Zero-Shot Evaluation of Conversational Language Competence in Data-Efficient LLMs Across English, Mandarin, and French

Sheng-Fu Wang
Academia Sinica
Institute of Linguistics
Taipei, Taiwan
sftwang@gate.sinica.edu.tw

Ri-Sheng Huang
Department of CSIE
National Taiwan University
Taipei, Taiwan
r13922102@csie.ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
shukaihsieh@ntu.edu.tw

Laurent Prévot
CNRS & Aix Marseille Univ., LPL, France
CNRS & MEAE, CEFC, Taiwan
laurent.prevot@univ-amu.fr

Abstract

Large Language Models (LLMs) have achieved outstanding performance across various natural language processing tasks, including those from Discourse and Dialogue traditions. However, these achievements are typically obtained thanks to pretraining on huge datasets. In contrast, humans learn to speak and communicate through dialogue and spontaneous speech with only a fraction of the language exposure. This disparity has spurred interest in evaluating whether smaller, more carefully selected and curated pretraining datasets can support robust performance on specific tasks. Drawing inspiration from the BabyLM initiative, we construct small (10M-token) pretraining datasets from different sources, including conversational transcripts and Wikipedia-style text. To assess the impact of these datasets, we develop evaluation benchmarks focusing on discourse and interactional markers, extracted from high-quality spoken corpora in English, French, and Mandarin. Employing a zero-shot classification framework inspired by the BLiMP benchmark, we design tasks wherein the model must determine, between a genuine utterance extracted from a corpus and its minimally altered counterpart, which one is the authentic instance. Our findings show that models pretrained on conversational data exhibit an advantage in handling discourse and interactional markers compared to those trained on written or encyclopedic text. Furthermore, the models, trained on small amount spontaneous speech transcripts, perform comparably to standard LLMs.[1]

[1]Source code, pretraining data and benchmarks are available at https://github.com/rszia/sp2bench

1 Introduction

The performance levels achieved by current large language models (LLMs) are impressive across all natural language processing tasks, including those from the Discourse and Dialogue traditions, which have traditionally been considered more challenging. However, these results are typically obtained by pretraining models on enormous amounts of data. In contrast, humans learn to speak and communicate through dialogue and spontaneous speech with only a fraction of the language exposure required by artificial models. More broadly, there is growing interest in evaluating whether much smaller and carefully selected pretraining datasets can support performance on specific tasks.

Initiatives like BabyLM (Warstadt et al., 2023b) explore this question by providing English pretraining datasets of 10M and 100M tokens, as well as evaluation benchmarks, including BLiMP (Warstadt et al., 2020a), which aims to evaluate the linguistic competence of LLMs. Such pretraining datasets include a majority of spontaneous speech transcripts originating from child-directed speech (MacWhinney, 2014) and everyday conversations.

The evaluation metrics employed, while a good starting point, appear however biased in two ways: (i) they tend to favor canonical written forms ; and (ii) prioritize syntactic, semantic, and commonsense pragmatic competence. Language and communicative competence include many other dimensions, in particular if one focuses on spontaneous speech in a conversational context. Indeed, most benchmarks concerned with language are centered on high-level tasks and relatively standard morphological and syntactic tasks revolving

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 32–47
Aug 25–27, 2025.

around the idea of grammaticality. Grammaticality is an ill-defined concept when used to analyze spontaneous speech productions. In this context, *discourse and interactional marking*, as well as linguistic performance-management like *disfluencies*, are crucial.

Moreover, most of these benchmarks are for English, whereas English constitutes the exception rather than the norm, at least with regard to data availability. Due to data scarcity, it is still impossible to gather a 100M-token dataset based solely on real spoken conversational data, but the 10M-tokens dataset is accessible for a few languages like English, French, Mandarin, and a few others.

Our stance focussing on 'Conversational speech' is adopted because it is the genre through which humans acquire their basic language skills, and it is also of special significance with regard to language emergence (Levinson, 2020; Christiansen and Chater, 2022). Moreover, it is quite distant from the usual written or web content on which LMs are trained, increasing the risk of biases against this crucial genre in the resulting models.

Our focus in this paper is to assess the adequacy of LLMs with respect to conversational and spontaneous speech phenomena. We construct "small" (10M-tokens) pretraining datasets from different sources (conversational vs. Wikipedia-style text). We also create evaluation benchmarks extracted from high-quality spoken corpora in English, French, and Mandarin. More specifically, we focus on discourse and interactional markers (Schiffrin, 1987) and fillers (Shriberg, 1994). Following the approach of BLiMP (Warstadt et al., 2020a), we design zero-shot classification tasks based on these phenomena. Our results show that pretraining data type has a clear impact on the model's ability to handle these features: conversational data offers an advantage over written encyclopedic text. Moreover, standard LLMs do not seem to have a strong advantage compared to our small models trained on spontaneous speech transcripts.

The paper is organized as follows. We present the related work (Section 2) and justify our proposal in more detail in Section 3. We then detail our methods and describe the creation of the pretraining datasets and benchmarks (Section 4). The next section presents the experiments, their hypotheses, and the results (Section 5). We conclude with a discussion of the limitations of this work and future directions.

2 Related Work

Since the emergence of large language models, there has been interest from the computational linguistics community in understanding why they are so successful. Warstadt et al. (2020b) explore the conditions (e.g., the amount of training data) under which RoBERTa develops and leverages linguistic features, such as part of speech (POS) and morphology, as opposed to relying on simpler surface-level features like position-based or length-based cues. More recently, several studies have probed LLMs to better characterize their performance across various domains, particularly with regard to linguistic competence versus commonsense reasoning. These studies have also examined the relationship between model performance and the amount of training data required for different tasks. In particular, Zhang et al. (2021) used training sets of varying sizes (1M, 10M, 100M, and 1B tokens) to show that syntactic and semantic competence becomes robust in the 10M–100M range, whereas larger datasets are needed to achieve strong results in pragmatic and commonsense reasoning tasks.

More broadly, there have been proposals for evaluating the performance of LLMs on diverse linguistic tasks. Warstadt et al. (2019b) leveraged a substantial body of generative syntax-semantics literature to develop benchmarks based on acceptability judgments, drawing either from the linguistic literature, as in the CoLA benchmark, or by exploiting more sources and data augmentation methods, as in BLiMP (Warstadt et al., 2020a). In addition to these binary decision tasks, Zhang et al. (2021) combined three other types of evaluation metrics: *classifier probing* (following (Ettinger et al., 2016; Adi et al., 2017)), which includes tasks from POS tagging to coreference resolution; *information-theoretic probing* based on the minimum description length (MDL) principle; and *fine-tuning on higher-level tasks* such as those in the SuperGLUE benchmark.

