Towards Fundamental Language Models: Evaluating Linguistic Competence Across Model Sizes

Anonymous ACL submission

Abstract

Fundamental Language Models (FLMs) propose a novel paradigm that separates linguistic competence from factual knowledge to address critical challenges in current language models, including hallucinations, data privacy concerns, and training-induced biases. This paper investigates whether FLMs can maintain robust language processing capabilities while externalizing factual knowledge. Through comprehensive evaluation of linguistic competence across model sizes using specialized benchmarks, we assess lexical, grammatical, and semantic capabilities. We also analyze how model size affects both linguistic and factual knowledge encoding. Our findings demonstrate that linguistic competence stabilizes at relatively modest model sizes, while factual knowledge continues scaling with model size. These results provide empirical support for FLMs as a promising research direction, suggesting that future work could effectively balance language understanding with external knowledge retrieval.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable linguistic capabilities, achieving state-of-the-art performance across various natural language processing tasks. However, these models often face critical challenges (Bengio et al., 2025) such as hallucinations—where they generate false or fabricated information—data privacy concerns, and the propagation of biases inherited from their training data. A key factor contributing to these challenges is the substantial volume of factual information that LLMs internalize during training, where models simultaneously encode both factual data and linguistic structures. This integration makes it difficult to update factual knowledge without retraining and complicates efforts to ensure accuracy and reduce bias.

This paper explores an alternative approach: Fundamental Language Models (FLMs). Instead of internalizing factual knowledge, FLMs aim to preserve the core linguistic competence of traditional LLMs while delegating factual retrieval to external knowledge sources. This separation could offer several advantages such as reducing model size, mitigating biases, and improving factual accuracy by relying on dynamically retrieved information rather than static, potentially outdated internalized knowledge.

Linguistic competence in this context refers to the model's ability to generate and comprehend language by understanding linguistic structures, such as grammar, vocabulary, and meaning, without reliance on embedded factual data. Drawing on linguistic theory as defined by the Council of Europe in its Common European Framework of Reference for Languages (CEFR)[1], FLMs should prioritize three key sub-competences:

1. **Lexical Competence**: Knowledge of, and ability to use, the vocabulary of a language consisting of word classes and fixed expressions.

2. **Grammatical Competence**: Knowledge of, and ability to understand and express meaning by producing and recognising well-formed phrases and sentences.

3. **Semantic Competence**: The capacity to generate and understand meaningful phrases, sentences, and text, including resolving ambiguity, paraphrasing, and interpreting nuanced meanings in context.

Phonological and orthoepic competences—related to spoken language—and orthographic competence—related to spelling—are less relevant to text-based models and thus remain outside the primary focus of FLMs.

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-and-its-language-versions

While traditional retrieval-augmented generation (RAG) systems enhance factual retrieval capabilities while maintaining full-scale language models, FLMs propose a more fundamental separation between linguistic and factual knowledge. Our research evaluates whether linguistic competence remains robust in smaller models, supporting FLMs as a viable direction for future development.

The paper is structured as follows: in Section 1, we introduce FLMs and their potential to disentangle linguistic competence from factual knowledge, addressing critical challenges in current LLM architectures. Section 2 examines related work, reviewing transformer models, linguistic evaluation methods, and theoretical frameworks. Section 3 presents the evaluation methodology, detailing the assessment of linguistic competencies. Section 4 presents our experimental results, with findings in both linguistic competence findings and factual knowledge analysis. Finally, Section 5 synthesizes our results, demonstrating that linguistic competence stabilizes at smaller model sizes while factual knowledge continues to scale, supporting the viability of the FLM approach.

