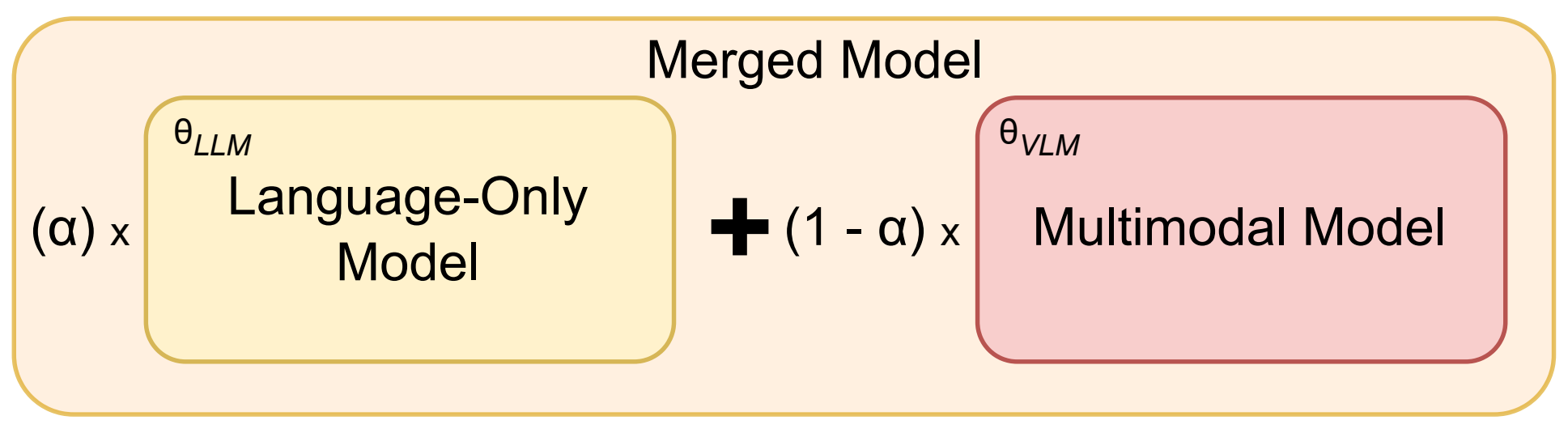
# Model Merging to Maintain Language-Only Performance in Developmentally Plausible Multimodal Models

Ece Takmaz, Lisa Bylinina, Jakub Dotlacil
Utrecht University

## Introduction

- **Multimodal Track of the BabyLM Challenge** [1, 2]
- Previous work, including BabyLM contributions, indicates that **multimodal data** has limited or no benefits in **language-only benchmarks** [3, 4, 5]
- We reach similar conclusions in our low-resource multimodal scenario
- Our multimodal models underperform in grammar-oriented benchmarks, although being exposed to the same language-only data as the language-only models (+ multimodal data from Conceptual Captions and Localized Narratives)
- **How can we mitigate this issue in developmentally plausible multimodal models and maintain language-only performance? Model merging**

## Model Merging



A technique that benefits multi-task and multi-language models, reducing the effects of catastrophic forgetting [6, 7]. We experiment with the **weighted linear interpolation** of language-only and multimodal models

## Training

- BabyLM data: language-only 50M and multimodal 50M words [2]. Trained a tokenizer: 1.36 word-to-subword ratio
- Modifying **LLaVa 1.5** [8]
- Vision encoder replaced with **mean-pooled representation from DINOv2-large** [9]
- Randomly initialized a 6-layer version of LlavaForConditionalGeneration and a multimodal projector
- Language-only input preceded by black image
- **Low-compute** setting with 2 parallel A10 GPUs, 10 epochs