BLiMP (Warstadt et al., 2019a) has inspired a series of language-specific benchmarks, such as CLiMP for Mandarin Chinese (Xiang et al., 2021), as well as benchmarks for other languages like Japanese (Someya and Oseki, 2023), Dutch (Suijkerbuijk et al., 2025), Russian (Taktasheva et al., 2024), German (Bunzeck et al., 2025), Italian (Suozzi et al., 2025), and now a multilingual BLiMP (Jumelet et al., 2025). These are important additions to the evaluation landscape. While

these benchmarks represent important extensions to the general evaluation framework, they all rely on syntax-semantics structures mostly derived from introspection and textbook data.

Turning now to conversational training data, even before the LLM era, Pannitto et al. (2020) trained a model on child-directed speech. In terms of specialized language models, Cabiddu et al. (2025) developed LMs based on child-directed speech transcripts and evaluated them on word-sense disambiguation tasks. They concluded that word acquisition trajectories could be better captured by multimodal models that incorporate acoustic features, among other aspects.

The discourse and dialogue phenomena we investigate to increase awareness and methods for making LLMs more conversation-aware[2] are well-studied phenomena. Discourse markers (Schiffrin, 1987) have been extensively studied in terms of their definition, analysis, and use in discourse parsing (Schilder, 2002). There are different sorts of discourse markers, ranging from discourse-semantic connectives (*but, because, ...*) to attitudinal or interpersonal markers (*ah, oh, ...*), adverbs (*frankly, actually, ...*), or conversational feedback markers (*yeah, ok, ...*). Recently, Sadlier-Brown et al. (2024) compared human and LLM performance in predicting the presence of *actually*.

Disfluencies (Shriberg, 1994) have long been known as a hallmark of spontaneous speech. There is a vast literature on their detection, processing, and, more recently, their generation by LLMs to produce fluent yet natural-sounding speech (Zohaib Hassan et al., 2024). These two dimensions (discourse markers and disfluencies) benefit from being studied jointly, as they are known to interact deeply (Crible and Pascual, 2020).

3 A Proposal for a New Source of Metrics

Most initiatives for linguistic LLM evaluation are grounded in text-based and/or handcrafted paradigms, potentially coupled with behavioral and/or physiological lab measures. In contrast, we propose using actual spontaneous conversational transcripts to build complementary benchmarks that test real-world language use in spontaneous speech. These metrics will remain fundamentally linguistic in nature rather than focusing on end-to-end evaluation.

Language is acquired, especially in its early stages, within spontaneous, conversational environments. While conversational language shares grammatical structures with other genres, its unique characteristics suggest that simply listing syntactic "errors" or semantic incongruities does not fully capture linguistic competence. Furthermore, in a conversational context, what may be considered a production error from a normative grammatical perspective is often perfectly acceptable and successfully achieves its communicative purpose. Therefore, we aim to develop a complementary approach that provides a broader set of metrics for evaluating language models from both cognitive and communicative perspectives.

Specifically, we propose using spontaneous speech corpora, as they offer insights into human language processing through various observable production phenomena. Our approach is a form of *classifier probing* (Ettinger et al., 2016; Adi et al., 2017; Warstadt et al., 2019b), but rather than focusing on meta-linguistic tasks (e.g., predicting syntactic categories), we aim to predict phenomena that serve as partial indicators of language processing. (Prévot et al., 2024; Wang et al., 2025) explored the prediction of speech-production variables, such as speech reduction and prosodic prominence; however, this approach requires fine-tuning a pretrained model, which obscures the overall interpretation of the results. We propose here a preliminary set of simple tasks exploiting lexical markers related to spontaneous speech, discourse, and interactional competence.

We approach the question by designing tasks involving several kinds of discourse markers and fillers (See Table 3 for examples). Different types of discourse markers are prototypical of different genres (e.g., conversational vs. textual) and therefore constitute an ideal testing ground for our experiments. Fillers are a crucial device and hallmark of spontaneous speech. Our work follows the idea that fillers are a crucial ingredient to consider when modeling language, especially when addressing the issue of written bias in linguistic and NLP models (O'Connell and Kowal, 2004). Moreover, fillers are closely associated with discourse markers (Crible, 2018), making their inclusion in our set of tasks an obvious choice.

[2]Conversational AI has become an important keyword nowadays. However, although conversational AI systems can be efficient and interactive, they do not exhibit many features of real human-human conversation.

4 Data and Method

We built the reference datasets from existing high-quality linguistic corpora and created the "tasks" based on these reference datasets; and, how we selected and processed the base pretraining data (from a wider range of sources) to create our LLMs. Crucially, there is no overlap between the pretraining data and the reference sources.

4.1 Reference Data Sets

lge	pretraining	benchmark
en	BNC, Switchboard CHILDES	Buckeye CANDOR
fr	CHILDES, ORFÉO OpenSubtitles	SUMM-RE
tw-zh	CHILDES, ASBC, NCCU, CallFriend, Parliament meetings, TV show	MCDC

Table 1: Spoken data in the pretraining data and the benchmarks. tw-zh correspond to Taiwan Mandarin.

The reference data for the benchmarks come from linguistic conversational corpora, as listed in Table 1. For English, we used two sources to build two different components of the benchmark. For the distribution of fillers, we used the Buckeye Corpus[3] (Pitt et al., 2005), which contains 38.1 hours of spontaneous speech (40 speakers) recorded in an interview format. For the task involving discourse markers, since Buckeye's interview format is less dynamic compared to true conversation, we used the CANDOR corpus (Reece et al., 2023), a 7M-token multimodal dataset of naturalistic conversation. For French, we used SUMM-RE (Hunter et al., 2024), a 1M-token meeting corpus, with about 20% of the corpus provided with manually checked fine-grained transcriptions. For Mandarin, we used the Sinica Mandarin Conversational Dialogue Corpus (Sinica MCDC8)[4] (Tseng, 2013).

The main reason for choosing these corpora is the high quality of their transcriptions, in which disfluencies are accurately rendered.

4.2 Tasks

Discourse Markers Task Organizing the typology of discourse and interactional lexical devices is a delicate issue (Crible and Degand, 2019). However, the markers we selected are among the most frequent items and are generally uncontroversial. More precisely, we work with two types: *semantic connectives* vs. *attitudinal markers*. See Table 2 for the complete list. These lists were established by examining the distributions of first tokens[5] of all utterances in the benchmarks, and selecting the most frequent discourse markers, excluding *"and"*, which has been considered problematic due to its systematic ambiguity between discourse-level and clause-internal usage.