2 Related Work

While transformer models exhibit impressive capabilities in handling linguistic tasks, they do not replicate traditional linguistic analysis methods. The models encode semantic roles and grammatical features in specific regions of sentence embeddings, rather than distributing this information evenly across the entire embedding (Nastase and Merlo, 2024). Anyhow, it seems that each layer in a transformer captures different levels of linguistic information, from local to global dependencies (Garnier-Brun et al., 2024). This was already found in not so large models like BERT, and several works have found that linguistic related information is encoded in a hierarchy way and some layers seem to focus on different aspects (Rogers et al., 2021). From these findings, the next question that arises is: How large a language model has to be to become linguistically competent? Although larger models perform better, yet smaller models can still achieve significant results (Steuer et al., 2023).

The evaluation of LLMs on linguistic competence is present in many benchmarks (Chang et al., 2024). Some studies have focused on this type of tests to identify the proficiency of language models in linguistic aspects like grammar, vocabulary or syntax, and compared it to reasoning capabilities (Atox and Clark, 2024). The work by (Dentella et al., 2024) examines LLMs' ability to understand uncommon meanings of common words, finding that even advanced models like GPT-4 perform worse than teenagers at this task. This reveals important limitations in LLMs' semantic understanding capabilities, despite their otherwise impressive language abilities. This could suggest that going to larger models may not scale up linguistic capabilities.

The BabyLM challenge has been engaging the research community to train language models on a limited set of texts, with the aim to emulate the way humans learn in their infancy (Hu et al., 2024). One of the most interesting findings was that, even with such a constrained set of training material, the performance of the models was not too far from models trained over trillions of tokens, like LLaMa2. Effective approaches included preprocessing of the training data, and some enhancements to the transformer architecture.

The Sapir-Whorf hyphothesis, also know as *linguistic relativity* proposes that language influences our understanding of the world and, even more, our cognitive skills (Penn, 2014). This hypothesis, which dates back to the middle of the 20th century, has been partially supported by the "emergent" abilities of large language models, though it is still an open discussion (Schaeffer et al., 2023). Studies with pre-linguistic infants have shown abilities to understand physical causality and object permanence (Hespos and Spelke, 2004). Early research work on chimpanzees showed they could solve complex puzzles and understand cause-effect relationships without linguistic abilities (Premack, 1959). We could conclude that reasoning is something more than language, as symbolic reasoning can occur without "talking" to ourselves.

The rising of the so-called *Agentic AI* paradigm is driven the evolution of artificial intelligence system far from monolithic approaches (Acharya et al., 2025). So larger is not necessary better, and the cooperation of several language models, with differentiated roles, is a promising path (Feng et al., 2025). Recent research has found that LLMs may have reached the peak in reasoning capabilities despite their size (Lin et al., 2025).

Linguistic relativity may not be fully right. Yet language strongly influences thought (Dong, 2022). Actually, large language models are still in the core of the most advanced solutions in artificial

intelligence. If language by itself could be such a powerful tool in natural thinking, the pursuit of linguistic competence isolated from factual knowledge is justified (Liu et al., 2024), as demonstrated by the rapid adoption of retrieval-augmented generative tools (RAG) (Lewis et al., 2020). To the best of our knowledge, no prior study exists focusing on the trade-off between model size and linguistic competence, due to the variety of architectures and training objectives across available models.

3 Evaluating Fundamental Language Models

Fundamental Language Models (FLMs) aim to separate linguistic competence from factual knowledge, ensuring that models retain strong language-processing abilities while externalizing factual retrieval. To explore this hypothesis, we assess linguistic competence and factual knowledge performance across various model families, including Llama-3, Qwen2.5, Gemma-2, and Yi-1.5, and sizes ranging from 0.5 to 9 billion parameters. Our evaluation follows a structured approach, leveraging well-established benchmarks present in the LM Evaluation Harness (Gao et al., 2024)—a unified framework to test generative language models— to analyze different competencies.

According to our definition of FLMs, these models should excel in language-related tasks while struggling with factual knowledge tasks unless supplemented with external retrieval mechanisms. However, defining strict boundaries between linguistic competence and factual knowledge is challenging, as effective communication often relies on shared world knowledge. Despite this complexity, we focus on benchmarks that best capture these two distinct abilities.