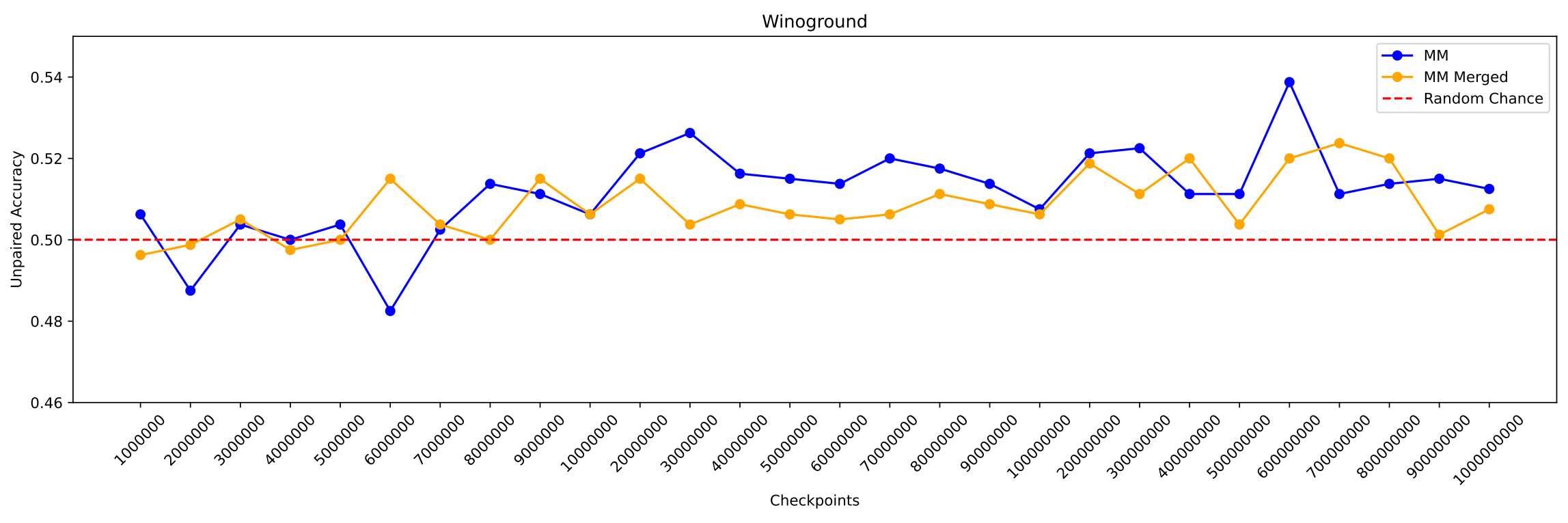
## Multimodal Benchmark

**Winoground** is a challenging benchmark, requiring fine-grained visual and linguistic analyses involving unusual images and texts [10]

painting the white wall red

painting the red wall white

Merging with $\alpha = 0.3$, in some checkpoints, can actually be beneficial without decreasing scores

