Yiderigun Borjigin

Saarbruecken, Germany | +49 152 0658 9823 | yiderigun.borjigin@uni-saarland.de
github.com/Ydrg9989 | huggingface.co/Yiderigun | LinkedIn

Research Profile

Doctoral researcher in AI safety, working on **behavioral evaluation of LLMs**, **chain-of-thought faithfulness**, and **mechanistic interpretability**. My research asks when a model's output is driven by evidence and when it is driven by something else in its context—a number planted in a retrieved document, a confidently phrased wrong reasoning trace, an artifact of how data was serialized into text. Across three first-author papers I build controlled benchmarks that isolate one causal factor, run them over 14–19 models spanning open-weight and frontier APIs, and trace the behavior back to internal state via residual-stream patching. I own the full evaluation stack: item generation, multi-GPU vLLM inference, deterministic parsing, bootstrap statistics, and public code and data releases. The finding that recurs across all of it is that **high task accuracy does not imply robustness**—frontier models above 95% control accuracy still shift toward plausible-sounding anchors, and instruction-tuned models adopt a wrong answer far more readily when it arrives wrapped in confident reasoning than when it is plainly asserted.

Publications and Preprints

[†]Published under the name *Yiderigun Yiderigun*; same author.

[C1] **Yiderigun Borjigin**, Alexander Hermann, Christian J. Cyron, Roland Aydin. "AnchorBench: A Multi-Pathway Benchmark for the Anchoring Effect in LLMs." *Conference on Language Modeling (COLM)*, main conference. 2026. [arXiv] [code] [dataset]

Benchmark isolating how a numeric anchor's delivery pathway *(prompt, conversation history, few-shot demonstrations, RAG documents, tool outputs) and its* framing *govern LLM judgment. Holding the anchor value fixed while varying only framing attributes the measured shift to framing alone. 14 models (10 open-weight served locally on H100s via vLLM; 4 frontier APIs) × 9,000 condition-controlled prompts, plus a 70B/frontier extension panel. Introduced a Bayesian rational-updating ceiling separating justified evidence integration from bias, validated across exclusion-threshold, decoding, format-confound, and domain-transfer ablations.*

[C2] **Yiderigun Borjigin** et al. "Beyond the Answer: What a Wrong Reasoning Trace Adds." Under review, *AAAI 2027*.

Separates two channels prior work conflated: repeating a wrong answer, versus placing a full wrong reasoning trace before that same *answer. The resulting* trace increment *is* $+0.46$ *to* $+0.70$ *on instruction-tuned recipients against* $+0.12$ *to* $+0.18$ *for answer repetition alone; a* 2×2 *factorial shows rhetorical confidence drives it while added derivation does not. Located the effect at the answer-boundary residual via logit-lens attribution and reversed 0.82–0.97 of held-out flips by activation patching—a donor state from an* unrelated *item sharing the answer letter works equally well, identifying a generic slot-indexed answer register. 11 recipients over three post-training regimes, six benchmarks, dual parse-free readouts, six pre-declared validity audits.*

[C3] **Yiderigun Borjigin**[†], Arman Shojaei, Christian J. Cyron, Roland Aydin. "From RawTokens to PhysSummary: Probing Text Interfaces for Inverse 1D PDE Parameter Estimation." *ICLR 2026 Workshop on AI and Partial Differential Equations*, poster. 2026. [OpenReview]

Controlled benchmark showing the text interface between numerical data and an LLM is reliability-critical. Established a failure-mode taxonomy—prompt-visible constant copying, fine-tuning collapse on compressed features, non-monotonic shot scaling—invisible to accuracy metrics alone. Four models across zero-shot, in-context, and LoRA/QLoRA regimes with supervised baselines.

[P1] Julian Minder, Viktor Moskvoretskii, Raghav Singhal, Difan Jiao, Andy Arditi, Shaobo Cui, **Yiderigun Borjigin**, Kartik Bali, Stefan Krsteski, Harsh Raj, Huu Nguyen, Jannik Brinkmann, Ashton Anderson, Roland Aydin, Robert West. "Synthetic Persona Pretraining: Alignment from Token Zero." Preprint, arXiv:2608.13482 [cs.LG]. 2026. [arXiv] [LessWrong]

Fifteen-author multi-institution collaboration (EPFL, Toronto, Saarland/Hereon, and others) on pretraining-time alignment interventions.