3.1 Linguistic Competence

To evaluate the linguistic competence of FLMs, we employ benchmarks that assess language processing abilities while minimizing reliance on factual knowledge. Using tasks from the LM Evaluation Harness, we examine three core linguistic competencies: lexical competence, grammatical competence, and semantic competence.

Each sub-competence is assessed through specific tasks designed to measure a model's ability to handle vocabulary, syntax, and meaning. This ensures that FLMs can generate and comprehend language effectively without the need for internalized factual knowledge.

3.1.1 Lexical Competence

Lexical competence refers to the model's ability to understand and use vocabulary effectively in different contexts. To evaluate this competence we use Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019):

- **WiC**: This dataset tests word sense disambiguation by presenting sentence pairs containing the same word. The task is to determine whether the word has the same meaning in both contexts. This task is primarily lexical, as it requires knowledge of word senses and their contextual variations. It uses accuracy as evaluation metric.

3.1.2 Grammatical Competence

Grammatical competence assesses the model's ability to generate and comprehend syntactically well-formed sentences. To this end, we find that The Benchmark of Linguistic Minimap Pairs (BLiMP) (Warstadt et al., 2020) evaluates this competence in depth:

- **BLiMP**: This benchmark consists of minimal sentence pairs, where one sentence is grammatically correct and the other contains a syntactic violation. The model must distinguish between the two, testing its grasp of linguistic rules, such as agreement, negation, and binding dependencies. It uses accuracy as evaluation metric.

3.1.3 Semantic Competence

Semantic competence concerns the model's ability to generate and comprehend meaningful phrases and sentences, which includes understanding sentence-level meaning, resolving ambiguity, and recognizing nuanced language use. Benchmarks from the LM Evaluation Harness that assess this competence include Recognizing Textual Entailment (RTE) (Dagan et al., 2005), Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018), and Quora Question Pairs (QQP)[2], which are all evaluated using accuracy and averaged for the final semantic competence score:

- **RTE**: This benchmark measures whether a model can determine if one sentence logically

[2]https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

follows from another. This task requires deep semantic understanding, as it tests the model's ability to grasp the meaning of sentences and their logical relationships.

- **MNLI**: This dataset tests whether a model can classify sentence pairs as entailment, contradiction, or neutral, evaluating its ability to capture meaning across different domains.

- **QQP**: This task involves determining whether two questions are semantically equivalent. It tests the model's ability to understand paraphrases and sentence-level meaning, making it a key benchmark for evaluating semantic competence.

3.2 Factual Knowledge

We categorize factual knowledge into two types: external factual knowledge, which involves reasoning over provided information, and internal factual knowledge, which assesses the model's memorization of factual data. This distinction is important because FLMs should maintain the ability to reason and extract relevant information from documents while minimizing reliance on memorized facts.

3.2.1 External Factual Knowledge

External factual knowledge requires reasoning based on given context rather than recalling stored facts. We evaluate this using datasets that provide a source passage or context to retrieve the answer from such as LAnguage Modeling Broadened to Account for Discourse Aspects (LAMBADA) (Paperno et al., 2016), BoolQ (Clark et al., 2019), Choice of Plausible Alternatives (COPA) (Gordon et al., 2011), Multi-Sentence Reading Comprehension (MultiRC) (Khashabi et al., 2018), and Reading Comprehension with Commonsense Reasoning Dataset (ReCoRD) (Zhang et al., 2018), which are all evaluated using accuracy excluding the last one, which is evaluated through exact matching (EM):

- **LAMBADA**: LAMBADA standard is a collection of narrative passages sharing the characteristic that human subjects are able to guess their last word if they are exposed to the whole passage, but not if they only see the last sentence preceding the target word.

- **BoolQ**: It is a question-answering dataset for yes/no questions where each example is a triplet of (question, passage, answer).

- **COPA**: This dataset assesses causal reasoning by presenting a premise and two alternative completions, requiring the model to select the most plausible one.

