MPhil ACS Project Proposals by Suchir Salhan (2025 -26): PICO and BabyLMs – Learning Dynamics, Interpretability and Multilinguality

Below are MPhil ACS/Part III CST proposals related to Small Language Models:

- **Pico** is our group's **open-source framework for hypothesis-driven pretraining of small language models**, based on a comprehensive analysis and comparison of model learning dynamics. For an introduction to Pico, visit www.picolm.io for guided tutorials and check out our launch video on YouTube.
- We are also interested in **cognitively-inspired language modelling** through BabyLMs – small language models trained on *developmentally-plausible volumes of training data.* Find out more about the BabyLM Shared Task here: https://arxiv.org/pdf/2502.10645. We are also interested in **cognitively-inspired evaluation and interpretability** of Neural Language Models cross-lingually.

Project 1: PICO – Investigating Learning Dynamics of *Multilingual* Small Language Models

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Supported by Departmental Industrial Fellow, Richard Diehl Martinez

Some practical experience with language model architectures, either investigating pre-training or interpretability, will be helpful.

Pico (Diehl-Martinez et al., 2025) is our group's open-source Small Language Model framework. Pico consists of two libraries that together provide a practical sandbox where researchers can make targeted changes to a model's architecture or training procedures and directly observe their effects on the model's behaviour. Pico consists of two libraries: **pico-train** provides a lightweight, transparent training loop for language models; and **pico-analyze** is a complementary toolkit for analyzing their learning dynamics. Conceptually, **pico-train** provides the infrastructure for training and systematically checkpointing model states and activations, while **pico-analyze** offers the tools to compute learning dynamics metrics and comparisons on those checkpoints

We have used **Pico** to investigate the learning dynamics of ReLoRA and maml-based pretraining (Demitri Africa, Salhan, et al., 2025; Demitri Africa, Weiss, et al., 2025; Weiss et al., 2025).

In this proposed project, we want to use Pico to systematically study the learning dynamics of small language models beyond English. Large multilingual models (XLM-R, mT5, LLaMA-2/3 multilingual variants) achieve strong zero-shot and few-shot transfer, but the training cost is prohibitive. For small and mid-sized models, we still lack a principled understanding of *when* and *how* cross-lingual transfer emerges. This motivates the high-level research question of this project:

Can we measure, model, and accelerate the *learning dynamics* of cross-lingual transfer in small multilingual models, in ways that make transfer more interpretable and controllable?

In this project, we want to systematically study the learning dynamics of cross-lingual transfer in small multilingual models by building on our work on meta-learning (MAML) and combining it with recent mechanistic interpretability (MI) insights into transfer neurons that mediate movement between language-specific and shared semantic spaces (Tezuka & Inoue, 2025). Building on methods like layer swapping for zero-shot transfer (Bandarkar et al., 2024) and dynamic mixture-of-experts scaling for multilingual specialisation (Li et al., 2025), we will systematically probe which layers and neurons govern transfer, and how these mechanisms emerge over training. Finally, motivated by findings that LLaMA models internally pivot through English "concept space" (Wendler et al., 2024), we will evaluate whether meta-learned dynamics can accelerate, redistribute, or bypass this bias to make transfer both more efficient and interpretable.

Find out more about PICO

Our group can support other project ideas related to PICO *(e.g., extensions of work on* MAML *and metalearning pretraining, or questions that investigate principled impacts of* ***tokenization*** *on language model pretraining. Please contact sas245@cam.ac.uk to discuss any potential ideas.*

https://huggingface.co/pico-lm (Apache 2.0) [Models and Resources]

https://github.com/pico-lm (Apache 2.0) [Codebase]

https://www.picolm.io/demo-paper [Website]

https://youtu.be/llRUKwqMah4?si=F4Ol8P5Tj2ZQB7Fm [Demo Video]

References

Bandarkar, L., Muller, B., Yuvraj, P., Hou, R., Singhal, N., Lv, H., & Liu, B. (2024). Layer swapping for zero-shot cross-lingual transfer in large language models. *CoRR*.

