Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures

Julius Steuer   **Marius Mosbach**   **Dietrich Klakow**
Department of Language Science and Technology
Saarland University
{jsteuer,mmosbach,dietrich.klakow}@lsv.uni-saarland.de

Abstract

Research on the cognitive plausibility of language models (LMs) has so far mostly concentrated on modelling psycholinguistic response variables such as reading times, gaze durations and N400/P600 EEG signals, while mostly leaving out the dimension of what Mahowald et al. (2023) described as formal and functional linguistic competence, and developmental plausibility. We address this gap by training a series of GPT-like language models of different sizes on the strict version of the BabyLM pretraining corpus, evaluating on the challenge tasks (BLiMP, GLUE, MSGS) and an additional reading time prediction task. We find a positive correlation between LM size and performance on all three challenge tasks, with different preferences for model width and depth in each of the tasks. In contrast, a negative correlation was found between LM size and reading time fit of linear mixed-effects models using LM surprisal as a predictor, with the second-smallest LM achieving the largest log-likelihood reduction over a baseline model without surprisal. This suggests that modelling processing effort *and* linguistic competence may require an approach different from training GPT-like LMs on a developmentally plausible corpus.

1 Introduction

In recent years several approaches have been taken to test LMs for cognitive plausibility. This is usually done by using output probabilities of the LM as a predictor for a model's preference towards certain linguistic structures (Roark et al., 2009; Wilcox et al., 2020). Another strain of research uses the output probabilities as a correlate of psycholinguistic measures, e.g., N400 and P600 EEG signals (Heilbron et al., 2019 and recently Li and Futrell, 2023) and (self-paced) reading times (Fernandez Monsalve et al., 2012). A natural question that arises is whether cognitive plausibility should be attributed to the model architecture itself, or to the training regime in combination with the training
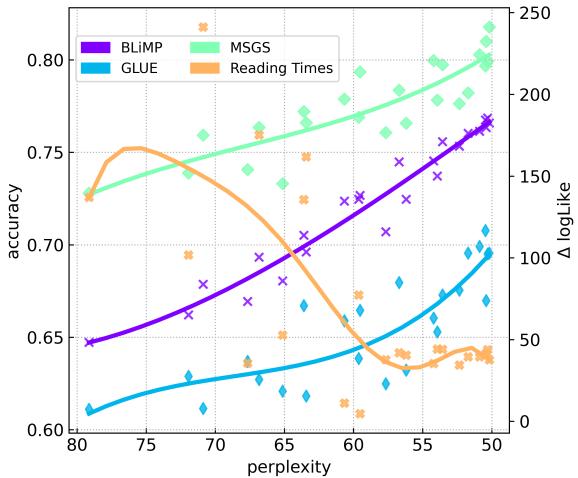

Figure 1: Our results show that LM performance on the BabyLM challenge tasks is negatively correlated with perplexity on the development set of the BabyLM corpus (lower perplexity leads to higher performance). In contrast, a *positive* correlation (Spearman's $\rho = 0.4784$, $p < 0.05$) was found between LM perplexity and the fit of LM surprisal to self-paced reading times from the Natural Stories corpus (Futrell et al., 2021) in terms of the difference in log-likelihood between a basline linear mixed-effects model and a model using LM surprisal as a predictor. Lines were fitted with 3 (challenge tasks) or 6 (reading times) degrees of freedom to the LMs' average performance on the task. See Section 6 for detailed results.

dataset. Little research has been done on the actual neurological plausibility of large LMs (LLMs), but Schrimpf et al. (2021) showed that the architecture of BERT-like models is already plausible for the next word prediction task before training: model predictions with only the language modelling head trained are already predictive of human brain activity during reading *and* correlate well with the predictions of the fully trained model. In contrast, no correlation between brain activity and model predictions was found for models trained on GLUE (Wang et al., 2019), a natural language understand-

ing (NLU) benchmark. This finding may mirror an underlying difference in language processing between *formal* and *functional linguistic competence* as introduced by Mahowald et al. (2023):

Formal linguistic competence is defined as the "capacity required to produce and comprehend a given language, i.e., the ability to distinguish grammatically correct from incorrect formations, based either on "knowledge of and flexible use of linguistic rules" or "non-rule-like statistical regularities" (Mahowald et al., 2023). An example for the former mechanism would be the regular formation of past tense verbs in English (*look:looked*), and for the latter the formation of irregular or ablauting past tense verbs (*go:went,tread:trod*).

Functional linguistic competence is defined as "non-language-specific cognitive functions that are required when we use language in real-world circumstances" (Mahowald et al., 2023) , i.e., the ability to perform cognitive tasks *with* language. GLUE is an example for a benchmark that test this dimension of linguistic competence, with some if its tasks (CoLA (Warstadt et al., 2019)) also testing for aspects of *formal linguistic competence*.

The dichotomy between formal and functional linguistic competence can be understood in terms of Wittgenstein's definition of the meaning of a word as its use in a language (Wittgenstein (1953), §43). The debate on whether statistical learners (i.e. LMs) can learn the meaning of a linguistic unit (word, phrase, text, etc.) in Wittgenstein's sense is still ongoing, with much division between positions that strongly deny that LMs can have such a property (Bender and Koller, 2020) and positions that advocate that they might have it, e.g., under the condition that the LM's predictions are grounded in extralinguistic reality (Bisk et al., 2020). Our study does not attempt to find arguments in favour of either position, but to study the implications of this dichotomy for the paradigm of cognitive modelling.

As stated earlier, the output probabilities of LMs lie at the basis of the application of LMs to cognitive language modelling, usually in the form of a probability distribution over a vocabulary of word forms given either surrounding words (masked language modelling) or preceding words (causal language modelling). Evidence for the use of surprisal (a word's negative logarithmic probability in con-

text) instead of the actual probablity comes from logarithmic effects of contextual probabilities on processing difficulty (Shain et al., 2022). Another approach is to evaluate the output probabilities of a LM over a number of classes that may or may not apply to the input sequence, usually after fine-tuning the LM. The reliance of research in this direction on the output probabilities of LMs has already been criticized from multiple sides. There is a growing body of evidence that the performance of a LM in the typical language modelling task, next word prediction, and measures of formal linguistic competence are not correlated. Hu et al. (2020) found no correlation between LM perplexity and measures of formal linguistic competence, while Huang et al. (2023) argue that LM surprisal should not be assumed to be a good predictor of psycholinguistic measures of processing difficulty that require more than just lexical information. This lack of correlation with psycholinguistic measures becomes more prominent with the increasing size of LMs (Oh and Schuler, 2022), and especially so in extreme cases of human processing difficulty: Arehalli et al. (2022) showed that surprisal from LSTM-based LMs underestimates garden-path effects on reading times, while successfully predicting reading times for most non-garden-path sentences. This finding has been corroborated for transformer-based LMs such as GPT-2 (Jurayj et al., 2022) and BERT (Irwin et al., 2023).

2 BabyLM

The BabyLM challenge (Warstadt et al., 2023) introduces a novel constraint to cognitively plausible language modelling by limiting the token budget for LM pretraining to 100 million (100M) tokens, roughly the same amount of tokens a 13-year old child has seen during language acquisition (Gilkerson et al., 2017). While the focus of the challenge is on the pretraining procedure, the evaluation pipeline consists of the BLiMP (Warstadt et al., 2020a), MSGS (Warstadt et al., 2020b) and GLUE benchmarks, each of which aims to test for a specific dimension of linguistic competence.

BLiMP BLiMP tests for *formal linguistic competence* by comparing model predictions at a critical word in pairs of grammatically acceptable and unacceptable sentences, with the sentence pair only differing with respect to a single feature, e.g., whether a determiner agrees with its antecedent in gender or not. A model succeeds at the task if it assigns a

higher probability to the critical word in the acceptable sentence.

GLUE GLUE is a benchmark that requires fine-tuning[1] of the LM. It tests for a wide range of NLU problems, e.g., question answering, natural language inference and linguistic acceptability judgements, and hence can be regarded as a proxy for the *functional linguistic competence* of a LM.

