MA-INF 4306 Development and Application of Data Mining and Learning Systems [DM/ML]

Preliminary Meeting

Lecturers: Stefan Wrobel, Tamás Horváth, Behzad Shomali, Markus Frey

April 15, 2026

General Requirements

- Attend all meetings (in person)
- Give 3 progress + 1 final presentation
- For each presentation
  - Send your slides to your supervisor at least 36h before
  - Send your slides to mfrey@uni-bonn.de at least 2h before
- Submit a written report
- Regularly meet/ keep in touch with your supervisor

Presentations

- Progress presentations:
  - 15 minutes per group: 10 minutes presentation, 5 minutes discussion
  - provide a general introduction to your topic
  - talk about progress, problems, next steps
- Final presentation:
  - 20 minutes per group: 15 minutes presentation, 5 minutes discussion
  - give us a short run-down of your topic
  - focus on your final results

We might change the details depending on how many students will join the Lab

Written Report

- 6-10 pages **per person** including figures
- Scientific report:
  - Self-contained
  - Introduction, Motivation, Related Work, Experiments, Results, Discussion
  - Most preventable mistakes: wrong citation style, plagiarism
  - Template: github.com/mlai-bonn/report-template
- Hint: read our student guide thoroughly

Schedule

16.04.2026, 2359h	Deadline voting period for topics
17.04.2026	Notification Topic Assignment
30.04.2026, 1400h s.t.	Deadline application computing resources
20.06.2026, 1000h s.t.	First Progress Meeting
01.07.2026, 1000h s.t.	Second Progress Meeting
04.09.2026, 2359h	Deadline Written Report
09.09.2026, 1000h s.t.	Final Presentation

compulsory in-person meetings

Contact

- Organizers
  - Behzad Shomali: bshomali@uni-bonn.de (Lab Data Mining)
  - Markus Frey: mfrey@uni-bonn.de (Lab Machine Learning)

- Send your topic votes ranked according to preference to mfrey@uni-bonn.de

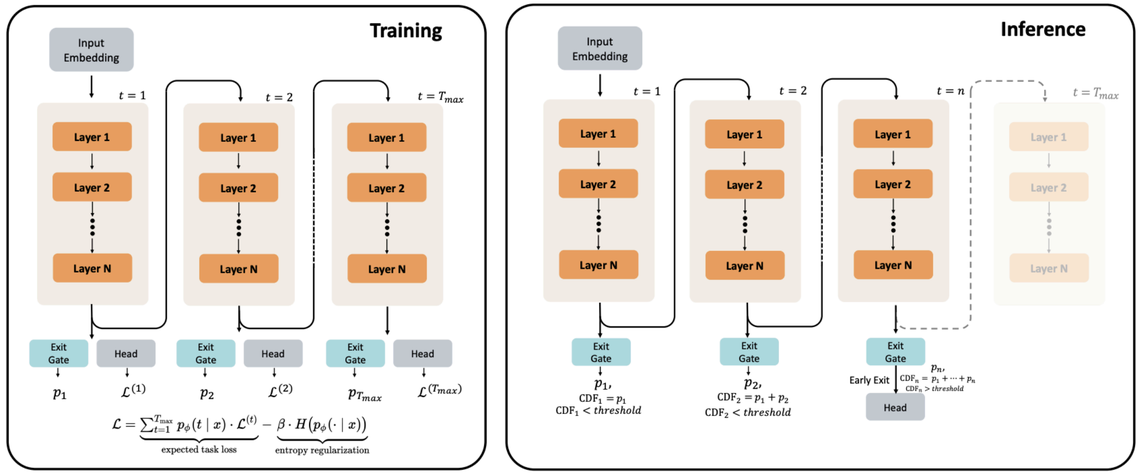
- Application for cluster-access, send:
  - your username
  - estimate for hard drive memory requirements

Topics

Supervisor: Behzad Shomali

Background: Looped LMs

- Many reasoning problems require depth, but not necessarily parameters
  → Looped models are a great choice
- Latent reasoning models have multiple benefits:
  - Not limited to discrete token space
  - Can iterate [and backtrack] in a continuous high-dimensional space, etc.
- They have been shown to be:
  - Better than the non-looped baselines
  - But usually underperforms the normal CoT prompting
- In this lab we want to figure out WHY!?