- **MultiRC**: It is a dataset of short paragraphs and multi-sentence questions that can be answered from the content of the paragraph.

- **ReCoRD**: Consists of queries automatically generated from CNN/Daily Mail news articles. The answer to each query is a text span from a summarizing passage of the corresponding news.

3.2.2 Internal Factual Knowledge

Internal factual knowledge refers to factual information that a model has memorized during training. This knowledge is particularly relevant for traditional LLMs, which internalize vast amounts of data. However, for FLMs, the goal is to minimize reliance on internalized facts, instead retrieving information dynamically from external sources. To evaluate internal factual knowledge, we use benchmarks that test the model's ability to recall specific facts without access to external context. These include TriviaQA (Joshi et al., 2017), which is evaluated using EM, and TruthfulQA (Lin et al., 2022), which provides several different metrics (i.e. BLEU, ROUGE-1, ROUGE-2, and ROUGE-L). To select one for the latter, we have considered a systemic hypothesis based on Konrad Lorenz's observation that *[...] the parts interacting in a system can be understood only simultaneously and together, or not at all* (Lorenz, 1960). The metric that fits best with this perspective is ROUGE-L, since this metric prioritizes recall over precision, acknowledging that missing crucial information fragments systemic understanding more severely than including supplementary details, while precision penalizes additional content that may actually enhance systemic comprehension. In our opinion, ROUGE-L better preserves the holistic nature of meaning by favoring that all essential components should be present in the answer, making it more aligned with how humans process information in complex linguistic systems:

- **TriviaQA**: It is a large-scale reading comprehension dataset that includes question-answer pairs authored by trivia enthusiasts. This dataset provides evidence documents automatically gathered that no dot guarantee to

contain all facts needed to answer the question. Consequently, the LM Evaluation Harness excludes these documents during evaluation, making TriviaQA a suitable benchmark for assessing internal factual knowledge.

- **TruthfulQA**: This benchmark is designed to test a model's ability to generate factually accurate responses while avoiding common misconceptions. It comprises three tasks: (1) *TruthfulQA Generation*, where the model generates a 1-2 sentence response to a given question; (2) *TruthfulQA MC1*, a multiple-choice task requiring the selection of the single correct answer from 4-5 options; and (3) *TruthfulQA MC2*, which presents a question along with multiple true/false reference answers and scores the model based on the normalized probability assigned to the correct responses.

4 Results and discussion

This section presents the results of our evaluation across linguistic competence, external factual knowledge, and internal factual knowledge. We also discuss key findings and their implications for the feasibility of FLMs as an alternative paradigm to traditional LLMs. Every experiment has been executed on a single NVIDIA Ampere A100 GPU.

Linguistic Competence Table 1 presents the results for linguistic competence. The model Qwen2.5-7B achieved the highest overall linguistic competence score of 0.7239, indicating strong performance across lexical, grammatical, and semantic tasks. Qwen2.5-3B also performed competitively, scoring 0.6909. These findings suggest that linguistic competence can remain stable even at moderate model sizes, challenging the assumption that larger models are always necessary for high language-processing performance. Semantic competence results are further described in Appendix A, Table 4.

External Factual Knowledge Table 2 presents the results for external factual knowledge evaluation. The model gemma-2-9b achieved the highest score (0.7961), followed by Llama-3.1-8B (0.7712). These results suggest that external factual knowledge continues to improve with model size, reinforcing the idea that factual retrieval is increasingly effective in larger models.

Internal Factual Knowledge Table 3 reports the results for internal factual knowledge. The model gemma-2-9b demonstrated the highest internal factual knowledge score (0.4598), followed closely by Yi-1.5-9B (0.4362). However, performance in this category was noticeably lower than in linguistic competence and external factual knowledge. This supports the hypothesis that internal factual knowledge is highly dependent on model size, as larger models tend to memorize more factual data.