MSGS MSGS is a benchmark of binary classification tasks that tests whether a LM prefers *surface generalizations* over *syntactic generalization* by first fine-tuning on data consistent with both types of generalization. At inference time, items are consistent with only one type, potentially revealing a bias towards either generalization type.

Previous studies mainly provided insights into the relationship of pretraining token budget and measures of formal and functional linguistic competence. Zhang et al. (2021) showed that encoder-only LMs already perform well on formal tasks such as BLiMP at a budget of 10-100M tokens, while requiring substantially larger token budgets to perform well on functional tasks such as GLUE. While this research established correlations for pretraining token budgets, similar relationships for *model size* at a fixed token budget have not yet been investigated. This study is dedicated to finding a relationship between model size and performance on these tasks, while simultaneously addressing the dimension of *processing effort*, which is not covered by the challenge tasks. This is done using the **strict** version of the BabyLM corpus, mainly because there is evidence that the fit with psycholinguistic measures profits from token budgets far larger than the 100M tokens in the corpus (Oh and Schuler, 2023).However, we also implicity evaluate on models that are trained on token budgets of 10M tokens, corresponding rather to the **strict-small** track in Section 7.

3 Research questions

The starting point of our work is Zhang et al. (2021)'s finding of an earlier saturation effect (in terms of pretraining tokens) for BLiMP as opposed to (Super)GLUE. If performance on BLiMP is already close to the optimum after pretraining for

100M tokens, we suspect that a model with relatively small capacity is sufficient to reliably learn the required syntactic and semantic features. In contrast, the larger pretraining token budget and model size needed for GLUE should also require a model with higher capacity.

Studies on reading time prediction generally use causal LMs trained on a next-word prediction task instead of masked LMs (Oh and Schuler, 2022; Arehalli et al., 2022; Jurayj et al., 2022) because of their closer similarity to human language processing. Although masked LMs such as BERT show some word order effects (Papadimitriou et al., 2022) and even garden-path effects (Irwin et al., 2023), they are cognitively implausible in the sense that they process all words in a sequence simultaneously when predicting a word at a masked position, rather than processing language sequentially. This *autoregressive* property mirrors human language processing, and is therefore desirable in studies with the primary goal of modelling human reading behaviour. We therefore employ decoder-only, GPT-like LMs (Radford et al., 2019) in our study, i.e., we want to answer the following research questions:

> **Research question A**
>
> Are GPT-like models cognitively plausible in the sense that they are able to acquire (a degree of) formal and functional linguistic competence, while being also predictive of human processing effort?

> **Research question B**
>
> Can such LMs be trained on the same data as a child has available during language acquisition (100M tokens)?

4 Previous work

Do we need transformers for cognitive plausibility? Despite promising findings by Hosseini et al. (2021), it has yet to be determined whether transformers, and decoder-only transformer LMs in particular, are cognitively plausible in the sense that they are data-efficient enough to acquire human-like[2] linguistic competence. Indeed, there are results that seem to partially contradict the necessity

---

[1]During fine-tuning, we train all parameters of the pre-trained LM as well as a randomly initialized classifier on top of the LM.

[2]Here, we do not use "human-like" to imply human-level performance, but rather that the model is *subject to similar processing constraints* as a human.

of LLMs with wide context windows in order for a model to exhibit human-like processing behaviour. Kuribayashi et al. (2022) showed that *reducing* context length of LLMs improves the fit of a linear mixed-effects model (LME) on gaze durations, with surprisal from a bigram GPT-2 model as a predictor yielding the largest log-likelihood reduction over the baseline model. Wilcox et al. (2020) failed to identify a relationship between psychometric predictive power (Δ log-likelihood) and syntactic generalization, concluding that different models are needed for modelling human processing effort versus syntactic generalization.

Linguistic competence vs. psycholinguistic measures It has long been clear that LM capacity, and subsequently LM perplexity, does not necessarily correlate with human-likeness (Kuribayashi et al., 2021). LLMs such as GPT-3 in particular were found to have considerable disadvantages when it comes to predicting psycholinguistic measures from their next-word predictions: Oh and Schuler (2022) found an inverse relationship between both perplexity and LLM capacity, versus fit to human reading times. The authors of this study hypothesize that this is because transformers have access to the full sequence context, and are trained on large enough corpora to make use of the information that they contain. This relationship between model perplexity and reading times is however not intrinsic to transformer-based LMs: Hu et al. (2020) found a similar relationship for LSTM LMs, though small GPT-like models have an advantage over recurrent models.

The impact of LM size on linguistic competence was investigated by Eldan and Li (2023), who found that relatively small GPT2-like models (<10M parameters) manage to produce fluent English and can be trained on relatively small corpora with a reduced vocabulary. Their study also shows that the relationship still holds for small models, while also identifying trade-offs between model width (hidden size) and depth (number of decoder layers).

As for training dataset size, Oh and Schuler (2023) found that surprisal from transformer-based LLMs gives the best fit to reading times at about 2B train tokens, across a wide range of model sizes. The corpus used in their study is very large (300B tokens), allowing for extensive training of a model without repeating any data. Reaching the same number of update steps with the much smaller BabyLM corpus would require training for multiple epochs.

Single- vs. multi-epoch training Since the BabyLM training data is substantially smaller than the 2B tokens suggested by Oh and Schuler (2023), training our models in a multi-epoch setting cannot be avoided. Previous research has shown that repeating the training data can have adverse effects: Xue et al. (2023) compared single-epoch vs. multi-epoch training in a limited data setting and show that multi-epoch training leads to overfitting, with little performance being gained after the first epoch. They also find that regularization can only partially alleviate the overfitting problem, with dropout having the largest effect. Not having to repeat the training data is advantageous for downstream tasks and psycholinguistic modelling, if a certain amount of training data is available: Oh and Schuler (2023) found that reading time fit deteriorates after 2B tokens over a wide range of model sizes. However, it is not clear if repeating the training data would lead to an even stronger deterioration. If the corpus is substantially smaller than 2B tokens, repeating the training data could have a different effect, especially if the optimum of the reading time fit depends on the availability of the 2B tokens.

5 Methodology

Modelling We use the OPT architecture by Zhang et al. (2022) with a language modelling head for pretraining. Following our intuition that BLiMP should require much smaller model sizes than MSGS and GLUE, we train a series of OPT models of different sizes, varying only model width (hidden size) and model depth (number of decoder layers). In total we train 24 models varying over 4 hidden sizes $l_{hidden} \in \{192, 384, 768, 1536\}$ and 6 numbers of decoder layers ($l_{decoder} \in \{1, 2, 4, 8, 16, 24\}$). We also adjust the dimension of the feedforward layers such that the size of the output vector $l_{forward} = 3 \times l_{hidden}$. Table 1 in Appendix A shows the resulting model sizes. The models and all code for pretraining are implemented with PyTorch (Paszke et al., 2019) and HuggingFace transformers (Wolf et al., 2020), starting from their implementation of OPT. We also trained a new tokenizer on the training set of the BabyLM corpus, using the same vocabulary size $|V| = 50272$ as the original OPT tokenizer. We report all results as averages over 3 random seeds (see Appendix D for full results and standard error).
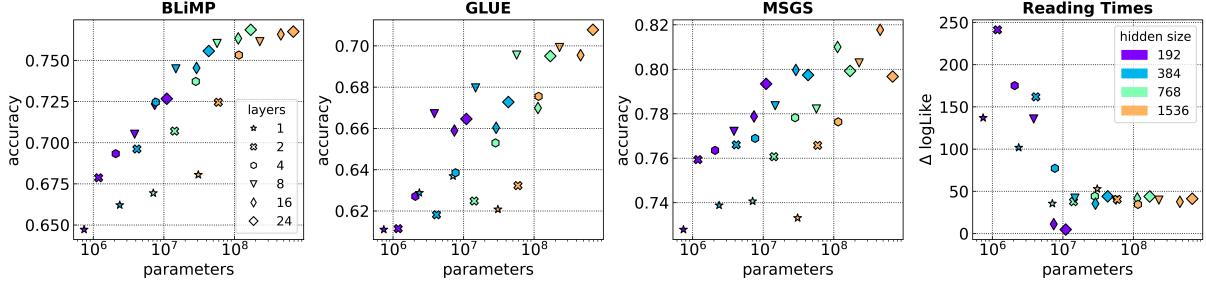
Figure 2: Task performance by model size (higher numbers are better). Baselines can be found in Appendix D.