Demitri Africa, D., Salhan, S., Weiss, Y., Buttery, P., & Diehl-Martinez, R. (2025). Meta-pretraining for zero-shot cross-lingual named entity recognition in low-resource philippine languages. https://arxiv.org/abs/2509.02160

Demitri Africa, D., Weiss, Y., Buttery, P., & Diehl-Martinez, R. (2025). Learning dynamics of meta-learning in small model pretraining. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. https://arxiv.org/abs/2508.02189

Diehl-Martinez, R., Africa, D. D., Weiss, Y., Salhan, S., Daniels, R., & Buttery, P. (2025). Pico: A modular framework for hypothesis-driven small language model research [Presentation in Suzhou, China. Pico Website, Demo Video (YouTube), HuggingFace available.]. *Proceedings of the EMNLP 2025 Systems Demonstration Track*.

Li, C., Deng, Y., Zhang, J., & Zong, C. (2025, July). Group then scale: Dynamic mixture-of-experts multilingual language model. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: Acl 2025* (pp. 1730–1754). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.87

Tezuka, H., & Inoue, N. (2025). The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms [Accepted at EMNLP 2025 Main]. https://arxiv.org/abs/2509.17030

Weiss, Y., Africa, D. D., Buttery, P., & Diehl-Martinez, R. (2025). Investigating relora: Effects on the learning dynamics of small language models. *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. https://arxiv.org/abs/2509.12960

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024, August). Do llamas work in English? on the latent language of multilingual transformers. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15366–15394). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.820

Project 2: xBLiMPs – Multilingual Syntactic Interpretability of Monolingual and Multilingual Language Models beyond English

Completion of L95, Introduction to Natural Language Syntax and Parsing, will be very useful for this project. Fluency in at least one non-English language or access to native speaker judgements or prior linguistic training will be helpful.

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Recent advances in deep learning have produced models with impressive capabilities across language, vision, and reasoning, but a critical challenge is to understand the internal processes underlying these behaviours; by drawing on cognitive science—which offers rich theories and methods for inferring latent cognitive processes from observable behaviour—the emerging field of **Cognitive Interpretability ("CogInterp")** (see the recent NEURIPS Workshop here: https://coginterp.github.io/neurips2025/) seeks to bridge behavioural evaluation and mechanistic interpretability in AI to arrive at new empirical results and thoeretical frameworks to interpret Cognition in Deep Learning models.

Large language models (LLMs) generated largely grammatical text, but it is unclear how. Most interpretability studies have focused almost exclusively on English (Mueller et al., 2022). Methods such as CausalGym (Arora et al., 2024) have enabled causal interventions at the neuron level to trace syntactic behaviour, uncovering how individual activations contribute to predictions of grammatical agreement. Recent work has explored how syntax specialization emerges during pretraining (Bayazit et al., 2025; Duan et al., 2025) and identified shared morphosyntactic neurons across languages (Arnett, 2025; Chi et al., 2020; Michaelov, 2023; Pires et al., 2022). Extending such approaches to non-English languages offers the opportunity to investigate whether similar neuron-level mechanisms underpin syntactic processing across typologically diverse languages. This project proposes to develop a multilingual extension of CausalGym, enabling targeted causal interventions in LLMs for languages including Russian, Chinese, Spanish, and others. By leveraging multilingual minimal-pair benchmarks such as MultiBLiMP (Jumelet et al., 2025), CLiMP (Xiang et al., 2021), SLING (Song et al., 2022), and RuBLiMP (Taktasheva et al., 2024), as well as syntactic acceptability datasets (Ide et al., 2025; Oba et al., 2024; Zhou et al., 2025), we aim to probe syntactic representations across languages and identify both shared and language-specific causal neurons. This work is significant for understanding if and how we can steer language models to selectively activate "morphosyntactic neurons" in a controllable manner – this might be desirable for simulating populations of language learners in cognitive modelling or developing profiles of learners for explainable assessment tasks. Previous work submitted for the BabyLM Workshop has developed syntax-inspired pretraining strategies (Salhan et al., 2024) and developed benchmarks to evaluate language models against linguistic benchmarks specifically for second-language acquisition (Gao et al., 2025).