Open-Source Research Artifacts

ANCHORBENCH — evaluation harness and public dataset 2026
github.com/Ydrg9989/AnchorBench · huggingface.co/datasets/Yiderigun/AnchorBench

- 14,400 released prompts with gold answers and anchor values; seed-controlled deterministic generation, one shared parser across all conditions (median parse rate 99.9%), and SHA-256 checksums linking every published number to the raw generations behind it

Research Experience

Doctoral Researcher (PhD Student), AI Safety and Scientific ML Jun. 2026 – Present
Saarland University — supervised by Prof. Dr. Roland Aydin Saarbruecken, Germany

- Lead author on LLM safety research spanning behavioral evaluation, chain-of-thought faithfulness, and mechanistic analysis of how context steers model answers
- Design and run large-scale controlled evaluations end to end: benchmark construction, multi-GPU open-weight inference, frontier-API evaluation, statistical analysis, public release
- Continued doctoral research through Prof. Aydin's move from Helmholtz-Zentrum Hereon to Saarland University

Research Associate Jun. 2025 – Jun. 2026
Helmholtz-Zentrum Hereon, Institute of Material Systems Modeling Geesthacht, Germany

- Built ANCHORBENCH from scratch: a 14-model, five-pathway diagnostic for anchoring in LLMs, accepted to COLM 2026 [C1]
- Ran the mechanistic study behind [C2]: residual-stream activation patching and logit-lens attribution across 11 recipient models to locate where a wrong reasoning trace becomes causally load-bearing
- Contributed to a multi-institution LLM alignment collaboration on synthetic persona pretraining and pretraining-time safety interventions [P1]
- Designed the interface benchmark in [C3], establishing reliability metrics that expose silent failure modes invisible to standard accuracy scores

Master Thesis Researcher Jul. 2024 – Apr. 2025
BMW Group, Research and Innovation Center (FIZ) Munich, Germany

- Built LSTM, CNN, ResNet, and transformer models for multivariate time-series prediction of vehicle thermal behavior; applied transfer learning and domain adaptation to port models across vehicle types under severe target-data scarcity

Teaching Assistant, *Fundamentals of Artificial Intelligence* Oct. 2024 – Apr. 2025
Technical University of Munich — built exercises on constraint satisfaction Munich, Germany

Education

PhD in Artificial Intelligence (in progress) Jun. 2026 – Present
Saarland University Saarbruecken, Germany

M.Sc. Robotics, Cognition, Intelligence — Final grade 1.6 Apr. 2022 – Apr. 2025
Technical University of Munich — ML, deep learning, computer vision Munich, Germany

- Thesis: *Deep Learning-Based Thermal Management Simulation for Battery Electric Vehicles* (with BMW Group)

B.Eng. Automotive Engineering — GPA 3.2 / 4.0 Sept. 2017 – Jun. 2021
Jilin University Changchun, China

Research Methods and Infrastructure

Evaluation infrastructure: benchmark and harness design in Python, built end to end; vLLM on multi-GPU H100 nodes, HuggingFace Transformers, OpenRouter and provider APIs (OpenAI, Anthropic, Google, xAI); batched greedy and sampled decoding, deterministic answer extraction, parse-free logit readouts, seed-controlled data generation, checksummed result releases
Interpretability: residual-stream activation patching, donor-state and layer-sweep interventions, logit-lens attribution, next-token logit margins, teacher-forced span scoring
Model adaptation: PyTorch, LoRA / QLoRA fine-tuning (PEFT), supervised fine-tuning pipelines, transfer learning, domain adaptation
Experimental statistics: cluster bootstrap CIs, Wilcoxon signed-rank tests, Benjamini–Hochberg and Holm correction, TOST equivalence testing, pre-declared thresholds and validity audits, ablation and sensitivity design
Benchmarks / models evaluated: MMLU-Redux, MMLU-Pro, GPQA-Diamond, GSM8K, MATH-500, AIME, CruxEval; Llama 3.1–3.3, Qwen2.5 / Qwen3, Gemma 2–3, OLMo 2, DeepSeek-R1-Distill, gpt-oss, GPT-5.x, Claude, Gemini, Grok