Training Following the Shortformer pipeline (Press et al., 2021), each model is trained for one epoch with an initial sequence length of 64, followed by 4 epochs with the full sequence length of 256. The full sequence length of 256 was chosen as a compromise between the relatively short test items in the challenge tasks (up to 128 tokens) In order to ensure that the model generalizes to longer sequences we use ALiBi (Press et al., 2022) instead of learned positional embeddings. This also ensures that our models generalize to the longer sequences in the Natural Stories corpus. We trained each model on a A100 GPU with 40 GB VRAM and an effective batch size of 128, using gradient accumulation for models that could not fit the full batch size. We used AdamW (Loshchilov and Hutter, 2019) as our optimizer with an initial learning rate of 0.001 and weight decay of 0.001 with 2000 linear warm-up steps. We use a dropout of 0.1 following the default HuggingFace transformers parameters for OPT.

Pretraining experiments We also experimented with changes to the pretraining regime. We trained models on multiple permutations of the training dataset: ordering sequences according to length (number of words), word length (number of characters), sequence-level perplexity from a 3-gram LM trained on the same data, and different orderings of the subcorpora as in Mueller and Linzen (2023). None of these approaches resulted in significant performance gains in terms of perplexity and performance on the challenge tasks over a baseline model trained on the concatenated BabyLM corpus with subsequent shuffling of the sequences.

Evaluation We evaluated all models on the downstream tasks of the BabyLM challenge. While these three tasks test for the linguistic competence of a model, they do not quantify the cognitive effort associated with language

processing. We therefore also evaluate all models on a reading time prediction task. For each model, we calculated surprisal on the items of the Natural Stories Corpus (Futrell et al., 2021). This corpus was chosen because its domain is close to at least one of the BabyLM subcorpora (Children's Stories). We fitted linear mixed-effects (LME) models with random intercepts for subject, word and item (the id of the story); surprisal, word frequency, word length and sentence position as predictors and log-normalized reading times as the response variable. The exact formula is

```
log(reading_time) ~
    word_surprisal + len(word)
  + log(word_frequency) + position
  + (1|word) + (1|subject) + (1|item)
```

For the reading time analysis we report the difference in log-likelihood between the models with surprisal as a predictor over a baseline model with only the control predictors. For all other tasks we report accuracy.

Code We used the evaluation code provided by the organizers of the BabyLM challenge[3], with some modifications to load custom models. The evaluation pipeline is based on the LM-Eval framework by Gao et al. (2021). Fine-tuning on GLUE and MSGS was done with the default hyperparameter settings, but we reduced the number of fine-tuning epochs to 3 as we did not observe any improvements after 3 epochs. The LME models were fitted using the lmerTest R library (Kuznetsova et al., 2017) via the pymer4 Python package (Jolly, 2018). The code to pretrain and evaluate all models is publicly available on GitHub[4]. The model with the highest BLiMP accuracy and detailed results for the LME models are made available at the same

---

[3] https://github.com/BabyLM/evaluation-pipeline
[4] https://github.com/uds-lsv/babylm

146

location, alongside instructions on how to run the training and evaluation pipelines.

6 Results

Fine-tuning GLUE   Fine-tuning on GLUE was overall very unstable and often failed to outperform the baseline. This was mainly due to the one-size-fits-all approach to the fine-tuning hyperparameters; we repeated several more fine-tuning runs with different hyperparameter settings on some of the GLUE tasks, and found that, e.g., RTE profited from a longer warm-up period (which is in line with the findings of Mosbach et al. (2021) for BERT-like models), but most other sub-tasks fine-tuned with the same hyperparameters showed a drop in performance. While we could have optimized hyperparameters for all sub-tasks, the main objective of the BabyLM challenge is to improve the pretraining part of the NLP pipeline. Thus, we decided to fine-tune with the default hyperparameters, only adjusting the number of epochs as we found that the fine-tuning runs already converged after a few epochs.

Model size   Figure 2 shows the relationship between model size and task performance: While GLUE (Spearman's $\rho = 0.7739$, $p < 1^{-4}$) and MSGS ($\rho = 0.7148$, $p < 1^{-4}$) performance scales with model size, BLiMP performance plateaus after reaching a model size of about 50M parameters ($\rho = 0.8835$, $p < 1^{-4}$). In contrast, reading time fit was negatively correlated with model size ($\rho = -51.39$, $p < 0.05$). All correlations are statistically significant with $p < 1^{-4}$. No single model performed best on all three challenge tasks, with large differences in the size of the best model. Figure 1 shows that similar correlations hold for model perplexity and task performance (BLiMP: $\rho = -0.9765$, $p < 1^{-4}$, GLUE: $\rho = -0.8287$, $p < 1^{-4}$, MSGS: $\rho = -0.8661$, $p < 1^{-4}$); negative correlations mean that lower perplexity leads to higher performance. We found strong positive correlations (pictured in Figure 7 in Appendix D) between performance on the challenge tasks (BLiMP and GLUE ($\rho = 0.8784$), BLiMP and MSGS ($\rho = 0.9182$) and GLUE and MSGS ($\rho = 0.815$) generally with $p < 1^{-4}$).

Model width vs. depth   While BLiMP performance was not found to be strongly correlated with either the number of decoder layers or hidden size, GLUE and MSGS showed some variability based on the number of layers. For GLUE the only configuration that showed a monotonic improvement in performance was a hidden size of 1536, with models with more decoder layers achieving higher accuracy in this setting. For MSGS we observed a drop in performance for the models with 24 decoder layers at the largest hidden sizes (384, 768). Overall, the effect of hidden size and number of layers was minor when compared to overall model size. In contrast, the best fit on the reading time data was achieved with the second smallest model with only 2 decoder layers and a hidden size of 192. Figure 3 illustrates this trend: for the challenge tasks, performance increases with the number of layers (though not monotonically), whereas Δ log-likelihood of the LME models decreases with the number of layers at $l_{hidden} = 192$ and, to a lesser extent, at $l_{hidden} = 384$, while deeper models with more decoder layers and larger hidden sizes perform considerably worse.

Possible confounds   The reading time analysis suffers from several potential confounding factors: Firstly, the domain of the training data differs considerably from the data in the Natural Stories corpus. While the training data also contains some longer texts (Wikipedia, Children's Stories), most of the corpora are more representative of spoken language (Open Subtitles, BNC Spoken, CHILDES). In addition, most sequences are relatively short, with a median sequence length of 8 in the Open Subtitles corpus, which accounts for >50% of the training data. This is considerably less than the median sequence length of 22 in the Natural Stories corpus. Another confounding factor might be the difference in exposure to language data of the model and that of the participants of the original reading time study. Futrell et al. (2021) do not provide demographic data of their participants, but since data collection was done via Amazon Mechanical Turk we can safely assume that the mean age of the participants was higher than 13, meaning that they were exposed to considerably more language data than the 100M tokens in the BabyLM corpus. Although a recent study by Oh and Schuler (2023) showed that reading time fit (in terms of Δ log-likelihood) from transformer models still profits from pretraining data multiple orders of magnitude larger than our corpus, with an optimum at 2B tokens, this is partially alleviated in this study by the multiple-epoch training regime, totalling about 500M tokens seen by each of our
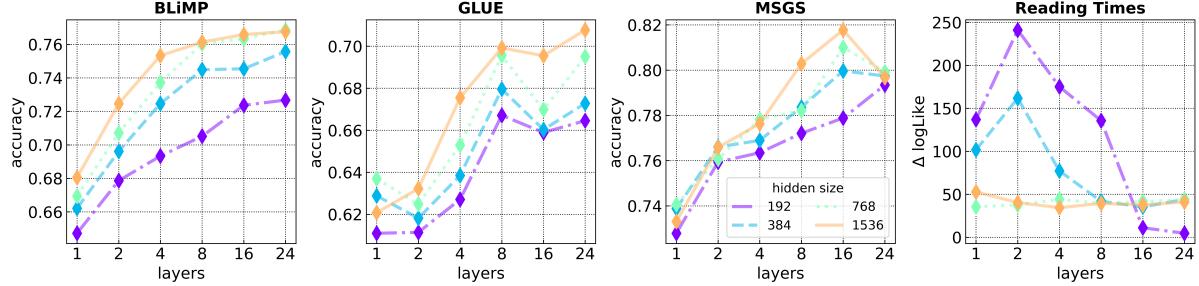
Figure 3: Task performance by hidden size, number of layers and task.
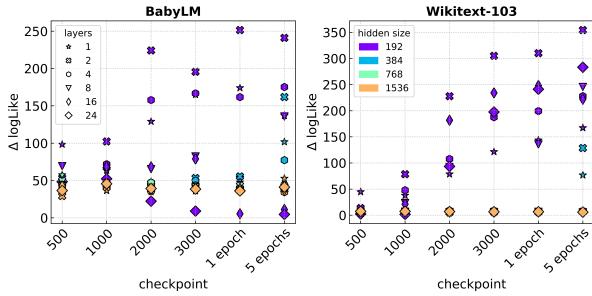
Figure 4: Reading time fit in terms of Δ log-likelihood over a base model without surprisal as a predictor, on the BabyLM and Wikitext-103 data after 500, 1000, 2000 and 3000 training steps (1/8, 1/4, 1/2 and 3/4 of an epoch) and 1 and 5 epochs.

