation. Condition, Model and Voice (active/passive sentence) are fixed effects, while Item is a random effect. We also computed the expected marginal means, in order to perform the pairwise comparisons (with Tukey correction).[3]

To further investigate model sensitivity, we computed the absolute MLL offset between the two sentences in each minimal pair, thereby providing a measure of the strength of the model's preference. This method is conceptually similar to the activation offset proposed by Zhou et al. (2025), but applied to log-likelihood differences. Specifically, for each sentence pair, the relevant measure corresponds to the difference between the MLL assigned to the preferred sentence and that assigned to the dispreferred one.

This metric is particularly informative in the context of strong and weak violations. For strong violations, a larger distance between possible and impossible sentences is expected, whereas a smaller distance is expected between plausible and implausible sentences in weak violations (with no distance expected for reversible sentences).[4] Employing this metric therefore allows us to determine whether models, beyond simply preferring the correct sentence in both conditions, exhibit qualitative sensitivity to different types of violations.

A linear mixed-effects model was then fitted to predict these sensitivity scores from model type and item-level linguistic features, which were treated as fixed effects, while including a random intercept for sentence pairs. Estimated marginal means were also computed in order to perform pairwise comparisons with Tukey correction. For this analysis, we included only the offsets corresponding to correct model predictions, ensuring that the sensitivity measure reflected reliable model choices and facilitating model fitting.

As a third metric, we employed the Expected Calibration Error (ECE: Guo et al. 2017; Boseak 2025; Pavlovic 2025) for each model, to further assess calibration quality and examine how model behavior varied with model size and training duration. ECE measures how closely a model's predicted probabilities align with its observed performance. A well-calibrated model should be correct approximately as often as it is confident. Accordingly, ECE is computed as a weighted average of the absolute difference between average confidence and observed accuracy (Pavlovic, 2025).

Finally, we analyzed the relationship between ECE and accuracy across strong and weak violations in order to better understand how confidence and correctness interacted in different semantic contexts and models.

All analyses were conducted in R (R Core Team 2024, Version 4.3.3) using the packages `lme4` (Bates et al., 2015) and `emmeans` (Lenth and Piaskowski, 2025).

3.4 Results

In the following sections, we present the results of the models concerning accuracy, MLL offsets, and ECE.

[3] Frequency was initially included in the model but removed later, as preliminary analyses showed it did not contribute significantly to explaining variance.

[4] Notably, this pattern is consistent with human judgments (cf. the Δ values reported in Section 3.1).

3.4.1 Accuracy

The accuracy achieved by the models is shown in Figure 1. Overall, all models perform above chance level, as they achieve an accuracy above 0.50 in AA and AI conditions. AA-rev condition is excluded from this analysis, as both sentences composing each minimal pair are correct.

Starting with the AA condition (plausible–implausible), Minerva_3B achieves the highest accuracy, despite being the least extensively trained among the large language models, followed closely by Cerbero_7B. Turning to the BAMBI models, they generally achieve slightly lower accuracies than the larger models, with the exception of the youngest model, BAMBI 0-6_CL, which shows the lowest accuracy overall. Surprisingly, Velvet_2B achieves a slightly above-chance accuracy despite its size. The same general pattern is observed in the AI condition (possible–impossible pairs), although all models achieve higher accuracy in this condition than in the AA condition.

The generalized linear mixed-effects model (cf. Section 3.3) reveals a significant interaction between Condition and Model: almost all models achieve significantly higher accuracy in the AI condition ($p < 0.001$). The only models for which the difference in accuracy is not significant are BAMBI 0-12_CL (AI vs AA, $p = 0.9102$); BAMBI 0-24_CL (AI vs AA, $p = 1$); BAMBI mc4 (AI vs AA, $p = 1$); Velvet_2B (AI vs AA, $p = 0.496$). In addition, pairwise comparisons show that Minerva_3B is the only model that consistently outperforms the BAMBI family: it is more accurate than any BAMBI variant in both AA and AI conditions and shows the largest benefit from violation strength (AA–AI log-odds gap ≈ -1.96, $p < 0.001$). Cerbero_7B only performs significantly better than the youngest BAMBI model (0-6_CL) in both conditions ($p = 0.0029$ and $p < 0.001$, respectively), while comparisons with older BAMBI variants yield non-significant differences ($p > 0.1$). By contrast, Velvet_2B is statistically indistinguishable from BAMBI in AA condition, and significantly worse than older BAMBI variants in AI condition. Finally, all models except Velvet are more accurate in AI than in AA condition, indicating sensitivity to semantic violation strength.[5] Finally, Voice (active *versus* passive sentences) has an overall significant negative effect on accuracy (β= -0.376; SE = 0.061; z = -6.160; $p < 0.001$): all models tend to perform worse in minimal pairs with passive sentences.