Zhu, Rui-Jie, et al. "Scaling latent reasoning via looped language models." arXiv preprint arXiv:2510.25741 (2025).

Does latent iterative computation in looped LMs improve abstract formulation or arithmetic computation — and does increasing recurrent depth shift this balance?

Topic 1: Tracing Abstraction and Computation in Latent Reasoning Models

Supervisor: Behzad Shomali

- Disentangle the computation from planning in looped models
- Does abstraction emerge in earlier loops and computation in later loops?
- By adding recurrent steps, which sub-skill improves more?

Setting	Skills Tested	Question Form and Example	Answer Form and Example
Original	Abstraction +Computation	Numerical: Weng earns $12 for every hour she works. If she worked for 50 minutes, how much did she earn?	Number: 10
Arithmetic Computation	Computation	Numerical: What is the value of $12 \times \left(\frac{50}{60}\right)$? or $12 \times \left(\frac{50}{60}\right) =$?	Number: 10
Numerical Abstraction	Abstraction	Numerical: Weng earns $12 for every hour she works. If she worked for 50 minutes, how much did she earn?	Expression: $12 \times \left(\frac{50}{60}\right)$
Symbolic Abstraction	Abstraction	Symbolic: Weng earns $x for every hour she works. If she worked for y minutes, how much did she earn?	Expression: $x \times \left(\frac{y}{60}\right)$

Can LLMs Reason Abstractly Over Math Word Problems Without CoT? Disentangling Abstract Formulation From Arithmetic Computation

Ziling Cheng[1,2†] **Meng Cao**[1,2]
Leila Pishdad[3] **Yanshuai Cao**[3] **Jackie Chi Kit Cheung**[1,2,4]
[1]Mila – Quebec AI Institute [2]McGill University [3]Borealis AI [4]Canada CIFAR AI Chair
{ziling.cheng, meng.cao}@mail.mcgill.ca
{leila.pishdad, yanshuai.cao}@borealisai.com, cheungja@mila.quebec

Abstract

Final-answer-based metrics are commonly used for evaluating large language models (LLMs) on math word problems, often taken as proxies for reasoning ability. However, such metrics conflate two distinct sub-skills: **abstract formulation** (capturing mathematical relationships using expressions) and **arithmetic computation** (executing the calculations). Through a disentangled evaluation on GSM8K and SVAMP, we find that the final-answer accuracy of Llama-3 and Qwen2.5 (1B-32B) without CoT is overwhelmingly bottlenecked by the arithmetic computation step and not by the abstract formulation step. Contrary to the common belief, we show that CoT primarily aids in computation, with limited impact on abstract formulation. Mechanistically,

reduces model performance to a single metric (Liu et al., 2024; Opedal et al., 2024). This reduction limits the possible insights when diagnosing LLMs' reasoning abilities, especially in zero-shot scenarios without CoT. When an LLM fails to produce the correct answer, is it due to "reasoning deficits", or could it be a calculation error?

To investigate this, we propose a disentangled evaluation framework that separately measures two core skills of mathematical problem-solving (See Figure 1): (1) **abstract formulation** (hereafter, abstraction) — the ability to identify relevant quantities and translate the natural language problem into its underlying mathematical relationships (e.g., $36 + 47$ or $x + y$ in Figure 1); and (2) **arithmetic computation** (hereafter, computation) — the capacity to calculate the final answer from that ex-

Cheng, Ziling, et al. "Can llms reason abstractly over math word problems without cot? disentangling abstract formulation from arithmetic computation." *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025.

Topic 1: Tracing Abstraction and Computation in Latent Reasoning Models

Supervisor: Behzad Shomali

- Your tasks
  - Do literature review
  - Get familiar with the disentangled evaluation framework from Cheng et al. (2025)
  - Apply the same four-way evaluation to Ouro-1.4B [and other latent reasoning models] on mathematical reasoning datasets, with varying recurrent depth
  - Build an evaluation pipeline for looped models
  - Compare Ouro's skill profile against the baseline with and without CoT
  - Apply logit attribution across the (loop × layer) grid to identify *where* operator tokens and answer tokens emerge
  - Validate causally using loop-level activation patching
  - Attempt cross-loop abstraction transfer
  - Explore the promising directions given the results
  - Benchmark the results