models. Since Oh and Schuler (2023) found that training on more *unseen* tokens after reaching the optimum leads to a quick deterioration of reading time fit proportional to model size, it is unclear what impact repeating the training data would have on the reading time fit.

7 Reading time prediction in a multi-epoch setting

Experimental setup In order to evaluate whether the negative correlation is an artifact of the domain mismatch between the BabyLM corpus and the items in the Natural Stories corpus or the repetition of the training data before reaching the optimal token budget, we conduct two additional experiments: First, we retrain all models on the BabyLM corpus for a single epoch, saving intermediate checkpoints at 100, 500, 1000, 2000 and 3000 training steps. Then, we use the intermediate models to fit LME models to the reading time data, using the same formula as given in Section 5. Second, we replicate these experiments on Wikitext-103, a corpus of similar size that does not have the same limitations of the BabyLM corpus (i.e. an average sequence length and a domain closer to the Natural Stories

corpus). The models trained on Wikitext-103 serve as a control for the experiments on the BabyLM corpus and were not included in the final submission. Since the results indicate that larger models yield a worse reading time fit, we restrict the experiment to small models (1-4 layers, all hidden sizes) and larger models with the smallest and largest hidden size (192 and 1536). The models are trained with the same hyperparameter settings as the original models, but sequence length is not reduced in the first epoch.

Results Figure 4 shows a somewhat different picture for the models trained on Wikitext-103, with reading time fit of smaller models increasing over the whole pretraining procedure, while models with $l_{hidden} > 192$ almost never improve over the baseline model. In contrast, the reading time fit of the LMs trained on the BabyLM data improves significantly over the baseline for shallower models ($<$ 2 decoder layers), while staying roughly constant for deeper and wider models (16, 24 decoder layers). However, the relationship between the number of training steps and reading time fit is not monotonic, with a slight decrease after training for 4 more epochs for the best model. While the models trained on the Wikitext-103 dataset yield a better fit to reading times in terms of Δ log-likelihood, the basic finding on the BabyLM data is corroborated: exposing a transformer model to multiple repetitions of the training data before reaching the optimal token budget does not lead to a decrease in reading time fit, but also does not improve over the single epoch setting in a meaningful way. The results also show that the improved reading time fit for $l_{hidden} = 192$ cannot be attributed to smaller model size alone, as the deepest model with that hidden size, 24*192 shows an improved fit over the baseline, while 1*384, a model with a comparable number of parameters, but a larger hidden size,

does not. In conclusion, we did not find a degradation of reading time fit when repeating the training data, with similar effects of LM size on reading time fit for Wikitext-103 and the BabyLM corpus (see Table 2 in Appendix C for Spearman's ρ's and p-values). We also found The BabyLM corpus to be advantageous for this task in the sense that – in contrast to Wikitext-103 – reading time fit from all models improved over the baseline LME model.

8   Discussion

Correlation between BLiMP, GLUE & MSGS
The experiments presented in Section 6 provide evidence for a correlation between LM performance on BLiMP, GLUE and MSGS tasks when pretraining on the BabyLM corpus. This correlation is in accordance with established effects of training dataset size (Zhang et al., 2021), and interactions of train corpus size and model capacity (Eldan and Li, 2023, Kaplan et al., 2020). However, no single model achieves the highest score on all three tasks: BLiMP shows diminishing returns for model sizes larger than 50M tokens, while the best model on MSGS (16*1536) is substantially smaller than the best model on GLUE (24*1536). This discrepancy between the best model on the BabyLM challenge tasks and on the reading times prediction task is illustrated by Figure 5. The correlation between BLiMP/MSGS and GLUE may be an artifact of the sub-optimal fine-tuning on GLUE, failing to outperform the baseline model. It cannot be ruled out that the results would change when determining the optimal hyperparameters for each sub-task individually. However, even if the correlation were an artifact of the pretraining data, the findings of a negative correlation between model size and reading time fit would still hold.

Cognitive plausibility of GPT-like models The best fit on self-paced reading times from the Natural Stories corpus was obtained with the second-smallest model, with models with $l_{hidden} > 192$ only slightly improving over the baseline. The second suite of experiments in Section 7 confirms that this is not solely caused by the multi-epoch training regime necessitated by the small token budget. The reason for the mismatch between measures of cognitive plausibility (reading times) and measures of formal (BLiMP, MSGS) and functional linguistic competence (GLUE) is rooted in the interaction of pretraining regime and model size: While it is feasible to train a model that performs com-
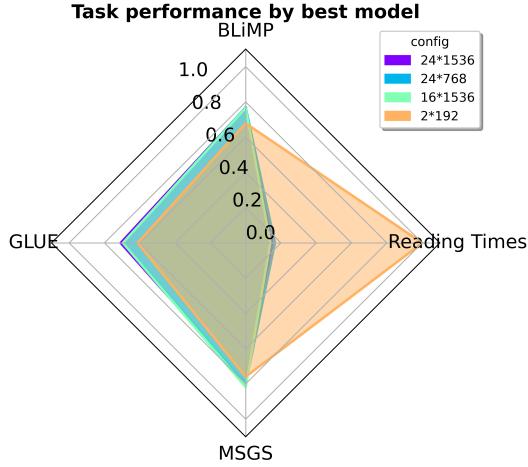
Figure 5: Performance of the best models by task. Reading times Δ log-likelihoods are normalized in the interval $[0, 1]$.

paratively well on all four tasks on a budget of 100M tokens, the sweet spot for model size and dataset size is reached much earlier for the reading time prediction task than for the BabyLM challenge tasks. This problem could easily be resolved by using one model when modelling reading times (or any other psycholinguistic measure), and another model when either of the forms of linguistic competence is the aim. This might be a valid and promising approach in a situation where the understanding of the research object does not depend on the connectedness of its experimental analoga. In the case of our research object – the human language faculty – it may not be necessary to find a single analogon that accounts for all its components, but since we *know* that the human language faculty is part of a unified cognitive system (with specialized sub-units) performing the tasks which the modern language modelling pipeline of pretraining and fine-tuning splits up into individual modules, it would be worthwile to move in the direction of a unified approach that accounts for both forms of linguistic competence and empirical evidence of processing effort. This could be achieved through adjustments to the pretraining regime (in terms of data, modelling objective etc.), as suggested by the BabyLM challenge, or through adjustments to the model architecture.