[5]The tables reporting the pairwise comparisons, both for overall accuracy and for accuracy by condition, will be included in the Appendix upon acceptance of the paper.

3.4.2 MLL offset

Table 4 reports the exact mean ($\pm$ SD) values of MLL offsets across the three conditions (AI, AA and AA-rev), while the distribution of MLL offsets across the three conditions (AI, AA and AA-rev) is given in Appendix A (Figure 3).

Model	AI	AA	AA-rev
BAMBI 0-6_Cl	0.68 ($\pm$ 0.57)	0.56 ($\pm$ 0.48)	0.53 ($\pm$ 0.48)
BAMBI 0-12_CL	0.63 ($\pm$ 0.53)	0.40 ($\pm$ 0.36)	0.42 ($\pm$ 0.37)
BAMBI 0-18_CL	0.58 ($\pm$ 0.43)	0.30 ($\pm$ 0.25)	0.29 ($\pm$ 0.24)
BAMBI 0-18_noCL	0.61 ($\pm$ 0.47)	0.36 ($\pm$ 0.28)	0.32 ($\pm$ 0.30)
BAMBI 0-24_CL	0.56 ($\pm$ 0.43)	0.32 ($\pm$ 0.28)	0.30 ($\pm$ 0.27)
BAMBI mc4	0.57 ($\pm$ 0.45)	0.37 ($\pm$ 0.33)	0.33 ($\pm$ 0.26)
Cerbero_7B	0.50 ($\pm$ 0.39)	0.33 ($\pm$ 0.27)	0.28 ($\pm$ 0.25)
Minerva_3B	0.63 ($\pm$ 0.38)	0.33 ($\pm$ 0.27)	0.20 ($\pm$ 0.16)
Velvet_2B	2.92 ($\pm$ 2.20)	2.50 ($\pm$ 1.79)	2.36 ($\pm$ 2.03)

Table 4: Mean ($\pm$ Standard Deviation) of the LLM offsets across all conditions for all Models

Model sensitivity towards different kinds of semantic violations (i.e., strong-AI *versus* weak-AA *versus* no violation-AA-rev) was assessed through absolute MLL offsets computed on correct items (that is, items in which models selected the correct sentence) and all reversible items. This provides a measure of the strength of the model's preference for the selected sentence within a minimal pair (higher values indicate stronger confidence). Table 4 shows that across models offsets were generally higher in AI condition (possible–impossible) and progressively lower in AA (plausible–implausible) and AA-rev condition. This reveals that models express greater sensitivity when detecting strong semantic violations than weak ones, for which they need to distinguish plausible from implausible sentences. Within the BAMBI family, absolute offsets tend to decrease as training coverage increases for all conditions (cf. Table 4), indicating that extensively trained models make less extreme predictions. However, the sensitivity gap between AI and AA conditions remains relatively constant ($\approx$ 0.20-0.30), suggesting that expanded exposure and curriculum progression do not reduce the semantic asymmetry observed in the measure. Among the larger LMs, Minerva_3B and Cerbero_7B show sensitivity magnitudes similar to the latest BAMBI variants but display a clearer increase under AI conditions, particularly for Minerva_3B, consistent with its higher accuracy on those items. In stark contrast, Velvet_2B exhibits extremely high offsets