To create minimal pairs involving discourse markers, we identified utterances from our reference corpora that were initiated by one of the semantic or attitudinal discourse markers. The utterance and its immediate successor were used as the *positive example* of the minimal pair. Notably, in addition to the type of discourse marker involved, the minimal pairs are categorized according to whether the two utterances (delimited by a discourse marker) belong to the same or different speakers, that is, whether the discourse marker starts a new turn (*monologic vs. dialogic*).[6] Overall, this gave us four different tasks: *semantic-monologic, semantic-dialogic, attitudinal-monologic*, and *attitudinal-dialogic*.

The *negative example* of the minimal pair was created by replacing the discourse marker with another one from the same category. Thus, the task evaluates a model's ability to prefer the authentic discourse marker given an utterance pair. As a sanity check, we also created a set of minimal pairs where the *negative example* consists of a shuffled version of the entire sequence.

While creating the final tasks, we faced a trade-off between benchmark size and balance. Discourse marker distributions are prototypically Zipfian, making it difficult to maintain both diversity and balance. We optimized this trade-off for each category, aiming to keep the discourse marker distribution as uniform as possible within each task, as illustrated in Table 2.

Fillers Task To complement our discourse markers tasks, we created another set of tasks focusing on *fillers*. Our tasks rely on the following list of conventionalized filler transcriptions: *'euh'* for

[3] https://buckeyecorpus.osu.edu/

[4] https://www.aclclp.org.tw/use_mat.php#mcdc

[5] Utterance initial position is known to have a discourse flavor.

[6] As mentioned earlier, since Buckeye's transcripts mainly focus on the interviewee, we used the CANDOR corpus to build the minimal pairs.

	en	zh	fr
dial-att	like, oh, well (3x167)	像(6), 喔(31), 就是* (21)	ah, ben (2x250)
dial-sem	because, but, so (3x158)	但是(17)*, 因為(11), 所以(11), 然後(2), 而且(2)	donc, mais, alors, après (4x75)
mono-att	like, oh, well (3x167)	像(7), 喔(4), 就是* (21)	ah, ben, oh, enfin, bon (5x15)
mono-sem	but, because, so, then (4x125)	但是* (16), 因為(8), 所以(10), 然後(14), 而且(3)	donc, mais (2x100)

Table 2: Discourse Markers used in the experiments. In parentheses, the number of ocurrences for each DM. It does not always sum to 500 (our target) due to lack of reference data in some cases. See Table 5 for translations. (The counts for 就是 include the counts of variants such as '就' and '就是說', while the counts for '但是' include '可是' and '不過')

French (Pallaud et al., 2019), *'um'* and *'uh'* for English (Clark and Tree, 2002), and *'uhn'*, *'en/un'*, and *'nage'* for Mandarin (Tseng, 2013).

More precisely, each minimal pair consists of a genuine utterance featuring one of these fillers (*positive example*) and an altered version in which the filler has been randomly moved to another position within the same utterance (*negative example*). Evaluation of a model involve using them to assign a probability to each sequence. If the model assigns a higher probability to the (*positive example*), it scores on that minimal pair.

BLiMP Tasks The recent proposal of a multilingual BLiMP benchmark (Jumelet et al., 2025) (which does not include Mandarin) and the existence of Chinese BLiMP (CLiMP) (Xiang et al., 2021) allow us to complement our experiments on discourse and dialogue phenomena with more standard linguistic tasks related to grammaticality. While this is not the main focus of our work, we considered it an interesting addition to better understand both the differences between the tested models and the nature of our own tasks. For French and Mandarin, we used the aforementioned benchmarks, while for English, we used the filtered and supplemented version of BLiMP (Warstadt et al., 2020a) employed in the 2024 BabyLM Challenge (Choshen et al., 2024).

For each language–model combination, we performed stratified bootstrapping over 1000 iterations, with the number of samples (with replacement) matching the benchmark sizes across subcategories (67 for BLiMP and 16 for CLiMP) in order to obtain a distribution of scores.

4.3 Pre-training LLMs

For each language, we tested five models, including three RoBERTa models pretrained from scratch on three types of data: *conversational*, *written*, and a mixture of these two. Motivated by the BabyLM initiative (Warstadt et al., 2023a; Choshen et al., 2024), these models were trained on datasets with moderate sizes (10M words for English and French, and 5M characters for Mandarin). The complete figures about the pretraining data can be found in Appendix Table 4. We also included XLM-RoBERTa-Base[7] and XLM-RoBERTa-Large[8] (Conneau et al., 2020) in the experiments to serve as potential topline performers. To accelerate the experiments, we performed vocabulary pruning (Yang et al., 2022) of the RoBERTa models on the corresponding training data.

Another purpose of using RoBERTa models was to better contextualize our proposed metrics as a form of sanity check. The underlying idea is that if full-fledged LMs like RoBERTa fail to perform the task, it is likely that the task cannot be achieved given the provided data.

For pretraining English RoBERTa models, we drew on the training data provided in the 2024 BabyLM challenge (Choshen et al., 2024). As the training data itself was a mix of spoken (Switchboard (Godfrey et al., 1992), BNC (Consortium et al., 2007) dialogues, and CHILDES) and written data (Gutenberg and Simple Wikipedia), we directly took the 10M version for our *mixed* model pretraining. The Switchboard and BNC dialogues were used as the pretraining data for the *spoken* model, while the Simple Wikipedia part was used

[7] https://huggingface.co/FacebookAI/xlm-roberta-base

[8] https://huggingface.co/FacebookAI/xlm-roberta-large

Discourse Marker (attitudinal)	
Original	you should go to new york # **well** when they open # i mean things are so different now
Replaced DM	you should go to new york # **oh** when they open # i mean things are so different now
Shuffled	you they now different when to should new york mean **well** i things # go are open # so
Discourse Marker (semantic)	
Original	that's what i wanted to say # **but** i didn't say # i'm like # yeah # that's my dog
Replaced DM	that's what i wanted to say, **then** i didn't say # i'm like # yeah # that's my dog
Shuffled	i yeah # say i'm what that's **but** say # like # didn't # to dog wanted that's i my #
Filler	
Original	i i **uh** # i would find that hard to believe * # personally
Moved Filler	i i # i **uh** would find that hard to believe * # personally
Shuffled	to i believe i i find that hard * # would personally **uh** #

Table 3: Example of corpus example manipulation to generate the disfluency benchmark

to pretrain the *written* model.