4.1 Discussion

Figures 1, 2, and 3 illustrate the primary competences examined in this study: linguistic competence, external factual knowledge, and internal factual knowledge, respectively. Each figure presents the trend line that best fits the observed data based on the highest R^2 value among linear, exponential, and logarithmic models. From these figures, we derive the following insights: (1) Linguistic competence follows a linear trend with a near-zero slope, indicating that increasing model size does not significantly improve linguistic capabilities beyond a certain threshold. This supports the viability of FLMs, as smaller models can retain strong language-processing abilities. (2) Both external and internal factual knowledge exhibit logarithmic growth with model size, suggesting that while factual knowledge retrieval and memorization improve as models scale up, they do so at diminishing returns.

5 Conclusions

Our evaluation of language models across different sizes reveals several significant findings regarding the relationship between model scale and language related capabilities. The results demonstrate that linguistic competence—encompassing lexical, grammatical, and semantic abilities—stabilizes at relatively modest model sizes, with Qwen2.5-7B achieving peak performance (0.7239) while smaller variants maintained strong capabilities. This finding challenges the assumption that increasingly large models are necessary for sophisticated language processing in tasks that do not require extensive background knowledge (like contextual question-answering, as it is needed in RAG systems).

The analysis of factual knowledge presents a different pattern, with external factual reasoning and comprehension showing consistent improve-

Model	Lexical	Grammatical	Semantic	Linguistic
Qwen2.5-0.5B	0.4937	0.8176	0.5528	0.6214
Llama-3.2-1B	0.4828	0.8246	0.4758	0.5944
Qwen2.5-1.5B	0.5313	**0.8251**	0.6596	0.6720
gemma-2-2b	0.4937	0.7710	0.5028	0.5892
Llama-3.2-3B	0.4969	0.8217	0.4212	0.5799
Qwen2.5-3B	**0.6254**	0.7270	0.7204	0.6909
Yi-1.5-6B	0.5000	0.6936	0.6252	0.6063
Qwen2.5-7B	0.5815	0.8225	**0.7676**	**0.7239**
Llama-3.1-8B	0.5110	0.8195	0.5398	0.6234
gemma-2-9b	0.5125	0.7799	0.5244	0.6056
Yi-1.5-9B	0.6129	0.7054	0.6045	0.6409

Table 1: Averaged accuracy scores for each linguistic subcompetence. Linguistic competence is computed as the average between lexical, grammatical, and semantic scores. Best scores are highlighted in **bold** and second best scores are underlined.

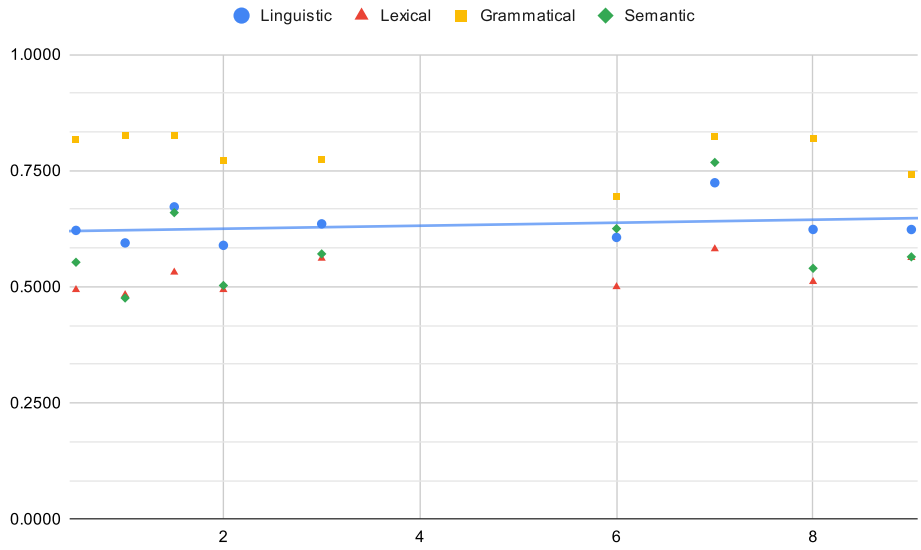

Figure 1: Evaluation for all linguistic subcompetencies. Linguistic competence is best approximated linearly. Models with the same size are averaged.