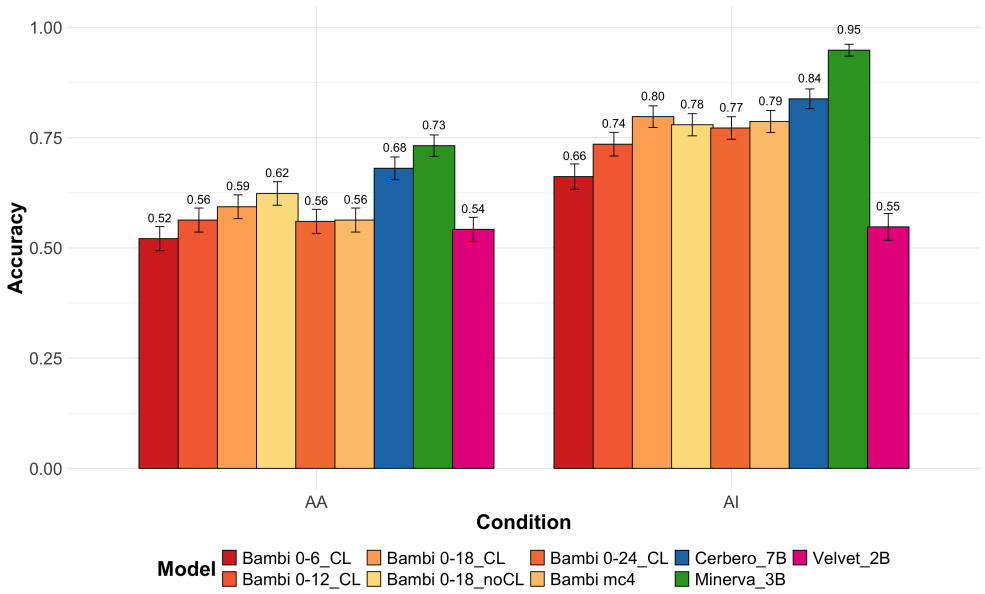

Figure 1: Distribution of accuracy across models and conditions.

across all conditions (≈ 2.50-3.00), several times greater than any other model. The linear mixed-effects model indicates that the most influential factor is Model, rather than Condition. Indeed, the pairwise comparisons reveal that smaller models (e.g., BAMBI 0-6_CL and BAMBI 0-12_CL) significantly differ from larger ones (especially Velvet_2B, but also Cerbero_7B and Minerva_3B) under all conditions ($p < 0.001$). On the contrary, Condition and Model interact when comparing smaller models. For instance, BAMBI 0-6_CL significantly differ from BAMBI 0-18_CL only when comparing the offsets of different conditions, while they do not differ when considering the same condition.

The data on Large LMs' (Velvet_2B, Minerva_3B, Cerbero_7B) sensitivity to semantic violations suggests uncalibrated confidence unaligned with accuracy, an instance of systematic over-confidence or under-confidence rather than genuine precision. ECE, the measure presented in the following section, aims to further explore this aspect.

3.4.3 Expected Calibration Error

Calibration was evaluated by calculating the Expected Calibration Error (ECE) over AI and AA items (cf. Section 3.3). As Table 5) shows, all BAMBI models showed low ECE values, indicating generally well calibrated confidence, with the best calibration displayed by BAMBI 0–12_CL. Larger models such as BAMBI 0-18_CL and 0-18_noCL exhibited slightly higher values, suggesting that increased training coverage did not systematically improve calibration. Overall, BAMBI models maintained balanced confidence and accuracy, with no clear trend toward over- or under-confidence as training scale increased. The web-trained BAMBI_mc4 performed comparably to the curriculum variants. In contrast, Minerva_3B and Cerbero_7B, despite higher accuracies, displayed higher ECEs, reflecting reduced calibration precision. Velvet_2B showed the poorest calibration, combining low accuracy with very high confidence, an instance of strong overconfidence misaligned with performance.

