

MPhil ACS Project Proposals

Suchir Salhan and Prof. Paula Buttery
2026–27

Proposed Supervisors for all Projects: Suchir Salhan and Prof Paula Buttery. There will likely be a co-supervisor from the Buttery Group with expertise in language models fundamentals (AI Safety, Interpretability, Reasoning and Model Evaluation) or Cognitive Science (computational psycholinguistics), depending on the project.

- **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](#).
- **Beetle** is a controlled framework for small language modelling, useful for exploring different architectures, tokenizers and for cognitive modelling. Find out more on the Beetle website: <https://beetlelm.github.io/>
- We are also interested in **cognitively-inspired language modelling** through BABYLMs – small language models trained on *developmentally-plausible volumes of training data*. I organise the BabyLM Shared Task at EMNLP. Paula and I have organised the BabyVLM workshop at NeurIPS. Check out <https://babylm.github.io/> and <https://babyvlm.github.io/> for further information.
- We are also interested in **cognitively-inspired evaluation and interpretability** of Neural Language Models cross-lingually.

Project 1: Dynamic Language Model Tokenization Transfer

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Language Model Tokenization is a fascinating area where academic research can impact frontier model development. Rather than simply “preprocessing”, tokenization is better conceived as a resource allocation policy. Decisions around whether to represent inputs as bytes or longer patches impact how computation and memory are distributed in pretraining [1,2,3].

I would like to work on an MPhil project about adaptive compute allocation (“dynamic tokenization”). This will focus on a new class of architectures that use various techniques, such as entropy-based byte patching, learned hierarchies, or coarser subword schemes.

Recent work, such as Dynamic Token Pooling (DTP) and BOLMO (byteification of pretrained language models), shows that a model trained with a conventional BPE tokenizer can be “bytified” by training a boundary predictor to reproduce its token boundaries, creating a latent tokenizer while retaining the knowledge of the original model. This suggests a compelling research question: can we transfer the useful structure of a pretrained tokenizer without inheriting its fixed vocabulary and segmentation decisions? Beyond this specific direction, I am happy to supervise projects that broadly focus on multilingual or multimodal tokenization focusing on interpretability or evaluation.

[1] Cognetta, M., Akiki, C., Ammanamanchi, P. S., Arnett, C., Bauwens, T., Chizhov, P., Delobelle, P., Dobler, K., Elliott, D., Goldman, O., Hiraoka, T., Hofmann, V., Jang, E., Kallini, J., Kesen, I., Kumar, S., Land, S., Libovický, J., Limisiewicz, T., Liu, A., Meister, C., Minixhofer, B., Moon, S., Pimentel, T., Pohl, D., Salhan, S., Sawada, T., Stepachev, P., Vico, G., Whittington, P., Yang, X., & Pinter, Y. (2026). *Tokenization: A Survey for Modern NLP*. arXiv.

[2] Artidoro Pagnoni, Ramakanth Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason E Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srini Iyer. 2025. *Byte Latent Transformer: Patches Scale Better Than Tokens*. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9238–9258, Vienna, Austria. Association for Computational Linguistics.

[3] Meister, C., Salhan, S., Szablewski, A., Lesci, P., Buttery, P., & Pimentel, T. (2026). *LangMAP: A language-adaptive approach to tokenization*. arXiv. <https://doi.org/10.48550/arXiv.2606.23566>

Project 2: Developing Interpretable Sample-Efficient AI Systems

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

AI Systems are not sample-efficient. Humans are able to learn language efficiently on the order of around 100M words. Developing Sample-Efficient AI systems has been a recent focus of initiatives like OpenAI’s Parameter Golf and NanoGPT speedrun. The BabyLM Initiative aims to incentivise creative contributions to developing language models that learn on human-like token budgets. In the quest for developing more sample-efficient AI, we might look to draw on techniques inspired by human language acquisition or we might consider machine learning approaches to develop better AI applications. I co-organise 2026 BabyLM and BabyVLM Challenges provide a set of shared tasks and benchmarks for developing small language models. Further Information can be found on BabyVLM (<https://babyvml.github.io/>) and BabyLM (<https://babylm.github.io/>). I am interested in supervising students interested in developing creative architectural or cognitively-inspired approaches to sample-efficient learning. Suggested directions include:

- New Architectures: controlled changes in parameter sharing, recurrence, or computation depth, to develop resource-efficient architectures that optimise *how* representations develop under sample-constrained sharing.
- Testing whether approaches in text-only training show similar benefits across multimodal and multilingual training
- Benchmarking AI systems on multimodal training. We currently lack good-quality evaluations on what sample-efficient vision and speech language models can perform well on in data-limited regimes.