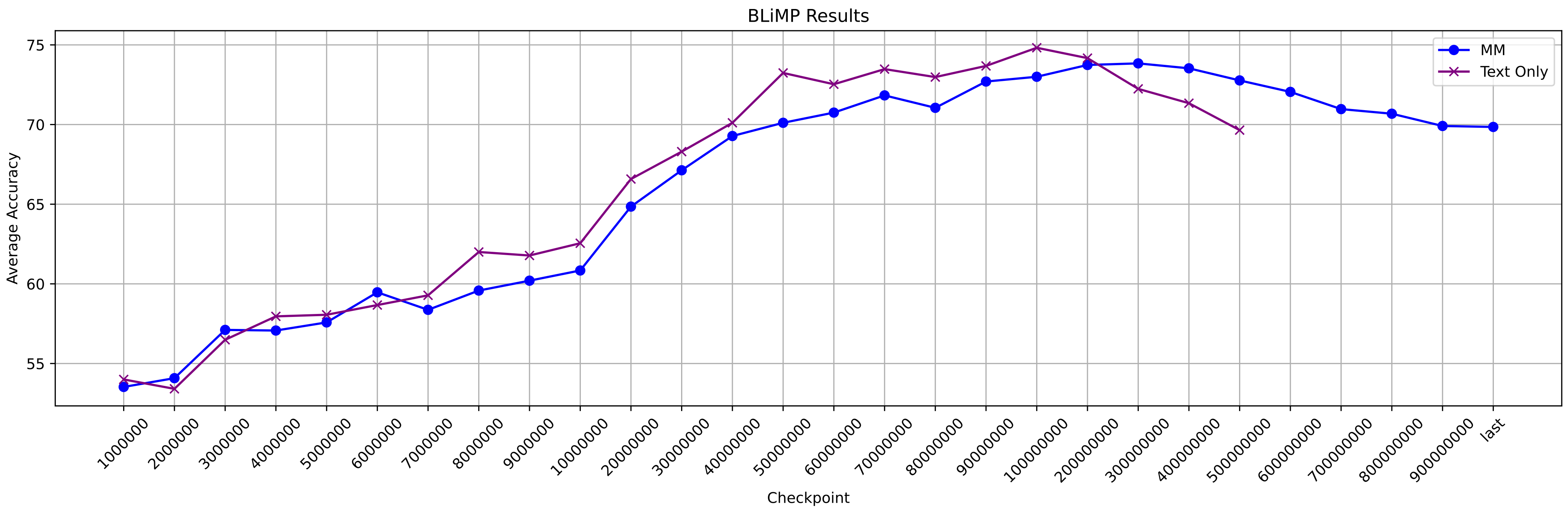


## Conclusion

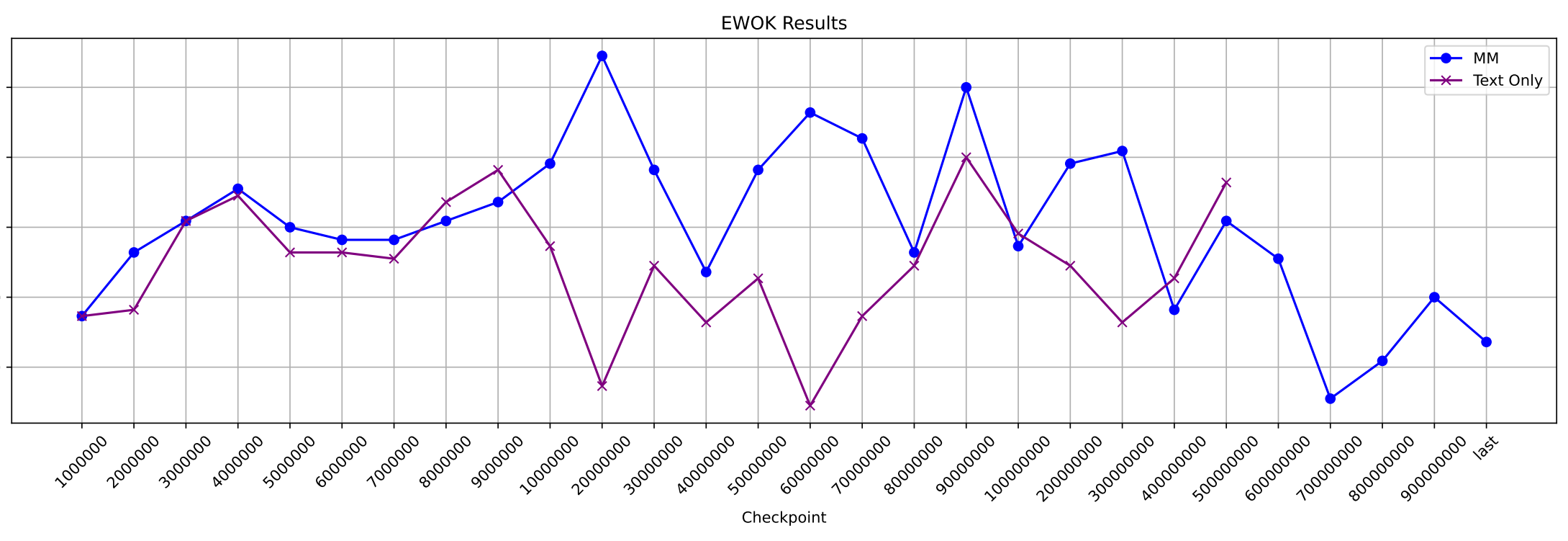
- Our multimodal BabyLM model surpasses previous baselines and submissions on the leaderboard
- Yet, it tends to underperform in text-only benchmarks that focus on grammar compared to language-only models
- Model merging with language-only checkpoints helps alleviate this issue to some extent, benefiting performance in language-only benchmarks and not disrupting accuracy in multimodal tasks heavily
- Future work can explore other model merging techniques and their effects in a wider set of benchmarks