For pretraining French models, we used a conversational dataset inspired by the original BABYLM data mix (ORFÉO[9] (Benzitoun et al., 2016), CHILDES-FR[10] (MacWhinney, 2014; Rose and MacWhinney, 2014)) and completed with OpenSubtitles (Lison and Tiedemann, 2016) and a 10M-token dataset from Wikipedia. 50% of the aforementioned two datasets were combined to pretrain the *mixed* model.

For Taiwan Mandarin, conversational data was derived from the NCCU Spoken Corpus of Taiwan Mandarin (Chui and Lai, 2008), the Taiwan Corpus of Child Mandarin (TCCM)[11] (Chang et al., 2011), the media interview sections of the Academia Sinica Balanced Corpus (Chen et al., 1996), and automatic transcripts from three sources: a subset of meeting recordings from Taiwan's Legislative Yuan[12], the CALLFRIEND Mandarin Chinese-Taiwan Dialect corpus (Canavan et al., 2020), and 116 episodes of 'PTS Theme Night SHOW,' a show featuring interviews and discussions[13]. The written data was from a subset of Traditional Chinese Wikipedia[14]. A similar mixed dataset was created to pretrain the third model.

To have more consistent formatting between conversational and written data, we inserted line breaks after all major punctuation marks (e.g., periods, exclamation marks, question marks), while also changing colons to commas. For the Mandarin data, we standardized the orthography of one of the common particles, *'eh'*.

Tokenization was performed with *SentencePiece* tokenizers (Kudo and Richardson, 2018) trained on the corresponding data with the Unigram model (Kudo, 2018). Vocabulary size was set to 10,000 with a minimum token frequency threshold of 2. Models were trained with hyperparameters similar to those used for languages with 'small' data (i.e., 5–10 MB) in the Goldfish LM project (Chang et al., 2024): 4 layers, 8 attention heads, and a hidden layer size of 512. Batch size was set to 64, with a learning rate of 5e-4 for English and French, and 1e-3 for Mandarin.

4.4 Evaluation

The evaluation of the models on the benchmarks follows the minimal-pair paradigm: Given a model and two sequences forming a minimal pair in the benchmark (one genuine, one "modified"), the model is considered correct if it assigns a higher probability to the positive sequence—defined in our benchmark as the naturally occurring sequence, and in BLiMP as the grammatical one. The mean log probability of each sequence was estimated using the `minicons` package (Misra, 2022) in Python. For English and French, within-word left-to-right masking was applied following (Kauf and Ivanova, 2023). A model's prediction is deemed correct when it assigns a higher probability to the positive example.

[9] https://hdl.handle.net/11403/cefc-orfeo
[10] https://phon.talkbank.org/access/French/
[11] https://lope.linguistics.ntu.edu.tw/tccm/
[12] https://www.parliamentarytv.org.tw/
[13] https://www.youtube.com/playlist?list=PLSzfF9jXmOMOrWt5brkz9FudYKJGbQuYG
[14] https://huggingface.co/datasets/zetavg/zh-tw-wikipedia/

5 Experiments

5.1 Hypotheses

The different categories of discourse markers, together with fillers and BLiMP tasks, allow us to form a complete set of hypotheses with regard to the models we test in this paper. Regarding the DMs, both their types (SEM-antic vs. ATT-itudinal) and their interactional context (MONO-logic vs. DIAL-ogic) correspond more or less closely to conversational spontaneous speech. More precisely, SEM DMs are important in all discourse genres, while ATT DMs are more present in genres in which the speaker tends to systematically signal their attitude toward discourse elements. The latter corresponds to dialogical situations in which new information is regularly brought in by the interlocutor, and it is expected to position oneself with regard to those. MONO-logical contexts are the prototypical ones for written genres, even if present to varying degrees in conversational contexts. On the contrary, DIAL-ogical situations are rare in written documents while being the majority of cases in many conversations. In addition, *fillers* are frequent in spontaneous speech (conversational or not) but are absent from written documents, and even removed to some extent from a wide range of transcripts and other written dialogical productions (Prevot et al., 2019). Finally, BLiMP tasks are based on normative and canonical resources, not spontaneous speech.

Taken together, these observations led us to make the following hypotheses based on four families of models: `written` (Wikipedia), `conversational`, `mixed`, and `roberta`, whose overall properties are summarized in Table 4 in the appendix.

First, we expect to observe a clear advantage for `conversational` over `written` (and *fillers*) for the categories DIAL-DM and ATT-DM, and the strongest advantage when the two are combined. We also expect to see our small models perform relatively well compared to larger `roberta` models.

There might be an advantage for the other categories as well due to the reduced distance between training and testing data. However, we do not form strong hypotheses about them, particularly for the combination MONO+SEM, for which the `written` (and `roberta`) models might even have some advantage, since it is the canonical case in written-based training data.

Regarding BLiMP, we only expect RoBERTa to be much better than our small models and do not form particular hypotheses about differences between our small models.

5.2 Results: DMs and fillers

The main results of discourse marker replacement are shown in Figure 1. Across languages, the systematic result is that models trained with conversational data outperformed the models trained with Wikipedia data. The only exception is the task with semantic discourse markers in Mandarin dialogues. This exception is actually in line with another general pattern in the results, i.e., the difference between `conversational` and `written` models is not as large for semantic discourse markers (e.g., because, so) as for attitudinal ones *(e.g., that is, well, oh)*, which likely reflects the fact that the use of semantic discourse markers tends to be similar across written and conversational genres, both in frequency and in function. Another anomaly concerns the results of ROBERTA for French MONO-logical contexts, for which we do not have a clear explanation, since this is not the case for the other languages.

Regarding the `mixed` categories, they do not seem to be very different from the `conversational` category for French and English but appear to perform worse for the Mandarin version.

Similar results are observed for fillers, with a clear difference between `written` and `conversational` models, a more complex situation for the `mixed` model, and a slight benefit for ROBERTA, as shown in Figure 2.

5.3 Results: BLiMP tasks

As hypothesized and illustrated in the Appendix figure 5, the results show that RoBERTa models display apparent advantages in the typical BLiMP-style minimal pair tasks that target syntactic, semantic, and morphological phenomena. Among the smaller models, the differences are more contrasted across languages, without a clear winner between `written`, `conversational`, and `mixed`.