ments with increased model size. The gemma-2-9b model demonstrated superior performance (0.7961) on external factual tasks, while internal factual knowledge scores remained notably lower across all models. This disparity supports the fundamental premise of FLMs: that factual knowledge can be effectively externalized while maintaining robust linguistic abilities.

These findings provide empirical support for the viability of Fundamental Language Models as an alternative paradigm to traditional LLMs. The demonstrated stability of linguistic competence at smaller scales, combined with the potential for external knowledge integration, suggests a promising direction for developing more efficient and reliable language models. This approach could address key challenges in current LLM architectures, including factual accuracy, bias mitigation, and computational efficiency. Future research should focus on optimizing the balance between model size and linguistic capabilities while developing effective mechanisms for external knowledge retrieval and integration.

Acknowledgments

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional

Model	LAMBADA	BoolQ	COPA	MultiRC	ReCoRD	EFK
Qwen2.5-0.5B	0.4349	0.6245	0.7400	0.3962	0.7704	0.5932
Llama-3.2-1B	0.5393	0.6404	0.7700	0.5670	0.8610	0.6755
Qwen2.5-1.5B	0.5861	0.7291	0.8300	0.2857	0.8442	0.6550
gemma-2-2b	0.6402	0.7343	0.8800	0.5588	0.8930	0.7413
Llama-3.2-3B	0.6423	0.7339	0.8600	**0.5720**	0.9012	0.7419
Qwen2.5-3B	0.5905	0.7722	0.8500	0.3851	0.8752	0.6946
Yi-1.5-6B	0.6802	0.8034	0.8500	0.3426	0.8971	0.7147
Qwen2.5-7B	0.6511	0.8468	0.9100	0.1588	0.8936	0.6921
Llama-3.1-8B	0.6738	0.8211	0.8700	**0.5720**	0.9193	0.7712
gemma-2-9b	**0.7231**	0.8398	**0.9300**	0.5668	**0.9207**	**0.7961**
Yi-1.5-9B	0.6990	**0.8584**	0.8900	0.1914	0.9095	0.7097

Table 2: Scores for each external factual knowledge (EFK) task. Reasoning is computed as the average between all the selected tasks. Best scores are highlighted in **bold** and second best scores are underlined.

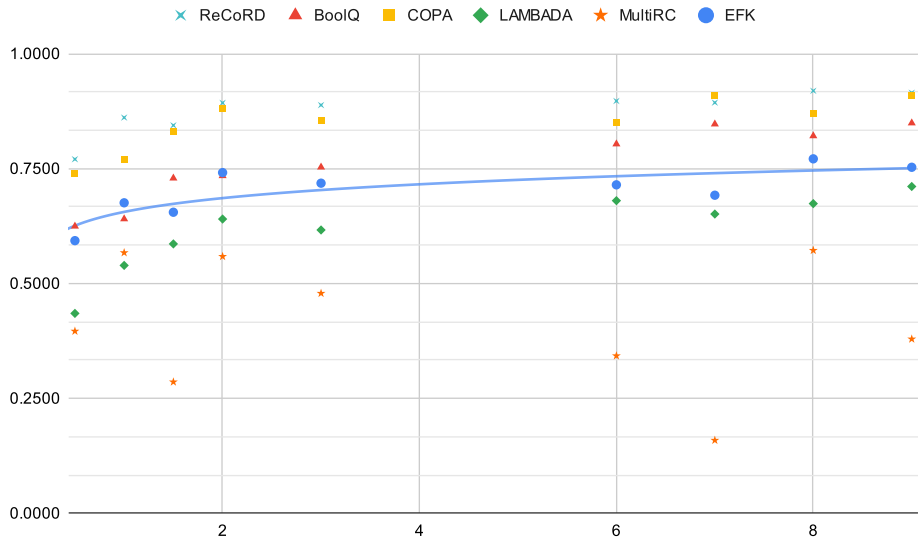

Figure 2: Evaluation for all external factual knowledge tasks. Averaged score is best approximated logarithmically. Models with the same size are averaged.