Size of transformer models The results of the reading time prediction study on the BabyLM corpus indicate that it in fact has an *advantage* over

Wikitext-103, although the LMs trained on the latter achieve larger Δ log-likelihoods on average: Since the largest models fail to improve over the baseline model if trained on Wikitext-103, it is possible that some properties of the language in the BabyLM corpus facilitate the learning mechanism that actuates the correlation of LM surprisal and reading times. The reason for the worse fit of surprisal from the larger models may be that both Wikitext-103 and the BabyLM corpus are not large enough to induce the learning bias needed to give good predictions of reading times in larger models, with Figure 4 showing that the results on the BabyLM corpus are much less stable than on Wikitext-103 and the improvements over the baseline much less sharply linear. In summary, our results lead to the following answers to our research questions:

> ### Result: Research question A
>
> GPT-like LMs can be cognitively plausible and display formal and functional linguistic competence, although not both at the same time...

> ### Result: Research question B
>
> ...under the constraint of a developmentally plausible training dataset.

9 Conclusion

Our study highlights the challenges of training a LM that performs well on tasks requiring some degree of formal and functional linguistic competence as defined by Mahowald et al. (2023), while also being predictive of the psycholinguistic measure of reading times. We find that small, shallow models of less than 5M parameters yield the best fit to the psycholinguistic measure, while performance on BLiMP, GLUE and MSGS improves with increasing model size, although to a different degree for each of the tasks. This has implications for research on cognitively or developmentally plausible models of human language processing: in the case of a small, domain-specific training corpus it is not feasible to pretrain an LLM that displays formal linguistic competence and performs well on a reading time prediction tasks, a conclusion also drawn by Wingfield and Connell (2022). Consequently, research in this direction has concentrated on fine-

tuning pretrained LLMs on domain-specific data, e.g., Škrjanec et al. (2023). A promising approach to a unified architecture could be relegating special tasks (such as classifying a sequence as in GLUE) to adapters (Houlsby et al., 2019), sub-networks within a pretrained LM. This approach is common in multilingual language modelling (Pfeiffer et al., 2022; Alabi et al., 2022), where its success is partially attributed to its ability to separate general linguistic knowledge from language-specific information. A similar modelling decision may be necessary for cognitively plausible language models.

Limitations

The results of the paper mainly hold for decoder-only transformer LMs. While these LMs are closer to human language processing in the sense that they process language incrementally, this has some disadvantages for reading time predictions, since humans do not attribute equal importance to each word, skipping some words in the process, and typically integrate words from the left- *and* right-hand context of a fixated word. While the first point can be addressed by explicitly modelling skipping behaviour (Hahn and Keller, 2016), the second could require a solution closer to masked language models.

A second limitation is the focus on self-paced reading time as the psycholinguistic response variable. Since the setup of self-paced reading studies, with the participants observing a single word at a time, distorts the natural reading process, the measure itself may be not that cognitively plausible. This could be addressed by repeating the experiments on corpora from eye-tracking studies such as the Dundee corpus (Kennedy and Pynte, 2005). There is evidence that much larger models than those tested in the current study still improve the fit to total reading times in less restricted experimental settings (de Varda and Marelli, 2023). The latter study also shows that the fit to psycholinguistic measures varies over languages and writing

systems.

Another option is modelling brain activity patterns directly by predicting N400 and P600 EEG signals, which have the additional advantage of providing a means of decomposing LM surprisal without the proxy of linguistic structure, as shown by Li and Futrell (2023).

Ethics Statement

The authors foresee no ethical concerns about the work presented in the paper.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English?

Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare yyntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 85–95, Austin, Texas. Association for Computational Linguistics.

Micha Heilbron, Benedikt Ehinger, Peter Hagoort, and Floris de Lange. 2019. Tracking Naturalistic Linguistic Predictions with Deep Neural Language Models. In *2019 Conference on Cognitive Computational Neuroscience*, Berlin, Germany. Cognitive Computational Neuroscience.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural Language Models for Nineteenth-Century English. ArXiv:2105.11321 [cs].

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. Publisher: arXiv Version Number: 2.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2023. Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark. preprint, PsyArXiv.

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. BERT Shows Garden Path Effects. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232, Dubrovnik, Croatia. Association for Computational Linguistics.

Eshin Jolly. 2018. Pymer4: Connecting R and Python for Linear Mixed Modeling. *Journal of Open Source Software*, 3(31):862.

William Jurayj, William Rudman, and Carsten Eickhoff. 2022. Garden-Path Traversal in GPT-2. ArXiv:2205.12302 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361 [cs, stat].

Alan Kennedy and Joël Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context Limitations Make Neural Language Models More Human-Like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower Perplexity is Not Always Human-Like. 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 5203–5217, Online. Association for Computational Linguistics.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Jiaxuan Li and Richard Futrell. 2023. A decomposition of surprisal tracks the N400 and P600 brain potentials. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Aaron Mueller and Tal Linzen. 2023. How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases. Publisher: arXiv Version Number: 1.

Byung-Doh Oh and William Schuler. 2022. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? Publisher: arXiv Version Number: 1.

Byung-Doh Oh and William Schuler. 2023. Transformer-Based LM Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens. ArXiv:2304.11389 [cs].

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. ArXiv:2203.06204 [cs].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Shortformer: Better Language Modeling using Shorter Inputs. 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 5493–5505, Online. Association for Computational Linguistics.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. ArXiv:2108.12409 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic

modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time. preprint, PsyArXiv.

Iza Škrjanec, Frederik Y. Broy, and Vera Demberg. 2023. Expert-adapted language models improve the fit to reading times.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. ArXiv:1804.07461 [cs].

Alex Warstadt, Aaron Mueller, Leshem Chohen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

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. 2019. Neural network acceptability judgments.

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.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. Publisher: arXiv Version Number: 1.

Cai Wingfield and Louise Connell. 2022. Understanding the role of linguistic distributional knowledge in cognition. volume 37, pages 1220–1270. Routledge.

Ludwig Wittgenstein. 1953. *Philosophische Untersuchungen*. Suhrkamp Verlag, Frankfurt am Main.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. 2023. To Repeat or Not To Repeat: Insights from Scaling LLM under Token-Crisis.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pretrained Transformer Language Models. Publisher: arXiv Version Number: 4.

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.

A OPT models

$l_{decoder}$	l_{hidden}	Parameters (non-embedding)
1	192	0.74
2	192	1.19
4	192	2.07
8	192	3.85
16	192	7.41
24	192	10.9
1	384	2.37
2	384	4.14
4	384	7.69
8	384	14.79
16	384	28.99
24	384	43.18
1	768	7.09
2	768	14.18
4	768	28.35
8	768	56.70
16	768	113.41
24	768	170.11
1	1536	30.69
2	1536	59.00
4	1536	115.69
8	1536	229.01
16	1536	455.67
24	1536	682.32

Table 1: OPT models sizes in million parameters by hidden size and number of decoder layers. The number of parameters does not include the embedding table, which is always of the size $l_{emb} \times |V| = 768 \times 50272 = 38.608.896$, as in OPT-128m.

B Validation perplexity
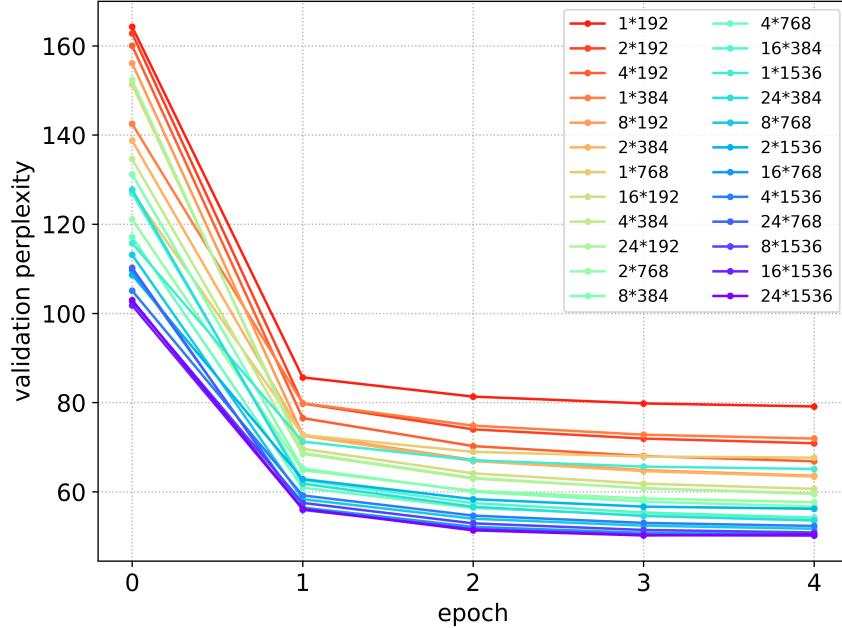
Figure 6: Validation perplexity by configuration and epoch on the development set of the BabyLM corpus.