Interestingly, when we ran the supplementary English BLiMP task (Warstadt et al., 2023b), which targets a wider variety of phenomena, we observed, in Figure 3, trends showing the conversational model's advantage in the following tasks: QA congruence *(Easy: What did you get? I got a chair/*teacher'; Tricky: Who studies? David/*Math studies.')*, turn-taking *(David: Should*

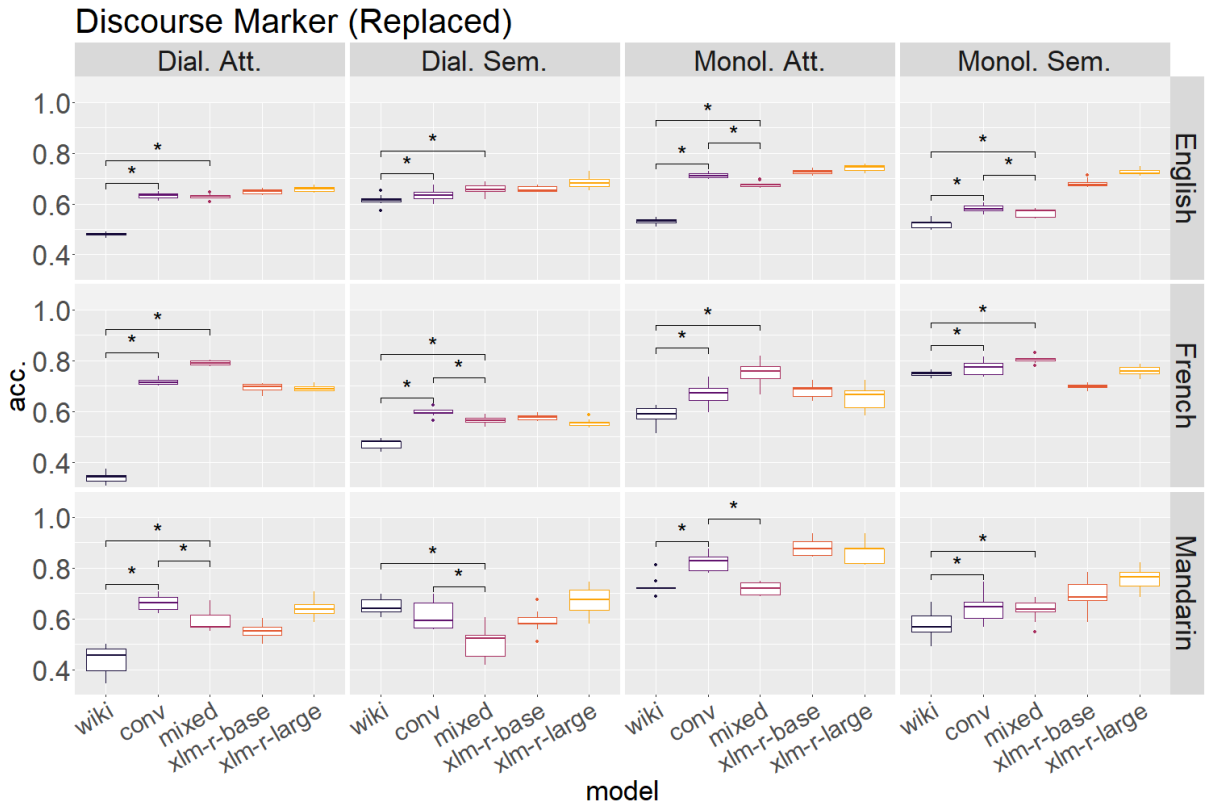

Figure 1: Overall Accuracy results on replacing discourse markers. The top and bottom of a box indicate the range between the first and the third quartiles, while the line in the box indicates the median. The whiskers (error bars) extend to the most extreme data point within 1.5 times the Interquartile Range (IQR) from the edges of the box. Model comparisons are based on Bayesian regression analyses where MODEL is the fixed predictor and SET is a random intercept. Dial. vs. Monol. = Dialogic vs. Monologic. Att. vs. Sem. = Attitudinal vs. Semantic. The models were run with weak (uniform) priors using the *brms* package in R, with post hoc testing comparing three small models (Wiki, conversational, mixed). Stars in the figure indicate a one-sided hypothesis with a posterior probability above 95%.

*you/*she quit? Sarah: No, I shouldn't')*, and subject-auxiliary conversion *(Is the novel he is putting away from the library' vs. '*Is the novel he putting away is from the library?')*, as well as the so-called *turn-taking* task, which concerns pronoun use coherence across turns. These tasks test semantic and anaphoric congruence in a dialogue setting (albeit very simple), suggesting a direct benefit of conversational data pretraining in tasks involving speaker interactions. The *hypernym* task from the supplement is a purely semantic task for which there was no reason to expect a benefit from more conversational data pretraining. It is actually surprising to observe that RoBERTa models do not perform better on that one.

6 Discussion and Conclusion

From a machine learning perspective, it might seem trivial that models trained on data similar to test sets perform better than models trained on other types of data. First of all, it is worth emphasizing that the pretraining datasets and benchmarks in our experiments are completely independent, as they do not come from the same raw corpora. Also, the pretraining datasets and the corpora used to build benchmarks have been curated by different teams and transcribed with different conventions. Nevertheless, we cannot deny that the conversational datasets are, in all aspects (sentence length distribution, lexical frequencies, etc.), more similar to the benchmarks than the Wikipedia datasets are. As trivial as it may seem, this is in fact one of our main points: to produce models more closely related to spontaneous speech, one should use datasets made of spontaneous speech (and not generic textual/web content).

Through this set of experiments, we aim to demonstrate the value of the proposed approach and to generalize it to other conversational, and spontaneous speech phenomena. From a broader

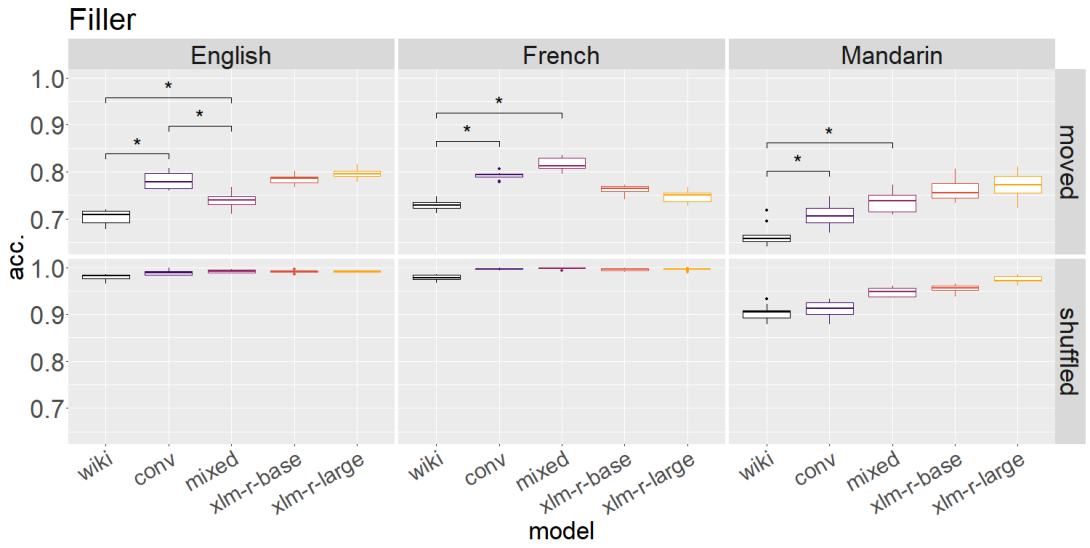

Figure 2: Overall results on fillers. Model comparison done with the same method as Figure 1.