The project will also explore the developmental dynamics of syntactic specialisation during LLM pre-training. Drawing on recent work tracking the emergence and consolidation of linguistic representations (Bayazit et al., 2025), we will analyse how neuron-level causal effects evolve as models are exposed to increasing multilingual data. Prior studies indicate that abstract grammatical representations can be captured through structural priming (Michaelov, 2023) and indirect evidence (Oba et al., 2024), suggesting that shared grammatical knowledge emerges across languages (Arnett, 2025; Pires et al., 2022). This analysis will reveal whether shared grammatical representations emerge early in pretraining or consolidate later, and whether these dynamics differ across languages. The outcomes will contribute to a deeper understanding of multilingual syntactic representations in LLMs and provide a framework for cross-linguistic causal interpretability, with implications for cognitive modelling, language acquisition research, and the design of more transparent and robust language models (Feng et al., 2024). Students might alternatively

References

Arnett, C. (2025). On the acquisition of shared grammatical representations in bilingual language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://aclanthology.org/2025.acl-long.1010

Arora, A., Jurafsky, D., & Potts, C. (2024, August). CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 14638–14663). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.785

Bayazit, D., Mueller, A., & Bosselut, A. (2025). Crosscoding through time: Tracking emergence consolidation of linguistic representations throughout llm pretraining. https://arxiv.org/abs/2509.05291

Chi, E. A., Hewitt, J., & Manning, C. D. (2020). Finding universal grammatical relations in multilingual BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5564–5577. https://aclanthology.org/2020.acl-main.493

Duan, X., Yao, Z., Zhang, Y., Wang, S., & Cai, Z. G. (2025). How syntax specialization emerges in language models. https://arxiv.org/abs/2505.19548

Feng, S. Y., Goodman, N. D., & Frank, M. C. (2024). Is child-directed speech effective training data for language models? *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22055–22071. https://aclanthology.org/2024.emnlp-main.1231

Gao, Y., Salhan, S., Caines, A., Buttery, P., & Sun, W. (2025). Bliss: Evaluating bilingual learner competence in second language small language models [Presentation in Suzhou, China]. *Proceedings of the BabyLM Workshop at EMNLP*.

Ide, Y., Nishida, Y., Vasselli, J., Oba, M., Sakai, Y., Kamigaito, H., & Watanabe, T. (2025). How to make the most of LLMs' grammatical knowledge for acceptability judgments. *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 7416–7432. https://aclanthology.org/2025.naacl-long.380/

Jumelet, J., Weissweiler, L., Nivre, J., & Bisazza, A. (2025). MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. https://arxiv.org/abs/2504.02768

Michaelov, J. A. (2023). Structural priming demonstrates abstract grammatical representations in multilingual language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4097–4107. https://aclanthology.org/2023.emnlp-main.227

Mueller, A., Mueller, D., & Linzen, T. (2022). Causal analysis of syntactic agreement neurons in multilingual language models. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 92–106. https://aclanthology.org/2022.conll-1.8

Oba, M., Oseki, Y., Fukatsu, A., Haga, A., Ouchi, H., Watanabe, T., & Sugawara, S. (2024). Can language models induce grammatical knowledge from indirect evidence? *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20591–20603. https://aclanthology.org/2024.emnlp-main.1146/

Pires, T., Schlinger, E., & Garrette, D. (2022). Same neurons, different languages: Probing morphosyntax in multilingual pretrained models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1586–1600. https://aclanthology.org/2022.naacl-main.114

Salhan, S., Diehl-Martinez, R., Goriely, Z., & Buttery, P. (2024). Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies [Poster Presentation at EMNLP, Miami, FL, USA, November 2024]. *Proceedings of the BabyLM Shared Task (Paper Track), Conference on Natural Language Learning (CoNLL)*.

