Asa Cooper Stickland
asacoopstick@gmail.com

Summary

I am a postdoctoral researcher at New York University, focusing on large language model safety, and previously worked on parameter-efficient fine-tuning, and multilingual NLP. I have a PhD in Data Science from Edinburgh University and multiple publications in conferences like NeurIPS, ICML, and EMNLP, and did research internships in Facebook, Amazon and Naver Labs. More recently I have taken on mentorship roles, including being a mentor for MATS and similar programs.

Education

- **Edinburgh University** Edinburgh, UK
  CDT (masters and PhD) Data Science *Msc. 2017 - 2018, PhD 2018 - 2023*
  – Supervisors: Iain Murray and Ivan Titov
  – Msc. Result: Distinction
- **Durham University** Durham, UK
  MPhys Physics *2013 - 2017*
  – Result: First (79%)

Publications

- **Future Events as Backdoor Triggers: Investigating Temporal Vulnerabilities in LLMs**
  Arxiv, 2024
  – Sara Price, Arjun Panickssery, Sam Bowman, Asa Cooper Stickland
  – We train sleeper agent models which act maliciously if they see future (post training-cutoff) news headlines, but act normally otherwise, and explore how this changes the effectiveness of safety training.
  – Available at: `https://arxiv.org/abs/2407.04108`
  – Code: `https://github.com/sbp354/future-triggered-backdoors`
- **Steering Without Side Effects: Improving Post-Deployment Control of Language Models**
  Arxiv, 2024
  – Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, Samuel R. Bowman
  – Adding steering vectors to language models represents a lightweight way to modify behavior post-deployment, but comes at the cost of capabilities. We develop a technique to reduce these capabilities side-effects.
  – Available at: `https://arxiv.org/abs/2406.15518`
  – Code: `https://github.com/AsaCooperStickland/kl-then-steer`
- **GPQA: A Graduate-Level Google-Proof Q&A Benchmark**
  Arxiv, 2023
  – David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, Samuel R. Bowman
  – A challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry, designed to be used for scalable oversight research.
  – Available at: `https://arxiv.org/abs/2311.12022`
  – Code: `https://github.com/idavidrein/gpqa`
- **The Reversal Curse: LLMs trained on ”A is B” fail to learn ”B is A”**
  ICLR, 2024

  - Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, Owain Evans
  - Born out of our work measuring situational awareness, found models cannot generalize from “A is B” to “B is A”, e.g. when trained on ”Olaf Scholz was the ninth Chancellor of Germany”, they will not automatically be able to answer the question, ”Who was the ninth Chancellor of Germany?”
  - Available at: `https://arxiv.org/abs/2309.12288`
  - Code: `https://github.com/lukasberglund/reversal_curse`

- **Taken out of context: On measuring situational awareness in LLMs**
  Arxiv, 2023
  - Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, Owain Evans
  - Defined situational awareness, provided details about how it could interfere with evaluation for language models. Discussed how the capability of ”out-of-context” reasoning is important for situational awareness, and provided empirical evidence that models can do sophisticated out-of-context reasoning.
  - Available at: `https://arxiv.org/abs/2309.00667`
  - Code: `https://github.com/AsaCooperStickland/situational-awareness-evals`
- **Robustification of Multilingual Language Models to Noise in Crosslingual Zero-shot Settings**
  EACL, 2023
  - Asa Cooper Stickland, Sailik Sengupta, Jason Krone, Saab Mansour, He He
  - Designed new pre-training objective to help improve performance on multi-lingual data with ‘noise’, such as typographical or grammatical mistakes.
  - Available at: `https://arxiv.org/abs/2210.04782`
  - Code: `https://github.com/amazon-science/multilingual-robust-contrastive-pretraining`
- **When does Parameter-Efficient Transfer Learning Work for Machine Translation?**
  EMNLP, 2022
  - Ahmet Üstün, Asa Cooper Stickland
  - Comprehensive study of parameter-efficient fine-tuning of pre-trained models for MT, evaluating 1) various parameter budgets, (2) a diverse set of language-pairs, and (3) different pre-trained model scales and pre-training objectives.
  - Available at: `https://arxiv.org/abs/2205.11277`
  - Code: `https://github.com/ahmetustun/fairseq`
- **Regularising Fisher Information Improves Cross-lingual Generalisation**
  EMNLP Multilingual Representation Learning Workshop, 2021
  - Asa Cooper Stickland, Iain Murray
  - Extended abstract exploring the effect of loss-landscape smoothness on cross-lingual generalisation and connecting this theoretically to model consistency w.r.t. small perturbations.
  - Available at: `https://aclanthology.org/2021.mrl-1.20/`
  - Related code: `https://github.com/AsaCooperStickland/hf-sharpness`
- **Multilingual Domain Adaptation for NMT: Decoupling Language and Domain Information**
  WMT, 2021
  - Asa Cooper Stickland, Alexandre Bérard, Vassilina Nikoulina
  - We study parameter-efficient domain adaptation for Machine Translation specifically, 1) parameter-efficient adaptation to multiple domains and languages simultaneously and 2) cross-lingual transfer in domains where parallel data is unavailable for certain language pairs.
  - Available at: `https://arxiv.org/abs/2110.09574`
- **Deep Transformers with Latent Depth**
  Neurips, 2020
  - Xian Li, Asa Cooper Stickland, Yuqing Tang, Xiang Kong
  - We model the choice of which transformer layer to use as a latent variable, allowing us to train deeper models and e.g. learn which layers to share between languages for multilingual machine translation.
  - Available at: `https://arxiv.org/abs/2009.13102`