I+D+i from the Spanish Government.

Limitations

While our research demonstrates the potential of Fundamental Language Models, several important limitations must be acknowledged. The separation of linguistic competence from factual knowledge presents challenges in cases where language understanding inherently requires world knowledge. For example, understanding metaphors, cultural references, or domain-specific terminology often depends on both linguistic and factual knowledge in ways that are difficult to disentangle. Our evaluation framework, though comprehensive, may not fully capture these interdependencies.

The performance stability we observed at smaller model sizes might not generalize across all linguistic tasks or languages. Our benchmarks focus primarily on English, and the relationship between model size and linguistic competence could vary significantly for other languages, particularly those with different syntactic structures or morphological complexity.

Our study also focuses on specific model architectures and sizes, and the findings might not extend to other architectural paradigms or scaling approaches. Future work should address these limitations through multilingual evaluation, real-world deployment testing, and investigation of hybrid approaches that better handle the linguistic-factual knowledge boundary.

Model	TriviaQA	TruthfulQA_gen	TruthfulQA_mc1	TruthfulQA_mc2	IFK
Qwen2.5-0.5B	0.1272	0.0379	0.2534	0.3973	0.2040
Llama-3.2-1B	0.2509	0.1848	0.2313	0.3768	0.2610
Qwen2.5-1.5B	0.2942	0.3696	0.3011	0.4661	0.3578
gemma-2-2b	0.5080	0.2521	0.2399	0.3624	0.3406
Llama-3.2-3B	0.5088	0.1934	0.2497	0.3922	0.3360
Qwen2.5-3B	0.4242	0.3696	0.3182	0.4894	0.4004
Yi-1.5-6B	0.4963	**0.4614**	0.2974	0.4405	0.4239
Qwen2.5-7B	0.5038	0.0575	**0.3905**	**0.5634**	0.3788
Llama-3.1-8B	0.6170	0.3464	0.2827	0.4517	0.4244
gemma-2-9b	**0.6803**	0.4064	0.2987	0.4539	**0.4598**
Yi-1.5-9B	0.5447	0.4137	0.3195	0.4667	0.4362

Table 3: Scores for each internal factual knowledge (IFK) task. Factual knowledge is computed as the average between all the selected tasks. Best scores are highlighted in **bold** and second best scores are underlined.