C Detailed resutls: Reading time experiments

Corpus	Step	Spearman's ρ	p-value
babylm	500	-0.5913	0.0097
babylm	1000	-0.6285	0.0052
babylm	2000	-0.7833	0.0001
babylm	3000	-0.7874	0.0001
babylm	1	-0.7915	0.0001
babylm	5	-0.614	0.0067
wikitext-103	500	0.0815	0.7478
wikitext-103	1000	-0.4241	0.0794
wikitext-103	2000	-0.7482	0.0004
wikitext-103	3000	-0.7441	0.0004
wikitext-103	1	-0.7172	0.0008
wikitext-103	5	-0.7523	0.0003

Table 2: Spearman's ρ of model size (in terms of number of parameters) and Δ log-likelihood over the baseline LME model. Steps 1 and 5 refer to the first and fifth epoch.

D   Detailed results: BabyLM challenge tasks
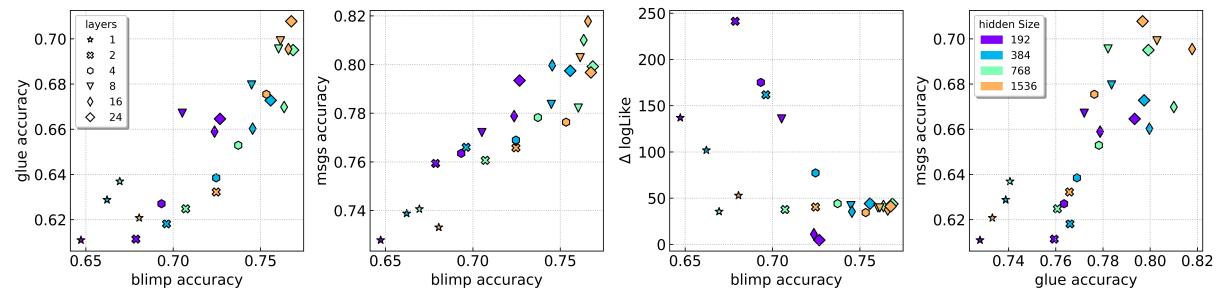
Figure 7: Correlation of LM performance on BLiMP vs. GLUE, BLiMP vs. MSGS, GLUE vs. MSGS.

Table 3: BLiMP accuracy by task and model

Task	1*192	1*384	1*768	1*1536	2*192	2*384	2*768	2*1536	4*192	4*384	4*768	4*1536	8*192	8*384	8*768	8*1536	16*192	16*384	16*768	16*1536	24*192	24*384	24*768	24*1536	OPT-125m (baseline)
anaphor-agreement	0.35±0.01	0.96±0.00	0.96±0.00	0.95±0.00	0.94±0.01	0.96±0.00	0.97±0.00	0.97±0.00	0.96±0.00	0.97±0.00	0.96±0.00	0.98±0.00	0.95±0.01	0.97±0.00	0.99±0.00	0.98±0.00	0.96±0.01	0.97±0.00	0.98±0.00	0.97±0.00	0.96±0.01	0.97±0.01	0.98±0.00	0.98±0.00	94.9
argument-structure	0.70±0.00	0.71±0.00	0.72±0.00	0.72±0.00	0.76±0.00	0.76±0.00	0.72±0.00	0.77±0.00	0.76±0.00	0.78±0.00	0.76±0.00	0.79±0.00	0.77±0.00	0.79±0.00	0.79±0.00	0.79±0.00	0.78±0.00	0.80±0.00	0.80±0.00	0.80±0.00	0.79±0.01	0.74±0.00	0.80±0.00	0.79±0.01	73.8
binding	0.67±0.00	0.67±0.00	0.68±0.01	0.68±0.00	0.69±0.00	0.69±0.02	0.72±0.00	0.71±0.02	0.72±0.02	0.74±0.01	0.74±0.00	0.72±0.00	0.72±0.00	0.74±0.00	0.74±0.00	0.75±0.00	0.73±0.03	0.72±0.00	0.76±0.00	0.76±0.01	0.74±0.00	0.74±0.00	0.76±0.01	0.77±0.01	73.8
control-raising	0.64±0.01	0.66±0.00	0.68±0.01	0.67±0.00	0.67±0.00	0.69±0.00	0.72±0.00	0.69±0.00	0.69±0.00	0.70±0.00	0.74±0.01	0.73±0.00	0.73±0.00	0.74±0.00	0.74±0.00	0.74±0.00	0.73±0.03	0.72±0.00	0.72±0.00	0.76±0.01	0.74±0.00	0.75±0.00	0.74±0.00	0.76±0.01	72.2
determiner-noun-agreement	0.85±0.00	0.85±0.01	0.87±0.00	0.88±0.00	0.91±0.02	0.91±0.00	0.92±0.00	0.93±0.00	0.90±0.00	0.93±0.00	0.93±0.01	0.94±0.00	0.92±0.00	0.95±0.00	0.94±0.00	0.95±0.00	0.93±0.00	0.94±0.00	0.94±0.00	0.95±0.00	0.94±0.00	0.95±0.00	0.94±0.00	0.94±0.00	93.1
ellipsis	0.57±0.01	0.62±0.01	0.63±0.00	0.64±0.01	0.62±0.02	0.63±0.00	0.70±0.00	0.70±0.00	0.68±0.00	0.69±0.01	0.75±0.00	0.78±0.02	0.71±0.00	0.76±0.01	0.82±0.01	0.82±0.01	0.80±0.00	0.80±0.01	0.82±0.03	0.75±0.01	0.76±0.01	0.81±0.00	0.84±0.02	0.82±0.01	80.5
filler-gap	0.64±0.01	0.65±0.00	0.65±0.00	0.65±0.01	0.67±0.00	0.68±0.00	0.70±0.01	0.72±0.00	0.68±0.00	0.71±0.00	0.72±0.00	0.73±0.00	0.70±0.00	0.72±0.00	0.74±0.00	0.74±0.00	0.72±0.00	0.74±0.01	0.75±0.01	0.75±0.01	0.73±0.00	0.74±0.00	0.76±0.00	0.75±0.01	73.6
irregular-forms	0.88±0.01	0.90±0.01	0.89±0.01	0.89±0.01	0.93±0.00	0.93±0.00	0.92±0.00	0.93±0.00	0.92±0.00	0.91±0.01	0.91±0.00	0.90±0.00	0.91±0.02	0.90±0.01	0.92±0.00	0.92±0.00	0.93±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.89±0.03	0.90±0.02	0.92±0.00	0.87±0.02	80.8
island-effects	0.54±0.01	0.55±0.00	0.56±0.01	0.61±0.02	0.50±0.03	0.52±0.02	0.57±0.02	0.57±0.02	0.56±0.03	0.60±0.01	0.65±0.01	0.65±0.00	0.55±0.03	0.64±0.00	0.66±0.02	0.67±0.00	0.61±0.02	0.65±0.01	0.69±0.01	0.69±0.00	0.65±0.02	0.65±0.00	0.68±0.00	0.70±0.00	51.6
npi-licensing	0.53±0.01	0.52±0.00	0.52±0.00	0.56±0.00	0.57±0.00	0.64±0.00	0.60±0.02	0.63±0.00	0.58±0.00	0.65±0.00	0.72±0.00	0.72±0.00	0.63±0.00	0.65±0.00	0.68±0.00	0.71±0.00	0.63±0.03	0.69±0.03	0.72±0.01	0.70±0.02	0.66±0.02	0.75±0.00	0.70±0.01	0.70±0.01	51.6
quantifiers	0.71±0.01	0.71±0.00	0.74±0.01	0.73±0.02	0.73±0.01	0.71±0.00	0.69±0.01	0.71±0.02	0.73±0.00	0.74±0.02	0.71±0.03	0.70±0.00	0.72±0.02	0.76±0.01	0.73±0.01	0.70±0.00	0.69±0.02	0.74±0.00	0.74±0.00	0.66±0.00	0.66±0.02	0.69±0.01	0.77±0.01	0.75±0.00	74.5
subject-verb-agreement	0.62±0.00	0.64±0.00	0.65±0.00	0.67±0.00	0.71±0.00	0.75±0.00	0.79±0.01	0.81±0.00	0.75±0.00	0.83±0.00	0.86±0.00	0.87±0.00	0.81±0.00	0.86±0.00	0.87±0.00	0.88±0.00	0.84±0.01	0.88±0.00	0.88±0.00	0.88±0.00	0.84±0.00	0.87±0.01	0.87±0.00	0.88±0.01	77.3
hypernym	0.51±0.00	0.50±0.01	0.49±0.00	0.48±0.00	0.49±0.00	0.49±0.00	0.47±0.01	0.48±0.00	0.48±0.01	0.50±0.01	0.50±0.00	0.48±0.01	0.48±0.01	0.48±0.00	0.48±0.00	0.48±0.00	0.46±0.00	0.46±0.00	0.48±0.00	0.46±0.00	0.48±0.00	0.47±0.01	0.47±0.00	0.46±0.00	46.3
qa-congruence-easy	0.47±0.03	0.55±0.02	0.52±0.01	0.59±0.00	0.50±0.02	0.58±0.02	0.60±0.01	0.66±0.00	0.58±0.00	0.64±0.02	0.67±0.04	0.71±0.00	0.59±0.02	0.66±0.01	0.69±0.01	0.68±0.03	0.59±0.02	0.66±0.02	0.70±0.04	0.70±0.00	0.62±0.01	0.69±0.03	0.72±0.00	0.68±0.03	76.5
qa-congruence-tricky	0.29±0.00	0.36±0.01	0.35±0.05	0.42±0.02	0.35±0.01	0.39±0.01	0.42±0.01	0.41±0.01	0.33±0.04	0.41±0.04	0.45±0.01	0.46±0.00	0.32±0.00	0.48±0.02	0.52±0.01	0.48±0.01	0.40±0.01	0.50±0.00	0.51±0.00	0.51±0.00	0.38±0.03	0.50±0.04	0.46±0.00	0.53±0.04	47.9
subject-aux-inversion	0.83±0.02	0.81±0.00	0.83±0.01	0.81±0.00	0.83±0.02	0.82±0.01	0.80±0.00	0.86±0.01	0.84±0.00	0.82±0.00	0.83±0.00	0.82±0.00	0.81±0.00	0.83±0.00	0.82±0.00	0.83±0.01	0.85±0.02	0.83±0.01	0.85±0.02	0.85±0.00	0.86±0.01	0.83±0.02	0.87±0.02	0.88±0.01	85.3
turn-taking	0.61±0.02	0.60±0.01	0.65±0.01	0.61±0.01	0.66±0.01	0.65±0.02	0.71±0.00	0.71±0.00	0.64±0.01	0.71±0.01	0.75±0.02	0.75±0.00	0.66±0.02	0.73±0.01	0.78±0.01	0.79±0.01	0.69±0.01	0.74±0.00	0.77±0.00	0.77±0.01	0.72±0.01	0.73±0.00	0.78±0.01	0.78±0.00	82.9
Average	0.65±0.03	0.66±0.03	0.67±0.03	0.68±0.03	0.68±0.03	0.70±0.03	0.71±0.03	0.72±0.03	0.69±0.03	0.72±0.02	0.74±0.02	0.75±0.02	0.71±0.03	0.74±0.02	0.76±0.02	0.76±0.02	0.72±0.03	0.75±0.03	0.77±0.02	0.77±0.02	0.73±0.03	0.76±0.02	0.77±0.02	0.77±0.02	73.11