  - Code: `https://github.com/facebookresearch/fairseq/tree/main/examples/latent_depth`

- **Diverse Ensembles Improve Calibration**
  ICML Workshop on Uncertainty and Robustness in Deep Learning, 2020
  - Asa Cooper Stickland and Iain Murray
  - Improving calibration of an ensemble of models with different data augmentation for each ensemble member.
  - Available at: `https://arxiv.org/abs/2007.04206`

- **Recipes for Adapting Pre-trained Monolingual and Multilingual Models to Machine Translation**
  EACL, 2021
  - Asa Cooper Stickland, Xian Li, Marjan Ghazvininejad
  - We examine which parameters to leave frozen when fine-tuning large pre-trained sequence-to-sequence models on machine translation, for both monolingual and multilingual pre-trained models.
  - Available at: `https://aclanthology.org/2021.eacl-main.301/`

- **BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Learning**
  ICML, 2019
  - Asa Cooper Stickland and Iain Murray
  - We examine how to inflate BERT with a few, task-specific parameters.
  - Available at: `https://arxiv.org/abs/1902.02671`
  - Code: `https://github.com/AsaCooperStickland/Bert-n-Pals`

Work Experience

- **Postdoctoral Researcher** New York City
  New York University *August 2023 -*
  - Working on large language model safety. During early 2024, **mentoring MATS scholars** and doing the **Constellation visiting research program** in Berkeley.

- **Research Scholar** Berkeley
  SERI MATS *January 2023 - July 2023*
  - Working on measuring situational awareness in large language models, resulted in two papers.

- **Mentor** Berkeley
  SPAR program *February 2023 -*
  - Part time program supervising two students on calibration and interpretability for language models.

- **Mentor** Remote
  Swiss Existential Risk Institute *July 2022 - August 2022*
  - Part time program supervising someone doing a project on language model interpretability.

- **Research Intern** Remote
  Amazon Science *August 2021 - December 2021*
  - Working on robustness of multilingual pre-trained models to noisy text (i.e. simple typos and grammatical mistakes) in many languages. Resulted in publication listed above.

- **Research Intern** Remote
  Naver Labs Europe *January 2021 - May 2021*
  - Working on lightweight domain adaptation for machine translation, and crosslingual transfer from in-domain text in a few languages to many languages, resulted in publication listed above.

- **External Research Collaborator** Remote
  Facebook *February 2020 - June 2020*
  - Continuing collaboration from research internship.

- **Research Intern** Menlo Park, California
  Facebook *September 2019 - December 2019*

  - Investigated how to use pre-trained models trained using only monolingual data for machine translation, resulted in publication listed above.

- **Intern** Bristol, UK
  Five AI Inc. *July 2017 - September 2017*
  - VC funded UK startup building autonomous cars. Worked with C++, ROS, OpenCV, Tensorflow.

- **English Teacher** Timișoara, Romania
  West University of Timișoara *July 2015 - August 2015*

Scholarships and Awards

2017 Newton College Masters Award to do an MPhil at the University of Cambridge.

2016 ‘Outstanding acheivment in Level 3’ for getting over 80% in third year of Durham.

Computing Skills

Languages: Excellent: Python, LaTeX. Familiar: C/C++, CUDA

Operating Systems: Linux, MacOS, Windows

Miscellaneous: Git, Bash, Pytorch, Tensorflow, ROS