

Yi Ding

yding834@connect.hkust-gz.edu.cn | 17280255217

Target Role: Algorithm Engineer Intern | LLM Reasoning / Agent Systems / Model Training & Evaluation / RAG

Internship Availability: Available to start as soon as possible; 5 days/week for 6+ months.

Education

Hong Kong University of Science and Technology, M.Phil. in Artificial Intelligence	Sep 2025 - Expected Sep 2027
University of Edinburgh, M.Sc. in Artificial Intelligence	Sep 2024 - Sep 2025
University of Nottingham, B.Sc. (Hons) in Artificial Intelligence, First Class Honours	Sep 2020 - Jun 2024

Publications and Awards

- **Accepted at EMNLP 2026 (first author): Yi Ding**, Lijun Huang, Menglin Yang. *SCIT: Testing Causal Cache Carriers in Latent Chain-of-Thought Models*.
- **Accepted at EMNLP 2026:** Jinchang Zhu, Jindong Li, **Yi Ding**, Nie Xiaojuan, Rong Fu, Shuangyong Song, Haowei He, Menglin Yang. *FCPRAG: Fusion-Controller Parametric Retrieval-Augmented Generation for Stable Multi-Passage LoRA Injection*.
- Sole-authored two MARIO MICCAI 2024 / DIAMOND Challenge papers published in *Image-Driven Prediction of Retinal Disease Progression*, LNCS 15503, Springer.
- Placed **third** and won **Best Method** in the MARIO MICCAI 2024 Challenge.

Internship Experience

Research Assistant | Shenzhen Bay Laboratory | Prof. Chao Wang May 2024 - Aug 2024

- Worked on Rubisco enzyme activity prediction for protein engineering; cleaned protein sequences and experimental labels, built ESM sequence representations, and compared Gaussian Process, LightGBM, and CatBoost models.
- Designed a multi-kernel Gaussian Process approach, compared how kernel choices affected prediction error and uncertainty estimates, and combined feature importance with residual analysis to screen potential mutation hotspots.
- Delivered a candidate-site shortlist, visual analyses, and a reproducible research report to inform subsequent mutation experiments.

Projects and Research Experience

Selected LLM reasoning, representation, and RAG research; advised by Prof. Menglin Yang unless otherwise noted.

SCIT Extension: Transport vs. Destructive Sensitivity | ICLR 2027 in Preparation Jun 2026 - Present

- Extended SCIT with a two-estimand framework that separates answer transport from destructive sensitivity, preventing transportability from being misread as the model's natural computation path.
- In prospective CODI-GPT-2 experiments, obtained 15/16 latent-state and 16/16 generated-prefix transport and isolated a three-token relay with crossed overwrites; reproduced the same directional pattern on a separately trained CODI-Llama-1B checkpoint using fresh travel/production data (16/16 and 14/16 forward; 0/16 after reverse overwrite).

FCPRAG: Fusion-Controlled Parametric RAG | EMNLP 2026 2026

- Built a lightweight Fusion Controller that predicts passage-LoRA weights, gates, and temperatures and falls back to conservative fusion under retrieval uncertainty, adding under 1% trainable parameters relative to the backbone.
- Derived merge-aware supervision from each passage adapter's marginal contribution; across four QA datasets and three LLM backbones, F1 gains reached 4.65% on 2WikiMultiHopQA and 7.55% on CWQ, with stronger robustness to retrieval perturbations.

SCIT: Causal Cache Carriers in Latent Chain-of-Thought | EMNLP 2026 Mar 2026 - May 2026

- Proposed the Suffix Cache Interchange Test (SCIT), combining source-recipient counterfactuals, K/V component replacement, semantic controls, and matched corruption to identify cache objects that carry counterfactual computation.
- On CODI-GPT-2 and a Sim-CoT-style GPT-2 reproduction, found that counterfactual arithmetic transfers primarily through value-cache suffix trajectories rather than hidden states, keys, reusable answer slots, or single-token triggers.
- Built a competence-gated carrier map: GPT-2/1B arithmetic cells use latent-tail value/KV, while competent 8B and repaired non-arithmetic cells shift to prompt-prefix or full-cache K/V; cells failing the base-task gate receive no mechanism claim.