Table 4: GLUE accuracy by task and model

Task	1*192	1*384	1*768	1*1536	2*192	2*384	2*768	2*1536	4*192	4*384	4*768	4*1536	8*192	8*384	8*768	8*1536	16*192	16*384	16*768	16*1536	24*192	24*384	24*768	24*1536	OPT-125m (baseline)
boolq	0.59±0.00	0.59±0.00	0.59±0.00	0.58±0.00	0.60±0.00	0.59±0.00	0.59±0.00	0.59±0.00	0.59±0.00	0.60±0.00	0.59±0.00	0.60±0.00	0.60±0.00	0.60±0.00	0.61±0.00	0.60±0.01	0.59±0.00	0.70±0.00	0.59±0.00	0.61±0.00	0.60±0.00	0.60±0.00	0.60±0.00	0.64±0.01	66.0
cola	0.69±0.00	0.70±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.70±0.00	0.70±0.00	0.69±0.00	0.69±0.00	0.70±0.00	0.70±0.00	0.70±0.00	0.70±0.00	0.70±0.00	0.74±0.00	0.74±0.00	0.70±0.00	0.71±0.00	0.75±0.00	0.70±0.00	36
mnli	0.55±0.00	0.58±0.00	0.58±0.00	0.53±0.01	0.54±0.00	0.55±0.00	0.57±0.00	0.56±0.01	0.59±0.00	0.64±0.00	0.65±0.00	0.67±0.00	0.71±0.00	0.72±0.00	0.74±0.00	0.74±0.00	0.66±0.00	0.66±0.00	0.67±0.00	0.74±0.00	0.64±0.00	0.66±0.00	0.73±0.00	0.75±0.00	70.1
mnli-mm	0.56±0.00	0.57±0.00	0.54±0.00	0.54±0.00	0.54±0.00	0.55±0.00	0.55±0.00	0.56±0.01	0.61±0.02	0.68±0.00	0.68±0.00	0.70±0.00	0.72±0.00	0.73±0.00	0.74±0.00	0.74±0.00	0.68±0.00	0.70±0.00	0.70±0.00	0.70±0.01	0.69±0.00	0.71±0.00	0.74±0.00	0.76±0.00	71.9
mrpc	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.69±0.00	0.70±0.01	0.69±0.00	0.73±0.00	0.69±0.00	0.70±0.00	82.1
multirc	0.50±0.00	0.51±0.00	0.50±0.01	0.50±0.00	0.49±0.02	0.52±0.01	0.51±0.01	0.52±0.01	0.51±0.01	0.52±0.01	0.52±0.00	0.53±0.00	0.50±0.01	0.53±0.00	0.50±0.01	0.51±0.00	0.48±0.01	0.52±0.02	0.50±0.00	0.52±0.02	0.50±0.00	0.50±0.01	0.50±0.00	0.51±0.01	61.1
qnli	0.60±0.00	0.61±0.00	0.60±0.00	0.60±0.00	0.61±0.00	0.61±0.00	0.61±0.00	0.62±0.00	0.62±0.00	0.78±0.00	0.80±0.00	0.81±0.00	0.79±0.00	0.81±0.00	0.82±0.01	0.82±0.00	0.80±0.00	0.80±0.00	0.81±0.00	0.82±0.00	0.78±0.00	0.80±0.00	0.82±0.00	0.83±0.00	80.1
qqp	0.76±0.00	0.76±0.00	0.77±0.00	0.71±0.00	0.73±0.00	0.74±0.00	0.75±0.00	0.75±0.00	0.75±0.01	0.78±0.00	0.80±0.00	0.81±0.00	0.83±0.00	0.84±0.00	0.80±0.00	0.85±0.00	0.79±0.00	0.79±0.00	0.80±0.00	0.85±0.00	0.78±0.00	0.79±0.00	0.84±0.00	0.86±0.00	77.8
rte	0.47±0.00	0.57±0.00	0.56±0.04	0.50±0.00	0.53±0.01	0.55±0.01	0.35±0.01	0.49±0.00	0.51±0.02	0.54±0.01	0.51±0.00	0.56±0.01	0.53±0.02	0.51±0.00	0.51±0.00	0.52±0.01	0.53±0.02	0.52±0.01	0.52±0.01	0.52±0.00	0.52±0.00	0.56±0.01	0.56±0.04	0.54±0.01	67.7
sst2	0.82±0.00	0.83±0.00	0.84±0.01	0.85±0.00	0.85±0.00	0.86±0.00	0.86±0.00	0.86±0.00	0.85±0.00	0.86±0.00	0.86±0.01	0.85±0.00	0.86±0.00	0.88±0.00	0.87±0.00	0.87±0.00	0.85±0.00	0.85±0.00	0.87±0.00	0.89±0.00	0.85±0.00	0.87±0.00	0.88±0.00	0.89±0.00	86.6
wsc	0.48±0.00	0.52±0.00	0.53±0.03	0.52±0.03	0.55±0.03	0.53±0.01	0.55±0.03	0.61±0.01	0.52±0.00	0.55±0.03	0.55±0.00	0.51±0.04	0.56±0.00	0.59±0.02	0.61±0.00	0.50±0.02	0.49±0.01	0.50±0.02	0.58±0.00	0.58±0.01	0.50±0.02	0.51±0.02	0.57±0.00	0.57±0.02	59.0
Average	0.61±0.03	0.63±0.03	0.62±0.02	0.61±0.02	0.62±0.03	0.62±0.02	0.62±0.02	0.63±0.02	0.63±0.02	0.65±0.03	0.68±0.02	0.67±0.02	0.68±0.02	0.70±0.02	0.70±0.03	0.71±0.03	0.66±0.03	0.67±0.02	0.67±0.03	0.70±0.03	0.66±0.02	0.67±0.02	0.70±0.03	0.71±0.03	68.95