Model	ECE	Accuracy	Confidence
BAMBI 0-6_CL	0.053	0.584	0.638
BAMBI 0-12_CL	0.026	0.641	0.616
BAMBI 0-18_CL	0.084	0.685	0.602
BAMBI 0-18_noCL	0.084	0.694	0.610
BAMBI 0-24_CL	0.054	0.656	0.602
BAMBI mc4	0.055	0.664	0.609
Velvet_2B	0.309	0.545	0.854
Minerva_3B	0.219	0.829	0.611
Cerbero_7B	0.155	0.752	0.597

Table 5: ECE, accuracy and confidence of the BAMBI family models and the baseline models.

Figure 2 plots models' accuracy against ECE across the two semantic conditions. A general positive slope emerges: models with higher accuracy also tend to exhibit higher ECE, indicating reduced calibration. The shift from plausible–implausible (AA) to possible–impossible (AI) items consistently moves models toward higher accuracy but

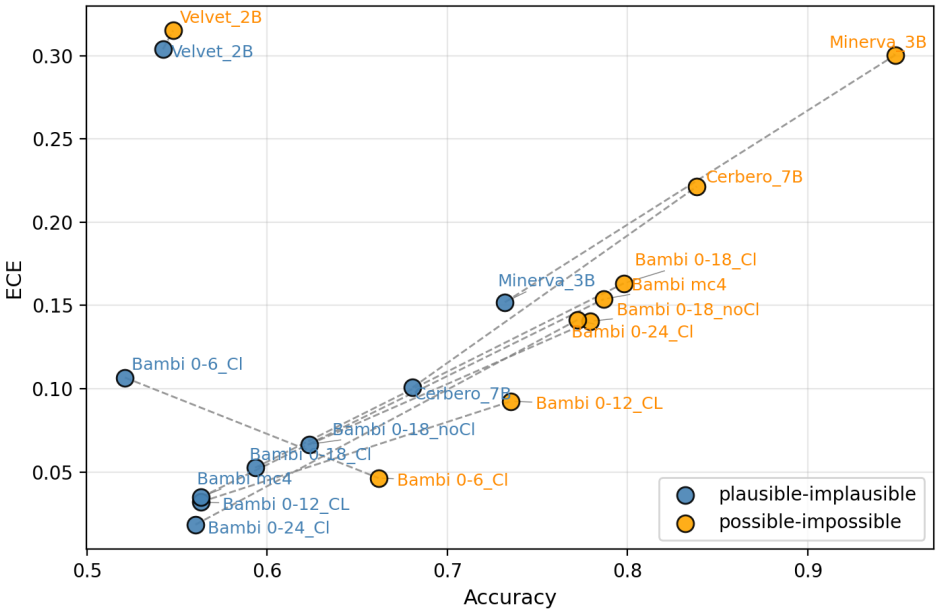

Figure 2: Accuracy and calibration of models across conditions.

slightly poorer calibration, suggesting greater overconfidence when the violation is easier to detect. The BAMBI models form a compact, low-ECE cluster with consistent condition effects, whereas Minerva-3B and Cerbero-7B extend the trend upward in accuracy and ECE. Velvet_2B, by contrast, combines the lowest accuracy with the highest ECE, standing out as a clear case of miscalibration. The joint inspection of accuracy, mean confidence, and ECE (Table 5) further clarifies model calibration patterns. The BAMBI models display a near-aligned accuracy–confidence ratio, with average confidences differing from empirical accuracies by less than 0.05-0.08, consistent with their low ECE values. This indicates balanced calibration across curriculum stages and training regimes. In contrast, Minerva_3B and Cerbero_7B, despite achieving the highest accuracies, show systematic underconfidence, with mean confidences well below observed accuracies, reflected in their higher ECEs. Velvet_2B represents the opposite extreme: its mean confidence vastly exceeds its accuracy, yielding the highest ECE and evidencing strong overconfidence and poor calibration. Overall, these results confirm that the BAMBI family maintains a stable, well calibrated accuracy-confidence ratio, while larger LMs tend toward miscalibration, underconfident in the case of Minerva and Cerbero, and overconfident in Velvet.

4 Discussions

The main aim of this study was to assess the semantic knowledge of Italian BabyLMs, focusing on their sensitivity to strong and weak semantic violations. To this end, we not only employed accuracy as a metric; rather, we examined general model behavior regarding semantic violations, quantified through MLL offset and ECE, as a function of training (i.e., data quality and quantity, and model size).