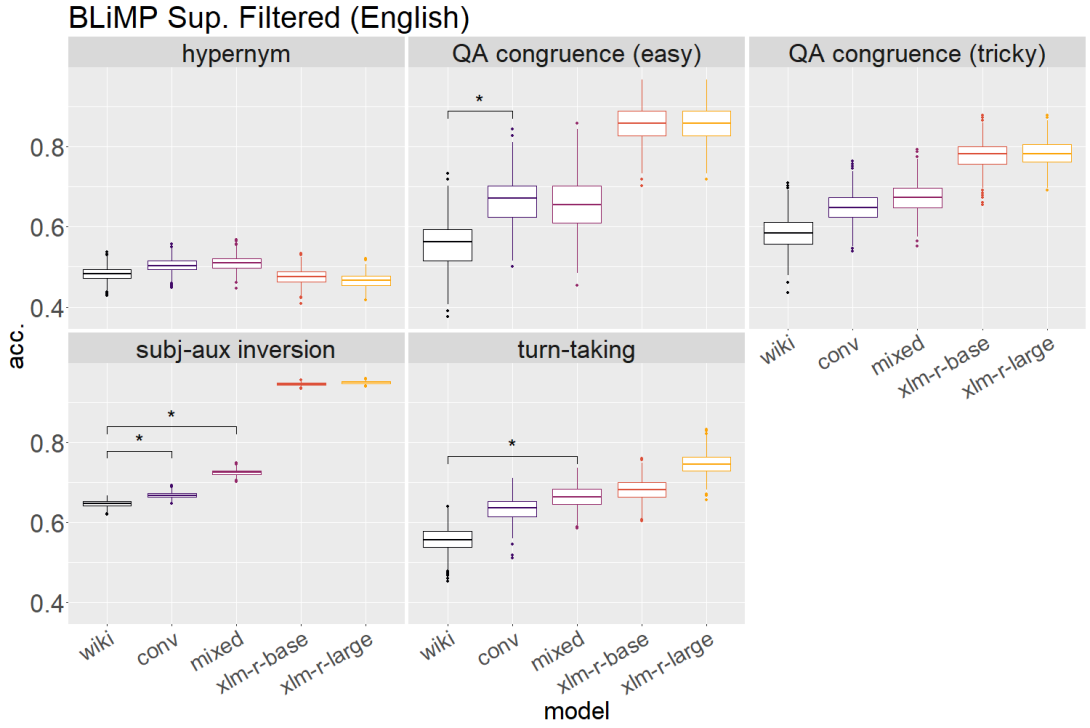

Figure 3: BLiMP supplement results

perspective, we hope to show that benchmarks from the rapidly expanding "BLiMP-family", which require significant amounts of expert and naive human input to build, can be complemented with benchmarks derived from numerous existing high-quality linguistic corpora, without additional human effort.

Acknowledgments

We would like to thank Shu-Chuan Tseng for discussions in relation to this paper. This study was supported by Taiwan's National Science and Technology Council (NSTC-112-2410-H-001-098-MY2), Institute of Linguistics, Academia Sinica (LING-114-DLR-01), ANR-Funded project SUMM-RE (ANR-20-CE23-0017) and from Institut Convergence ILCB (ANR-16-CONV-0002) managed by the ANR and A*MIDEX.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*, Toulon, France.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? *arXiv preprint arXiv:2503.11593*.

Francesco Cabiddu, Mitja Nikolaus, and Abdellah Fourtassi. 2025. Comparing children and large language models in word sense disambiguation: Insights and challenges. *Language Development Research*, 5(1).

Alexandra Canavan, George Zipperlen, and John Bartlett. 2020. CALLFRIEND Mandarin Chinese-Taiwan Dialect Second Edition LDC2020S06. Web Download.

Lih-Huei Chang, Zhong-Ru Chang, Yan-Chang Ke, and Su-Hui Xiao. 2011. Taiwan Child Language Corpus (TCCM). National Science Council Research Project, NSC96-2420-H-002-030. Available at http://lope.linguistics.ntu.edu.tw/tccm/.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Language, Information and Computation: Selected Papers from the 11th Pacific Asia Conference on Language, Information and Computation: 20-22 December 1996, Seoul*, pages 167–176. Waseda University.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. The 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Morten H Christiansen and Nick Chater. 2022. *The language game: How improvisation created language and changed the world*. Random House.

Kawai Chui and Huei-ling Lai. 2008. The nccu corpus of spoken chinese: Mandarin, hakka, and southern min. *Taiwan Journal of Linguistics*, 6(2).

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

BNC Consortium and 1 others. 2007. British national corpus. *Oxford Text Archive Core Collection*.

Ludivine Crible. 2018. Discourse markers and (dis) fluency: Forms and functions across languages and registers. In *Discourse Markers and (Dis) fluency*. John Benjamins Publishing Company.

Ludivine Crible and Liesbeth Degand. 2019. Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Ludivine Crible and Elena Pascual. 2020. Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics*, 156:54–67.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for NLP*, pages 134–139.

John J. Godfrey, Edward C. Holliman, and Jane M. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development.

Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, and Laurent Prévot. 2024. MEETING: A corpus of French meeting-style conversations. In *Actes de JEP-TALN-RECITAL 2024. 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position*, pages 508–529. ATALA & AFPC.

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

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Stephen C Levinson. 2020. On the human "interaction engine". In *Roots of human sociality*, pages 39–69. Routledge.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Daniel C O'Connell and Sabine Kowal. 2004. The history of research on the filled pause as evidence of the written language bias in linguistics (linell, 1982). *Journal of Psycholinguistic Research*, 33:459–474.

Berthille Pallaud, Roxane Bertrand, Laurent Prevot, Philippe Blache, and Stéphane Rauzy. 2019. Suspensive and disfluent self interruptions in French language interactions. *Fluency and Disfluency across Languages and Language Varieties. Corpora and Language in Use–Proceedings*, 4:109–138.