Table 5: MSGS accuracy by task and model

| Task	1*192	1*384	1*768	1*1536	2*192	2*384	2*768	2*1536	4*192	4*384	4*768	4*1536	8*192	8*384	8*768	8*1536	16*192	16*384	16*768	16*1536	24*192	24*384	24*768	24*1536
main-verb-control	0.63±0.01	0.66±0.01	0.65±0.00	0.66±0.00	0.75±0.04	0.90±0.01	0.91±0.00	0.93±0.01	0.92±0.00	0.95±0.01	0.99±0.00	0.99±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00	1.00±0.00	1.00±0.00	1.00±0.00
control-raising-control	0.82±0.01	0.85±0.00	0.84±0.02	0.90±0.00	0.88±0.01	0.91±0.01	0.93±0.00	0.93±0.00	0.92±0.00	0.94±0.00	0.94±0.01	0.96±0.00	0.96±0.00	0.93±0.01	0.96±0.00	0.93±0.00	0.97±0.00	0.95±0.01	0.95±0.00	0.97±0.00	0.94±0.00	0.95±0.00	0.96±0.01	0.96±0.00
syntactic-category-control	0.82±0.02	0.82±0.00	0.84±0.00	0.79±0.01	0.90±0.01	0.88±0.00	0.82±0.02	0.83±0.01	0.81±0.01	0.78±0.01	0.83±0.00	0.83±0.00	0.84±0.01	0.83±0.00	0.83±0.00	0.78±0.00	0.78±0.00	0.83±0.01	0.84±0.02	0.87±0.00	0.83±0.01	0.84±0.02	0.83±0.00	0.82±0.01
lexical-content-the-control	0.79±0.01	0.79±0.00	0.80±0.02	0.76±0.00	0.80±0.02	0.76±0.00	0.75±0.00	0.56±0.01	0.81±0.01	0.78±0.00	0.79±0.01	0.83±0.01	0.79±0.01	0.79±0.01	0.77±0.00	0.78±0.01	0.78±0.00	0.78±0.01	0.76±0.00	0.76±0.00	0.83±0.01	0.84±0.02	0.76±0.01	0.75±0.00
relative-position-control	0.96±0.01	0.98±0.00	0.95±0.01	0.92±0.00	0.92±0.00	0.97±0.00	0.96±0.00	0.97±0.00	0.98±0.00	0.98±0.00	0.97±0.00	0.99±0.00	0.99±0.00	0.99±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.01	1.00±0.00	0.99±0.00	0.98±0.01	1.00±0.00	0.99±0.00
main-verb-lexical-content-the	0.67±0.00	0.66±0.00	0.66±0.00	0.67±0.00	0.67±0.00	0.67±0.00	0.67±0.00	0.67±0.00	0.68±0.00	0.67±0.00	0.68±0.00	0.68±0.01	0.73±0.02	0.67±0.00	0.68±0.00	0.68±0.00	0.69±0.00	0.81±0.01	0.71±0.01	0.71±0.00	0.71±0.01	0.73±0.00	0.77±0.02	0.76±0.01
main-verb-relative-token-position	0.67±0.03	0.66±0.00	0.66±0.00	0.67±0.00	0.67±0.00	0.66±0.00	0.67±0.00	0.67±0.00	0.68±0.00	0.67±0.00	0.69±0.01	0.68±0.01	0.75±0.01	0.69±0.01	0.70±0.01	0.75±0.01	0.74±0.01	0.74±0.00	0.76±0.01	0.78±0.01	0.72±0.02	0.74±0.01	0.77±0.02	0.78±0.02
syntactic-category-lexical-content-the	0.67±0.03	0.70±0.00	0.73±0.01	0.71±0.00	0.70±0.01	0.67±0.01	0.67±0.03	0.67±0.03	0.68±0.01	0.65±0.02	0.68±0.01	0.68±0.02	0.68±0.02	0.67±0.02	0.68±0.00	0.67±0.03	0.68±0.03	0.64±0.01	0.64±0.01	0.68±0.00	0.64±0.01	0.67±0.01	0.67±0.01	0.66±0.01
syntactic-category-relative-position	0.66±0.00	0.66±0.00	0.66±0.00	0.66±0.00	0.65±0.01	0.67±0.01	0.66±0.00	0.66±0.00	0.65±0.01	0.63±0.00	0.68±0.01	0.65±0.02	0.64±0.01	0.64±0.01	0.63±0.00	0.63±0.01	0.63±0.00	0.62±0.00	0.63±0.00	0.63±0.00	0.62±0.00	0.62±0.00	0.62±0.00	0.61±0.00
control-raising-lexical-content-the	0.66±0.00	0.67±0.00	0.67±0.00	0.66±0.00	0.67±0.00	0.67±0.00	0.66±0.00	0.67±0.00	0.70±0.01	0.67±0.01	0.68±0.01	0.68±0.00	0.79±0.01	0.68±0.00	0.70±0.02	0.75±0.00	0.82±0.05	0.72±0.00	0.72±0.00	0.75±0.00	0.72±0.00	0.75±0.03	0.71±0.02	0.82±0.01
control-raising-relative-token-position	0.67±0.00	0.67±0.00	0.67±0.00	0.68±0.00	0.67±0.00	0.67±0.00	0.67±0.00	0.67±0.00	0.69±0.00	0.67±0.00	0.69±0.00	0.69±0.01	0.69±0.00	0.67±0.00	0.69±0.00	0.73±0.01	0.73±0.00	0.74±0.01	0.69±0.01	0.73±0.01	0.69±0.01	0.71±0.01	0.72±0.00	0.71±0.00
Average	0.73±0.02	0.74±0.02	0.74±0.02	0.73±0.02	0.76±0.02	0.77±0.03	0.77±0.03	0.76±0.02	0.76±0.03	0.77±0.02	0.78±0.03	0.78±0.03	0.80±0.03	0.77±0.03	0.78±0.03	0.78±0.03	0.80±0.03	0.81±0.03	0.79±0.03	0.82±0.03	0.79±0.03	0.8±0.03	0.8±0.03	0.8±0.03