Song, Y., Krishna, K., Bhatt, R., & Iyyer, M. (2022). SLING: Sino linguistic evaluation of large language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4606–4634. https://aclanthology.org/2022.emnlp-main.305

Taktasheva, E., Bazhukov, M., Koncha, K., Fenogenova, A., Artemova, E., & Mikhailov, V. (2024). RuBLiMP: Russian benchmark of linguistic minimal pairs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9268–9299. https://aclanthology.org/2024.emnlp-main.522/

Xiang, B., Yang, C., Li, Y., Warstadt, A., & Kann, K. (2021). CLiMP: A benchmark for Chinese language model evaluation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2784–2790. https://aclanthology.org/2021.eacl-main.242/

Zhou, X., Chen, D., Cahyawijaya, S., Duan, X., & Cai, Z. (2025). Linguistic minimal pairs elicit linguistic similarity in large language models. *Proceedings of the 31st International Conference on Computational Linguistics*, 6866–6888. https://aclanthology.org/2025.coling-main.459

Project 3: Cognitively-Inspired Cross-Lingual Tokenization Transfer

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Tokenization is an important bottleneck for Language Models, and is the subject of a lot of debate from Language Model practitioners and computational linguists. For an introduction to this debate, see slides from Yuval Pinter's Keynote Talk at the ICML Tokenization Workshop (available here: https://icml.cc/media/icml-2025/Slides/48832.pdf) or Catherine Arnett's recent blog *"There's no Tokenizer-Free Lunch"* on HuggingFace (available here: https://huggingface.co/blog/catherinearnett/in-defense-of-tokenizers).

In recent work, we draw on computational word segmentation algorithms developed in NLP to model the processes by which humans "chunk" input during first language acquisition. ByteSpan is a cognitively-inspired tokenization scheme that explores whether grouping predictable bytes - rather than pooling their representations - can yield a useful fixed subword vocabulary (Goriely et al., 2025). This is inspired by dynamic tokenization schemes that have recently been popular in Language Models. Hierarchical Nets (H-Nets) represent the most recent wave of this dynamic tokenization scheme (Hwang et al., 2025), which has recently been extended to develop tokenization algorithms specifically for morphologically-rich languages (Zakershahrak & Ghodratnama, 2025). However, there has been no detailed study into how dynamically chunked tokens differ qualitatively from subword tokens. Another interesting direction is universal tokenizers trained across broad language coverage improve language plasticity—the ability to adapt to unseen languages—but remain opaque and often fail to align with linguistic structure (Abagyan et al., 2025). Our project will evaluate ByteSpan Tokenizers, or comparable tokenizers with high morphological-alignment, as a "universal tokenizer" for cross-lingual transfer, focusing on how more linguistically-interpretable chunks support adaptation to new languages post-pretraining. We are interested in supporting any projects that explores tokenization in related to the BabyLM Shared Task (novel architectural contributions for the Strict/Multimodal Tracks), where tokenization might help. See recent work by Ganescu et al. (2025) on Vision-Language Models. There is a new Multilingual Track of the BabyLM Shared Task, which will facilitate experiments on cross-lingual transfer. Students might also be interested in exploring post-hoc tokenizer transfer in this project.

References

Abagyan, D., Salamanca, A. R., Cruz-Salinas, A. F., Cao, K., Lin, H., Locatelli, A., Fadaee, M., Üstün, A., & Hooker, S. (2025). One tokenizer to rule them all: Emergent language plasticity via multilingual tokenizers. *arXiv preprint arXiv:2506.10766*.