Ludovica Pannitto, Aurelie Herbelot, and 1 others. 2020. Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176. Association for Computational Linguistics.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Laurent Prevot, Pierre Magistry, and Pierre Lison. 2019. Should we use movie subtitles to study linguistic patterns of conversational speech? a study based on French, English and Taiwan Mandarin. In *Third International Symposium on Linguitic Patters of Spontaneous Speech*.

Laurent Prévot, Sheng-Fu Wang, Jou-An Chi, and Shu-Kai Hsieh. 2024. Extending the BabyLM initiative : Promoting diversity in datasets and metrics through high-quality linguistic corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 147–158, Miami, FL, USA. Association for Computational Linguistics.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Yvan Rose and Brian MacWhinney. 2014. The phonbank project: Data and software-assisted methods for the study of phonology and phonological development. In *The Oxford handbook of corpus phonology*, pages 380–401.

Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. How useful is context, actually? comparing llms and humans on discourse marker prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Frank Schilder. 2002. Robust discourse parsing via discourse markers, topicality and position. *Natural Language Engineering*, 8(2-3):235–255.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–35.

Alice Suozzi, Luca Capone, Gianluca E. Lebani, and Alessandro Lenci. 2025. Bambi: Developing baby language models for italian. *Preprint*, arXiv:2503.09481.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, and Ekaterina Artemova. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*.

Shu-Chuan Tseng. 2013. Lexical coverage in Taiwan Mandarin conversation. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*.

Sheng-Fu Wang, Laurent Prevot, Jou-An Chi, Ri-Sheng Huang, and Shu-Kai Hsieh. 2025. Spontaneous speech variables for evaluating LLMs cognitive plausibility. *arXiv preprint arXiv:2505.16277. Presented at CMCL 2025.*

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and 1 others. 2023b. 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. 2019a. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Ziqing Yang, Yiming Cui, and Zhigang Chen. 2022. TextPruner: A model pruning toolkit for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 35–43, Dublin, Ireland. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. 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, Online. Association for Computational Linguistics.

Syed Zohaib Hassan, Pierre Lison, and Pål Halvorsen. 2024. Enhancing naturalness in LLM-generated utterances through disfluency insertion. *arXiv e-prints*, pages arXiv–2412.

A Characteristics of Pretraining Data

Table 4 provides statistics on pretraining data, including the number of utterances, words, and unique words. Figure 4 presents the utterance length distributions for the training splits.

Table 4: The characteristics of pretraining data. For Taiwan Mandarin (tw-zh), each character is counted as one "word". The "segments" column counts the separated by punctuations (in particular, `.,?!` and their Chinese counterpart)

Lang	Split	Source	Utterance Count		Word Statistics			Unique words
			Lines	Segments	Total words	Avg words per line	Avg words per segment	
en	train	wiki	737,776	1,214,526	8,910,644	12.08	7.34	212,758
		conv	981,747	1,734,651	9,531,387	9.71	5.49	67,713
		mixed	1,253,464	2,232,350	9,349,902	7.46	4.19	137,325
	dev	wiki	69,726	116,903	860,230	12.34	7.36	53,219
		conv	105,261	183,273	917,598	8.72	5.01	21,503
		mixed	128,562	252,385	903,845	7.03	3.58	29,725
fr	train	wiki	607,959	1,244,188	8,972,811	14.76	7.21	250,722
		conv	1,708,359	1,884,569	9,143,990	5.35	4.85	84,821
		mixed	1,278,525	1,597,684	9,047,415	7.08	5.66	189,327
	dev	wiki	64,799	143,112	1,152,831	17.79	8.06	69,227
		conv	239,965	260,580	979,949	4.08	3.76	12,658
		mixed	179,018	215,966	1,080,586	6.04	5.00	43,493
tw-zh	train	wiki	210,376	424,018	4,494,896	21.37	10.60	46,826
		conv	570,457	599,189	4,644,140	8.14	7.75	19,364
		mixed	392,829	513,037	4,594,104	11.69	8.95	41,614
	dev	wiki	22,673	46,690	479,870	21.16	10.28	13,860
		conv	64,668	66,593	506,526	7.83	7.61	4,973
		mixed	44,074	56,886	490,082	11.12	8.62	10,208