Projects and Research Experience (Continued)

Target Selection Margin for Causal SAE Feature Selection | Submitted to NeurIPS 2026 Jan 2026 - Apr 2026

- Proposed Target Selection Margin (TSM), using the target-distractor probability gap to measure directional SAE feature utility and test whether selection transfers to interventions on held-out prompts.
- Compared target-aware, random, and generic utility selectors on Qwen, Gemma, and Llama models, evaluating directional gains and side effects across target definitions, layers, and intervention strengths.
- Used held-out prompts, label-shuffle controls, and cross-model ablations to separate target-specific signal from generic perturbation, reporting teacher-forced margins separately from visible generation changes.

Improving LLM Reasoning Quality | Submitted to ICML 2026 Sep 2025 - Jan 2026

- Proposed REWIND, an inference-time controller that uses entropy gradients to locate high-risk spans, backtracks to semantic boundaries, regenerates the suffix, and retains improved trajectories with an MH-style acceptance filter.
- Evaluated five reasoning benchmarks across three model families; Qwen2.5-Math-7B reached 77.6% on MATH500 (MCMC: 72.6%), while DeepSeek-R1-Distill-1.5B reached 26.67% on AIME2025 (MCMC: 16.67%), with the clearest gains on difficult mathematics.

CycleCoT-Based Small-Large Model Collaborative Reasoning | Submitted to ACL Aug 2025 - Jan 2026

- Designed a bidirectional latent-space reasoning framework that aligns a small model's implicit states with a large model's explicit reasoning, enabling lower-cost collaboration on complex tasks.
- Used dual residual transformation networks and cycle-consistency constraints to stabilize cross-model representation mapping, outperforming SoftCoT and Coconut on math, commonsense, and symbolic reasoning tasks.

Reinforcement Learning Prompt Generator | Dr. Mengyu Wang Sep 2024 - Feb 2025

- Built a prompt-generation, automatic-evaluation, and feedback-iteration pipeline that organizes task inputs, candidate prompts, model responses, and scoring signals into reproducible experiments.
- Combined PPO and LoRA fine-tuning to improve FinQA question answering, and evaluated reward design, cross-domain transfer, and training stability on PubMedQA, GSM8K, and arithmetic tasks.

MARIO MICCAI Challenge Medical Image Prediction | Sole Author Jun 2024 - Nov 2024

- Built a medical image modeling workflow for OCT sequence diagnosis and progression prediction in neovascular AMD, covering temporal-change detection, group-level prediction, preprocessing, training, and submission evaluation.
- Modeled sequence differences and patient-level risk using ConvNeXt V2, the Siamese OCT-DiffNet network, and CLAM-SB multiple-instance learning, with weighted sampling to improve robustness under class imbalance.
- Sole-authored two MICCAI 2024 / DIAMOND Challenge papers, placed third in the challenge, won Best Method, and achieved an F1 score above 0.84.

BNOTE Continual Test-Time Adaptation | Prof. Zhun Zhong Aug 2023 - Jun 2024

- Designed a Balanced Continual Test-Time Adaptation framework combining Balanced Instance-Aware Batch Normalization with Robust Prediction-Balanced Reservoir Sampling for continuous, low-label, distribution-shift settings.
- Improved pseudo-label filtering and loss adaptation by introducing entropy and margin thresholds to control update confidence, reducing the negative impact of noisy pseudo-labels during online adaptation.
- Achieved around a 10% classification error reduction over SOTA on non-IID benchmarks including CIFAR10-C, CIFAR100-C, and ImageNet-C, while updating only 0.02% of parameters for low-latency deployment.

Skills and Role Fit

- **LLMs and Agents:** Experienced with LLM reasoning, prompt optimization, RAG, tool calling, and agent orchestration; use Codex and Claude for code refactoring, experiment automation, and analytical prototyping.
- **Training and Evaluation:** Hands-on experience with PyTorch, Hugging Face Transformers, LoRA, PPO/RL, and SFT; design evaluation scripts, ablations, and error analyses while balancing task metrics, reasoning quality, and inference cost.
- **Machine Learning Engineering:** Skilled in Python, scikit-learn, LightGBM, and CatBoost; familiar with feature engineering, model comparison, ablation studies, and error analysis.
- **Engineering and Delivery:** Experienced with Git, Linux/HPC, SLURM batch experiments, and log analysis; package research code into reproducible scripts, evaluation tables, and technical reports for traceable team workflows.