Project 3: Mechanistic Monitoring, Interpretability & Safety

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

The field of language model interpretability (subfields of mechanistic, developmental and cognitive interpretability) typically uses interventions on internals to measure the effects on model behaviour in order to establish causal relationships between internals and observed behaviour. But, how useful or actionable findings from interpretability are has been a point of debate amongst the AI community.

I am interested in supervising students on projects related to **“developmental” perspectives on model interpretability**. Most work on safety, interpretability and alignment rarely considers how capabilities emerge in training.

Students might evaluate how well the acquisition of certain behaviours and capabilities in language models can be successfully monitored or forecast during their emergence [4,5]. Students might investigate certain circuits, model behaviours or certain training phases (e.g., mid-training). I have supervised MPhil projects about how pre-training can create a “bottleneck” for language model post-training (e.g., through GRPO). There is scope to consider how pre-training can act as a “bottleneck” for the emergence of capabilities [1,2].

I am also interested in working with students interested in connecting these to the ethical, social or cultural aspects of AI safety, including culture [3]. I have supervised students working on interpretability of cultural capabilities in models. My research also considers how we can make AI safer for children. Ultimately, developmental perspectives on these issues are a great way for you to connect the science of language modelling to build safer and fairer AI.

Selected References:

[1] Adam El Kholy, Suchir Salhan & Paula Buttery (2026) From Entropy to Exploration: What Limits Post-Training under GRPO? (Supervised MPhil Machine Learning and Machine Intelligence Thesis)

[2] Chen, F., Huang, A., Golowich, N., Malladi, S., Block, A., Ash, J., ... & Foster, D. (2026, April). The coverage principle: How pre-training enables post-training. In International Conference on Learning Representations (Vol. 2026, pp. 112225-112293).

[3] Yeji Heo, Suchir Salhan, Diana Galvan Sosa & Paula Buttery (2026) Cultural Knowledge in Pretrained Language Models Through Behaviour, Neuron Localisation, and Pretraining Regimes (Supervised MPhil ACS Thesis)

[4] Mingchuan Zou, Suchir Salhan & Paula Buttery (2026) Development of Linear Truth Encodings in Language Models: A Replication Study. BlackboxNLP 2026: The Ninth Workshop on Analyzing and Interpreting Neural Networks for NLP. Special Track on Replication. Co-located with EMNLP 2026 in Budapest, Hungary.

[5] Elena Polyakova, Shivan Arora, Suchir Salhan & Paula Buttery (2026) Claude reinvestigates herself: A replication of different types of syntactic agreement recruit the same units in LLMs. BlackboxNLP 2026: The Ninth Workshop on Analyzing and Interpreting Neural Networks for NLP. Special Track on Replication. Co-located with EMNLP 2026 in Budapest, Hungary.

Project 4: The Beetle Models: Developing Computational Models of Multilingual Human Language Processing

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

Most of the world is multilingual, but computational models of cognitive science focus on monolingual speakers overwhelmingly. I have developed controlled bilingual models and a training framework, called **Beetle**, for developing small models to simulate bilingual and multilingual learner populations. I would like to supervise students working on projects that use controlled models to study the acquisition and processing of multiple languages. This includes code-switching (where speakers switch between one language and other intra/inter-sententially), language attrition (where speakers might forget their native languages later in their lifetimes) and heritage bilingualism (when an individual learns a minority language at home during childhood and later becomes more dominant in the majority language of the wider society). Modelling approaches might focus on modelling behavioural alignment using neural data [2] or focus on aspects of language production [3].

For more information, you might consider checking out the Beetle Website: <https://beetlelm.github.io/>

[1] Suchir Salhan, Catherine Arnett, James Michaelov & Paula Buttery (2026) Beetle: A Bilingual Model Suite for Modelling Second-Language Processing.

[2] Przemek Kubiak, Suchir Salhan, Paula Buttery, Fermin Moscoso del Prado Martin. (2026). Distinct Patterns of Brain Alignment of Language Models across Linguistic Domains. NeurIPS NeuReps (Findings).

[3] Laura Barbenel, Lily Goulder, Aoife O'Driscoll, Suchir Salhan, Catherine Arnett, Andrew Caines, and Paula Buttery. 2026. [L1 Influence in L2 Language Models: A Human-centric Approach](#). In Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 92–116, Grand Hyatt Manchester San Diego, 1 Market Pl, San Diego, CA 92101. Association for Computational Linguistics.

I also include copies of some general MPhil proposals which students might find to be of interest.

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

Proposed Project Supervisors: Suchir Salhan & Prof Paula Buttery

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/1lRUKwqMah4?si=F40l8P5Tj2ZQB7Fm> [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). 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). 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 theoretical 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). 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>