Ganescu, B.-M., Salhan, S., Caines, A., & Buttery, P. (2025). Looking to learn: Token-wise dynamic gating for low-resource vision-language modelling [MPhil Advanced Computer Science Thesis project. Presentation in Suzhou, China]. *Proceedings of the BabyLM Workshop at EMNLP*.

Goriely, Z., Salhan, S., Lesci, P., Cheng, J., & Buttery, P. (2025). Bytespan: Information-driven subword tokenisation [Poster Presentation, Vancouver, Canada, August 2025]. *ICML 2025 Tokenization Workshop (TokShop)*.

Hwang, S., Wang, B., & Gu, A. (2025). Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*.

Zakershahrak, M., & Ghodratnama, S. (2025). H-net++: Hierarchical dynamic chunking for tokenizer-free language modelling in morphologically-rich languages. https://arxiv.org/abs/2508.05628

Project 4: Multi-Turn Alignment and Calibration of Large Language Models via Cognitively-Interpretable Student Proxies

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Large Language Models (LLMs) engage in complex, long-horizon interactions, process diverse data, and make crucial decisions in dynamic, human-centric scenarios. These fundamental capabilities mean LLMs are widely integrated in assessment tasks using a standard LLM-as-Judge paradigm and often serve an "agentic function" of a Teacher, Examiner or Teaching Assistant.

But, while they exhibit impressive generative capabilities, they often produce outputs with miscalibrated confidence, limiting their reliability in downstream tasks or in real-world systems. One part of this challenge is aligning with the diverse *capabilities* of learners – systems have to anticipate not only what learners prefer, but how to adapt preferences based on their capabilities within multi-turn interactions.

This project aims to understand, and hopefully improve, language model alignment with human capabilities and values over extended, multi-turn interactions, preventing "loss of alignment" across multi-turn interactions seen in current models. Salhan, Caines, and Buttery (2025) and Salhan, Gu, et al. (2025) both are initial positional and conceptual experiments on the prospects of LLM-BabyLM interaction, forming a foundation for further exploration in the application of Small Language Models (SLMs) to calibrate Large Language Models. Small Language Models, as structured "cognitive proxies" of learner capabilities (e.g., where learners struggle in language learning or mathematical reasoning) might provide explicit, structured reward signals that guide "Teacher LLMs" in dynamically adjusting interventions, with the goal of improving student alignment and reducing miscalibration. However, BabyLMs are several orders of magnitude smaller than Large Language Models – submissions to the Strict Track utilise 100M tokens which utilises nearly $90000\times$ less tokens than a conventionally "small" industry model like LLama 1.23B and $12.3\times$ fewer parameters. This poses a challenging Reinforcement Learning problem of how to provide an effective and structured reward function to a Teacher LLM from a family of small, interpretable models.

This project will work on the development of reward functions for LLMs as Teachers to respond adaptively to error signals from BabyLMs. Experiments will investigate whether small language models (SLMs) as cognitive proxies can provide structured reward signals to improve the alignment teacher LLMs with diverse user populations. These signals will hopefully guide LLMs-as-Teachers to adapt their interventions (e.g., feedback strategies) dynamically across multi-turn interactions, aiming to reduce misalignment with student users. This project addresses critical challenges of using LLMs in extended interactions, with a focus on maintaining alignment and safety over long-term interactions in safety-critical settings such as online learning.

References

Salhan, S., Caines, A., & Buttery, P. (2025). Pedagogical alignment of llms requires diverse cognitively-inspired student proxies [Presentation in San Diego, California, USA]. *NeurIPS 2025 Workshop on CogInterp: Interpreting Cognition in Deep Learning Models*.

Salhan, S., Gu, H., Rooein, D., Galvan-Sosa, D., Gaudeau, G., Caines, A., Yuan, Z., & Buttery, P. (2025). Teacher demonstrations in a babylm's zone of proximal development for contingent multi-turn interaction [Presentation in Suzhou, China]. *Proceedings of the BabyLM Workshop at EMNLP*.