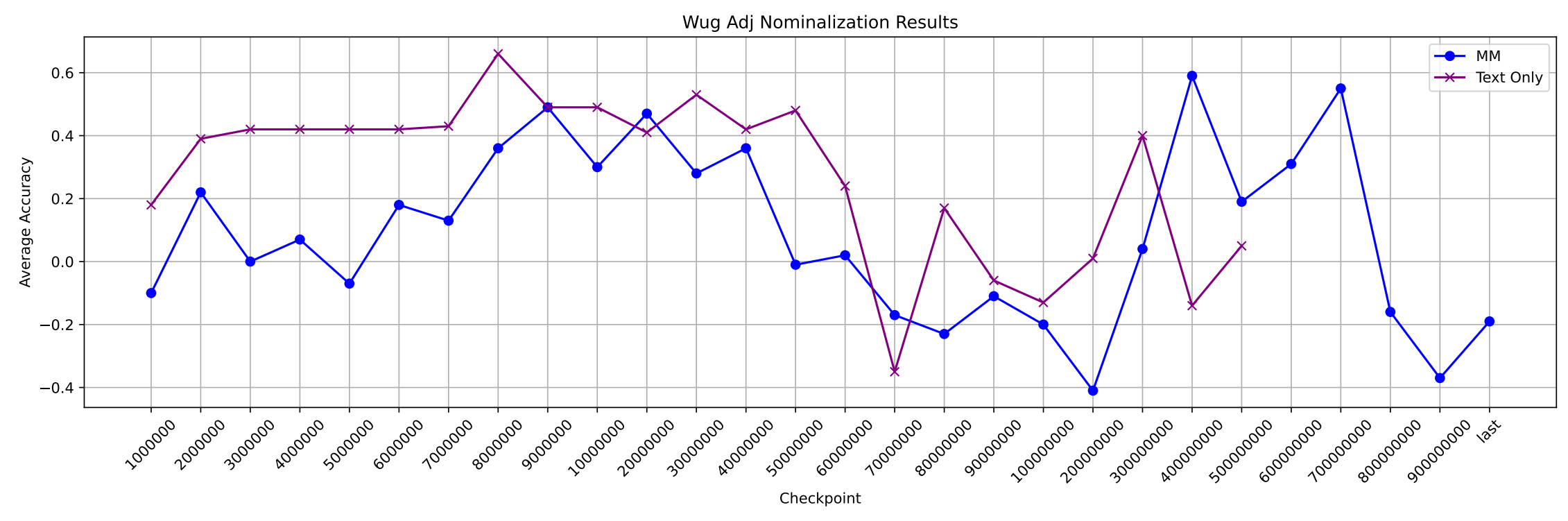
## Language-Only Benchmarks

- Our multimodal model outperforms the multimodal BabyLM baselines and the current submissions on the multimodal leaderboard on the BLiMP benchmark [11]