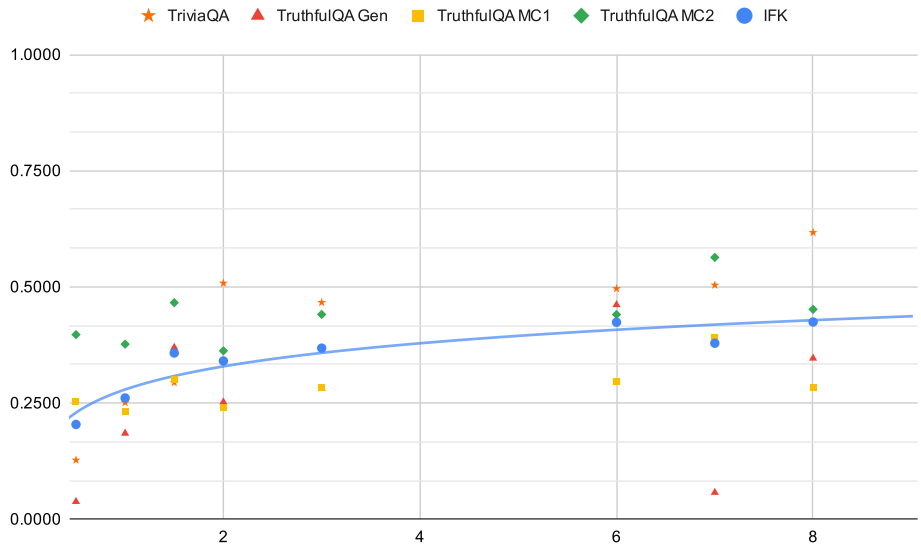

Figure 3: Evaluation for all internal factual knowledge tasks. Averaged score is best approximated through logarithmically. Models with the same size are averaged.

6 Ethical considerations

The development of Fundamental Language Models raises some ethical considerations. While FLMs aim to reduce hallucinations and biases through external knowledge retrieval, this approach introduces new ethical issues. The selection and curation of external knowledge sources could perpetuate or amplify existing biases if not carefully managed. Additionally, the separation of linguistic and factual knowledge raises questions about transparency and accountability - users must understand which parts of the model's responses come from its linguistic processing versus external sources. Therefore, the separation between knowledge and linguistic competence does not ensure the avoidance of already existing problems in LLMs, but could help to identify and mitigate them.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*.

Nathan Atox and Mason Clark. 2024. Evaluating large language models through the lens of linguistic proficiency and world knowledge: A comparative study. *Authorea Preprints*.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, and Yejin Choi. 2025. International ai safety report. *Preprint*, arXiv:2501.17805.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

Vittoria Dentella, Fritz Guenther, and Evelina Leivada. 2024. Language in vivo vs. in silico: Size matters but larger language models still do not comprehend language on a par with humans. *Preprint*, arXiv:2404.14883.

Jianmei Dong. 2022. A study on the relationship between language and thought based on sapir-whorf. *Journal of Global Humanities and Social Sciences*, 3(3):83–85.

Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, et al. 2025. When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2502.04506*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Jérôme Garnier-Brun, Marc Mézard, Emanuele Moscato, and Luca Saglietti. 2024. How transformers learn structured data: insights from hierarchical filtering. *arXiv preprint arXiv:2408.15138*.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Susan J Hespos and Elizabeth S Spelke. 2004. Conceptual precursors to language. *Nature*, 430(6998):453–456.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *Preprint*, arXiv:2412.05149.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. 2024. Ddk: Distilling domain knowledge for efficient large language models. *Preprint*, arXiv:2407.16154.

Konrad Lorenz. 1960. Methods of approach to behaviour problems. *The Harvey Lectures*, 19581959.

Vivi Nastase and Paola Merlo. 2024. Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 203–214.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset.

Julia M Penn. 2014. *Linguistic relativity versus innate ideas: The origins of the Sapir-Whorf hypothesis in German thought*, volume 120. Walter de Gruyter.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

David Premack. 1959. Toward empirical behavior laws: I. positive reinforcement. *Psychological review*, 66(4):219.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36:55565–55581.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, abs/1810.12885.

A Appendix

Model	RTE	MNLI	QQP
Qwen2.5-0.5B	0.5884	0.3869	0.6831
Llama-3.2-1B	0.5668	0.3585	0.5022
Qwen2.5-1.5B	0.7004	0.5254	0.7530
gemma-2-2b	0.6137	0.4338	0.4610
Llama-3.2-3B	0.5451	0.3462	0.3722
Qwen2.5-3B	0.7581	<u>0.5505</u>	<u>0.8527</u>
Yi-1.5-6B	0.7401	0.5437	0.5917
Qwen2.5-7B	**0.8159**	**0.6265**	**0.8605**
Llama-3.1-8B	0.6968	0.5084	0.4141
gemma-2-9b	0.6787	0.4849	0.4096
Yi-1.5-9B	<u>0.7834</u>	0.5077	0.5224

Table 4: Scores for each semantic competence task. Best scores are highlighted in **bold** and second best scores are <u>underlined</u>.