Overall, BAMBI and larger models can detect semantic violations, achieving above-chance accuracy in both AI and AA conditions. They also display varying confidence depending on the violation type (strong vs. weak), mirroring human-like responses reported in prior psycho- and neurolinguistic studies (e.g., Fedorenko et al. 2020; Ivanova et al. 2021, cf. also Section 3.1). Smaller models, such as BAMBI 0–6_CL or 0–12_CL, exhibit the same quantitative and qualitative patterns despite lower overall accuracy.

Focusing on BAMBI variants, accuracy highlights the impact of training data quality: BAMBI mc4 performs slightly worse than BAMBI 0–18_CL, consistent with the effect of reporting bias in models trained exclusively on written corpora (cf. Section 1). Differences related to curriculum learning (CL) are negligible. Moreover, improvement rates vary by semantic condition: in "older" BAMBI models, the gap between AA and AI accuracy widens progressively.

For MLL offset, BabyLMs generally show a

decrease with training. Mc4 and noCL variants have slightly broader offset distributions than their CL counterparts (cf. Figure 3, Appendix A). CL training and data quality do not significantly affect calibration, which scales straightforwardly with accuracy. Consequently, models exhibit higher ECE on possible-impossible items, where they achieve highest accuracy, except BAMBI 0–6_CL, which shows higher ECE on the more challenging AA (plausible–implausible) items, reflecting less calibrated predictions. This suggests that sensitivity to weak semantic violations may require more training than available to this model.

Finally, our results underscore the utility of sentence minimal pairs, a state-of-the-art method for assessing both Large LMs and BabyLMs (a.o. Warstadt et al. 2020 and subsequent versions). However, while models typically achieve maximum accuracy on syntactic tasks, our findings indicate that minimal-pair tasks are more challenging when probing semantic competence.

5 Conclusion

This study examined the event semantic knowledge of Italian BabyLMs by testing their reactions to semantic violations through an Italian minimal-pair benchmark. Analyses of accuracy, MLL offsets, and calibration reveal that even the smallest BabyLMs show sensitivity to strong semantic violations, indicating that core event-level knowledge and knowledge of selectional restrictions appear in the very first stages of pretraining. In contrast, the ability to discriminate plausible from implausible sentences - a subtler, world knowledge–driven skill - requires a larger word budget, becoming calibrated only beyond the 0–6M word exposure. Calibration results further show that model confidence grows in line with accuracy, with all BabyLMs maintaining balanced behavior across training stages. Finally, the slightly superior performance of models trained on child-directed data, compared to those exposed to web-scraped text (and same amount of words), provides a first indication of reporting bias: input that explicitly encodes everyday experiences seems to foster more stable and contextually grounded semantic representations.

As for future research, the dataset used in this study is part of a larger benchmark designed to assess different aspects of linguistic knowledge in Italian (Baby)LMs, with a particular focus on the syntax–semantics interface (e.g., prototypicality of arguments, argument structure alternations, and syntactically optional arguments, among others). This benchmark is currently being validated with Italian-speaking adults.

References

Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2025. BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. Turblimp: A turkish benchmark of linguistic minimal pairs. *arXiv e-prints*, pages arXiv–2506.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Julie E. Boland, Michael K. Tanenhaus, and Susan M. Garnsey. 1990. Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language*, 29(4):413–432.

Julie E. Boland, Michael K. Tanenhaus, Susan M. Garnsey, and Greg N. Carlson. 1995. Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language*, 34(6):774–806.

Christopher Boseak. 2025. Evaluating log-likelihood for confidence estimation in llm-based multiple-choice question answering. *Innovative Journal of Applied Science*, 2(4):29.

Luca Capone, Alice Suozzi, Gianluca E. Lebani, and Alessandro Lenci. 2025. Bambi goes to school: Evaluating italian babylms with invalsi-ita. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*.

Rui P. Chaves and Stephanie N. Richter. 2021. Look at that! bert can be easily distracted from paying attention to morphosyntax. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 28–38.