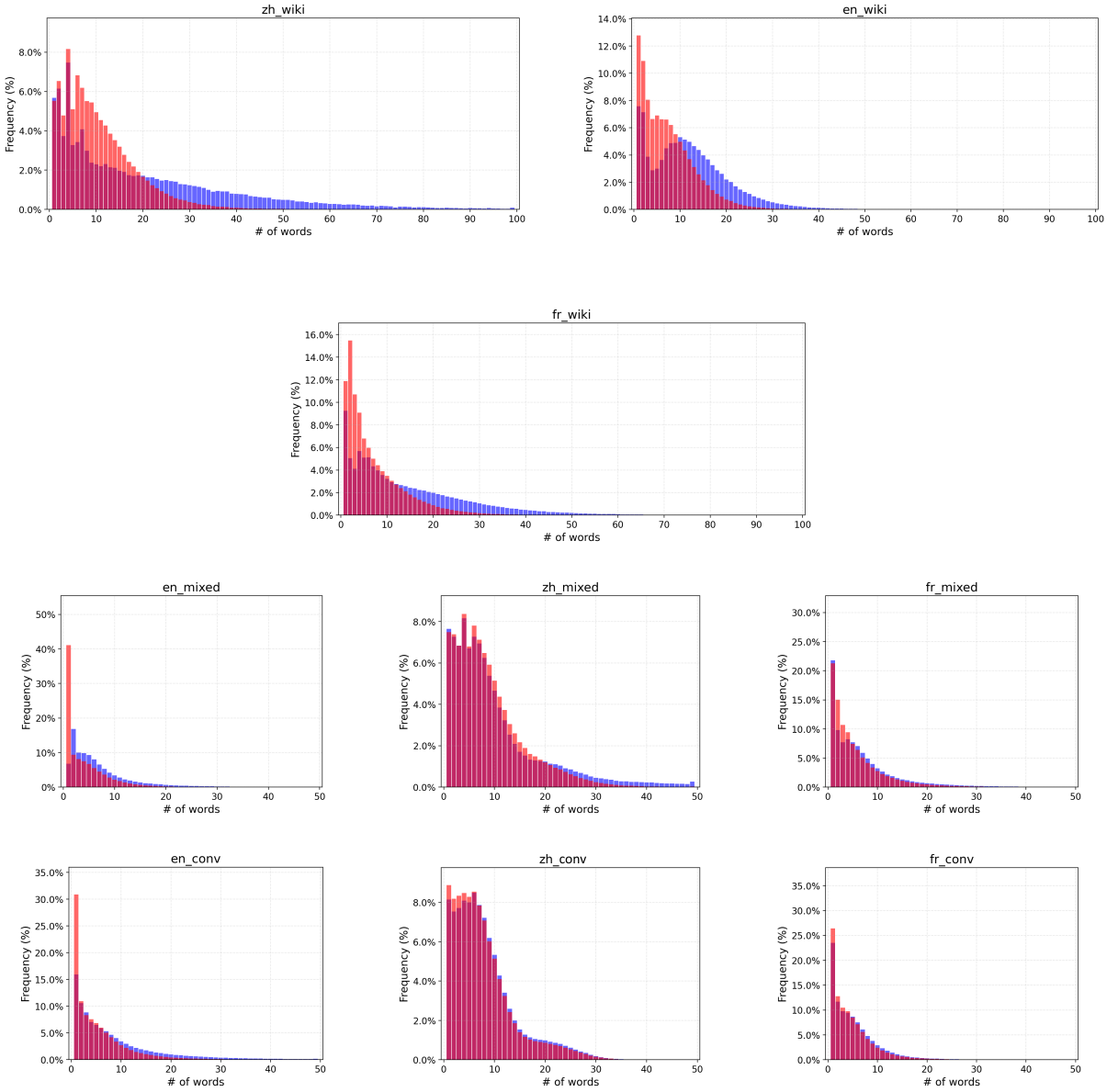

Figure 4: Utterance length distributions for the training splits. Blue bars show distributions based on line-separated utterances; red bars show distributions based on punctuation-separated segments.

B Markers used in our experiments

Table 5: Translation of the markers used in our experiments. In **bold** markers actually used. The other ones are only provided for exhaustivity.

english	mandarin	french	type
but	**dan4-shi4** 但是**, ke3-shi4** 可是**, bu2-guo4** 不過	**mais**	sem
because	**yin1-wei4** 因為	parce que	sem
then	**ran2-hou4** 然後	**après, alors**	sem
and	**er2-qie3** 而且	et	sem
so	**suo3-yi3** 所以	**donc**	sem
oh		ah, oh	att
well		**bon, ben**	att
well		**enfin**	att
like	**xiang4**	genre	att
that-is (to say)	**jiu4** 就**, jiu4-shi4** 就是**, jiu4-shi4-shuo1** 就是說		att
um, uh	**uhn, en, un** 嗯**, nage** 那個	**euh**	filler

C Additional results

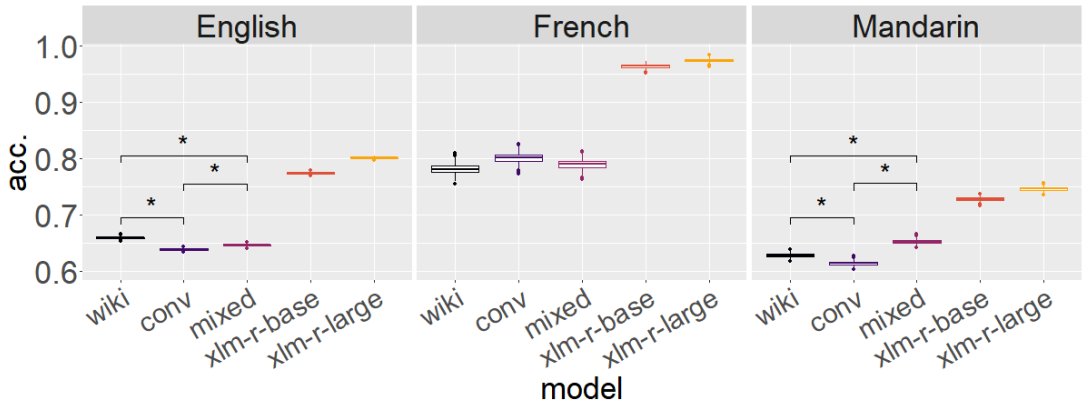

Figure 5: BLiMP results across languages.

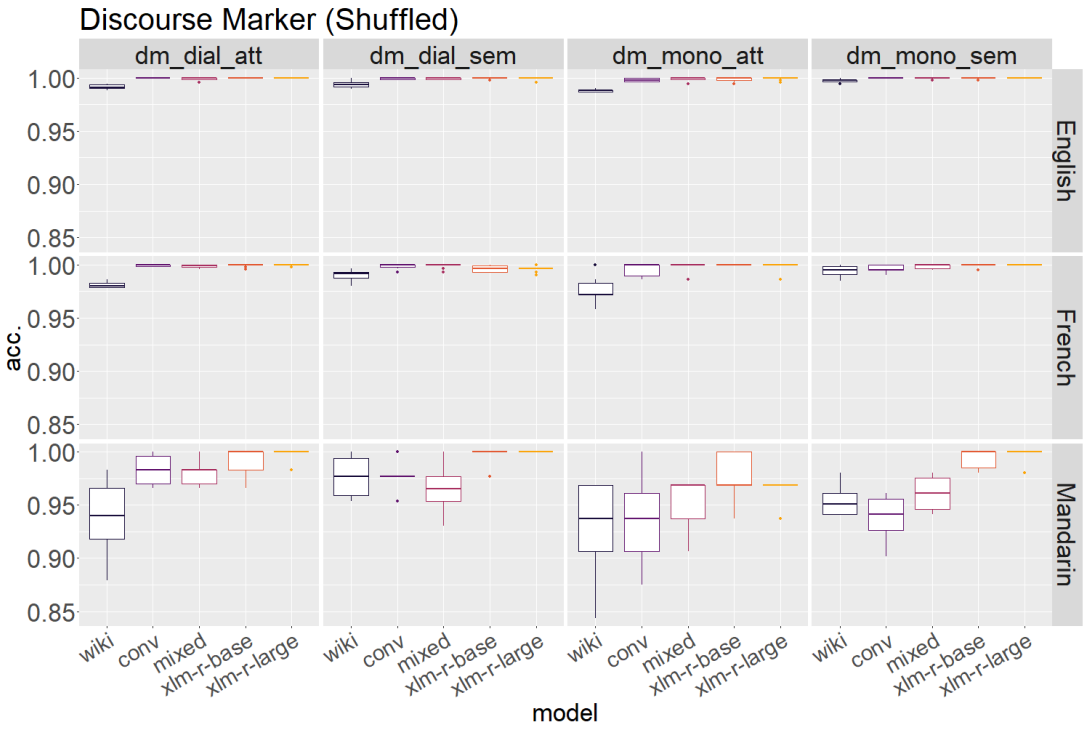

Figure 6: Overall results on the DM-shuffling baseline task.