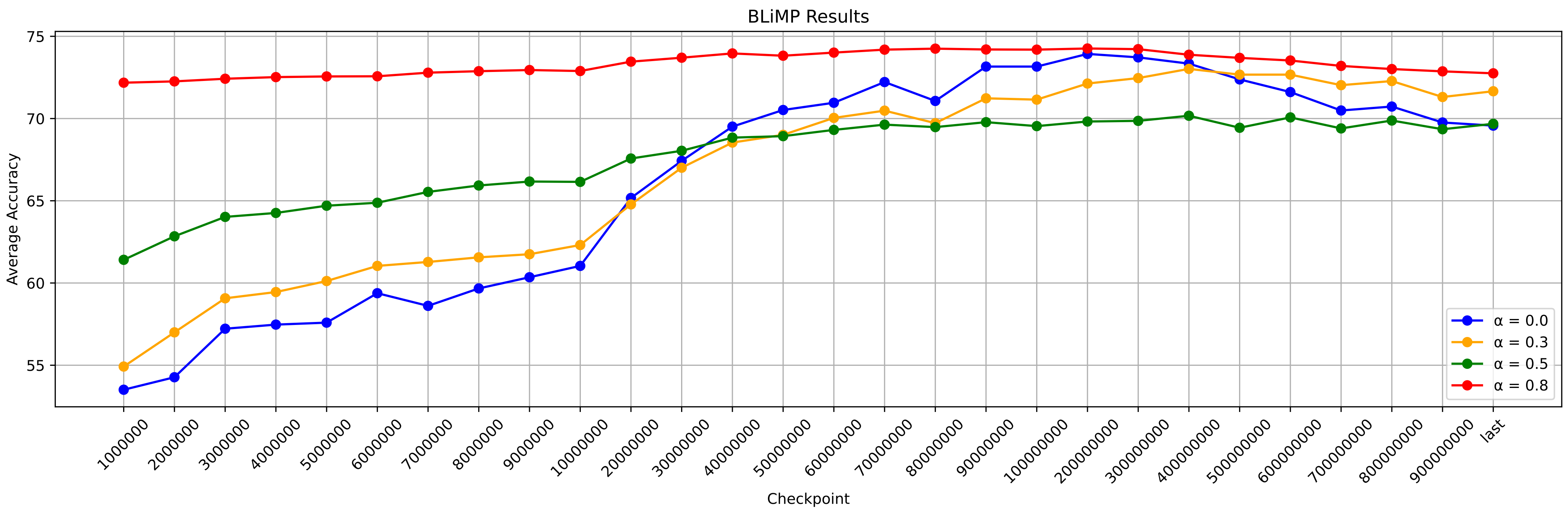
- Benchmarks focusing on grammar **(BLiMP, Wug past tense and Wug adjective nominalization)**: Our multimodal model performs worse than our language-only model
- Knowledge-, semantics- and reasoning-oriented benchmarks **(EWOK and Entity Tracking)**: Multimodal model performs better





### Merging with the best language-only checkpoint

- Early checkpoints: **Merging the multimodal checkpoint with a language-only model leads to better results in BLiMP**
- Later checkpoints: When the multimodal model's language-only capabilities begin to drop, merging can help **maintain language-only capabilities**



## References

[1] Alex Warstadt et al., 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. CoNLL-BabyLM 2023.
[2] Lucas Charpentier et al., 2025. BabyLM turns 3: Call for papers for the 2025 BabyLM workshop.
[3] Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024. Visual grounding helps learn word meanings in low-data regimes. NAACL.
[4] Theodor Amariucai and Alexander Scott Warstadt. 2023. Acquiring linguistic knowledge from multimodal input. CoNLL-BabyLM 2023.
[5] Alina Klerings, Christian Bartelt, and Aaron Mueller. 2024. Developmentally plausible multimodal language models are highly modular. CoNLL-BabyLM 2024.
[6] Enneng Yang et al., 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities
[7] Charles Goddard et al., 2024. Arcee's MergeKit: A toolkit for merging large language models. EMNLP.
[8] Haotian Liu et al., 2024. Improved baselines with visual instruction tuning. CVPR.
[9] Maxime Oquab et al., 2024. DINOv2: Learning robust visual features without supervision. TMLR.
[10] Tristan Thrush et al., 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. CVPR.
[11] Alex Warstadt et al., 2020. Blimp: The benchmark of linguistic minimal pairs for English. TACL

European Research Council
Established by the European Commission