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the influence of verb aspect on the activation of typical event locations with BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2922–2929.

Moreno I. Coco, Antje Nuthmann, and Olaf Dimigen. 2020. Fixation-related brain potentials during semantic integration of object–scene information. *Journal of Cognitive Neuroscience*, 32(4):571—-589.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203(104348).

Todd R. Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Federico A Galatolo and Mario GCA Cimino. 2023. Cerbero-7B: A Leap Forward in Language-Specific LLMs Through Enhanced Chat Corpus Generation and Evaluation. *arXiv preprint arXiv:2311.15698*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21.

Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.

Anna A. Ivanova, Zachary Mineroff, Vitor Zimmerer, Nancy Kanwisher, Rosemary Varley, and Evelina Fedorenko. 2021. The language network is recruited but not required for nonverbal event semantics. *Neurobiology of Language*, 2.

Anne-Lise Jouen, Nicolas Cazin, Carol Madden-Lombardi Sullivan Hidot, Jocelyne Ventre-Dominey, and Peter Ford Dominey. 2019. Beyond the word and image: III. Neurodynamic properties of the semantic network. *bioRxiv*, 62:621–647.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47(11).

Marta Kutas and Kara D. Federmeier. 2011. Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, 62:621–647.

Marta Kutas and Steven A. Hillyard. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Russell V. Lenth and Julia Piaskowski. 2025. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 2.0.0.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561.

LimeSurvey GmbH. 2026. Limesurvey: An open source survey tool.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, and 1 others. 2024. Zhoblimp: a systematic assessment of language models with linguistic minimal pairs in chinese. *arXiv e-prints*.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Ken McRae, Mary Hare, Jeffrey E. Elman, and Todd R. Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7):1174–1184.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312.

OpenAI. 2023. *ChatGPT (March 14 version)*.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719.

Maja Pavlovic. 2025. Understanding model calibration-a gentle introduction and visual exploration of calibration and the expected calibration error (ECE). In *The Fourth Blogpost Track at ICLR 2025*.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 1–11.

Prolific. 2026. Prolific.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Version 4.3.3.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Julien Romero and Simon Razniewski. 2022. Do children texts hold the key to commonsense knowledge? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10954–10959.

Alice Suozzi, Luca Capone, Gianluca E Lebani, and Alessandro Lenci. 2025. Bambi: Developing baby language models for italian. *Lingue e linguaggio, Rivista semestrale*, 1/2025:83–102.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and 1 others. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888.

A Appendix A

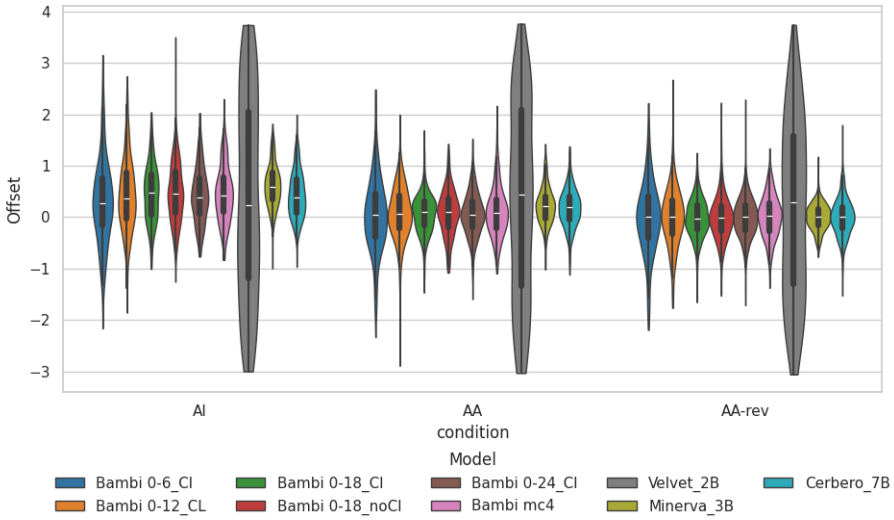

Figure 3: Distribution of MLL offsets across models and conditions.