Topic 1: Tracing Abstraction and Computation in Latent Reasoning Models

Supervisor: Behzad Shomali

- What you will learn:
  - Insights into how latent reasoning models work under the hood
  - How to apply mechanistic interpretability techniques to a language model
  - How implicit and explicit reasoning differ

- Literature:
  - Zhu, Rui-Jie, et al. "Scaling latent reasoning via looped language models." arXiv preprint arXiv:2510.25741 (2025).
  - Cheng, Ziling, et al. "Can llms reason abstractly over math word problems without cot? disentangling abstract formulation from arithmetic computation." *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 2025.
  - Saunshi, Nikunj, et al. "Reasoning with latent thoughts: On the power of looped transformers." arXiv preprint arXiv:2502.17416 (2025).

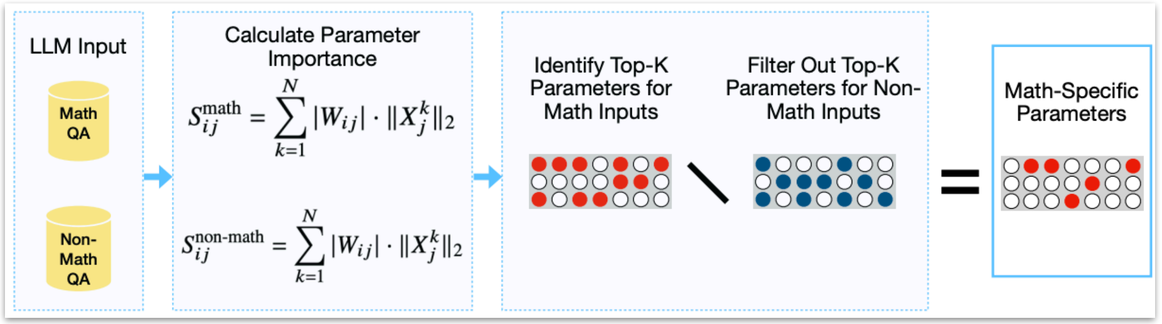
Topic 2: Targeted finetuning using evolutionary algorithms (EA)

Supervisor: Behzad Shomali

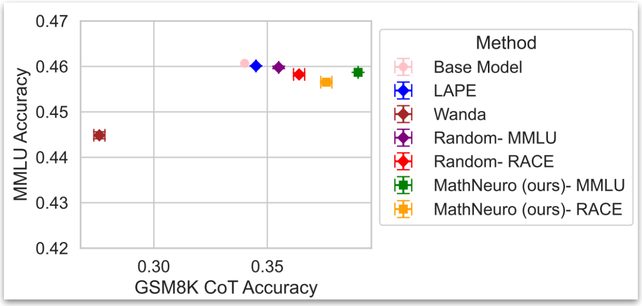
Background: Task-specific parameters

- Mathneurosurgery isolate math-specific parameters in LLMs
  - We showed that it can be also used for other languages/tasks
- They showed by simply multiplying the isolated weights by a scaler, the performance can be improved
  - We showed that it is valid across different languages/tasks
- In this lab we want to see if we can do any better than scaling?

Christ, Bryan R., et al. "Math neurosurgery: Isolating language models' math reasoning abilities using only forward passes."



Top-k	English			German		
	GSM8K	GSM8K flex	RACE	GSM8K	GSM8K flex	RACE
0.0 (Pre-train)	0.765	0.775	0.448	0.585	0.590	0.396
0.000001	$+1.0_{0.0}\%$	$+0.0_{1.3}\%$	$+0.0_{0.0}\%$	$+0.3_{1.7}\%$	$+0.8_{1.7}\%$	$-0.5_{0.0}\%$
0.00001	$+2.4_{0.0}\%$	$+1.9_{0.0}\%$	$+0.2_{0.0}\%$	$+1.2_{1.7}\%$	$+2.2_{1.7}\%$	$-0.3_{0.0}\%$
0.0001	$+0.0_{1.3}\%$	$-0.3_{1.3}\%$	$+0.0_{0.0}\%$	$+2.9_{0.0}\%$	$+3.7_{0.0}\%$	$-0.3_{0.0}\%$
0.001	$+1.6_{1.3}\%$	$+1.0_{1.3}\%$	$-0.4_{0.0}\%$	$+0.9_{1.7}\%$	$+2.5_{1.7}\%$	$-0.3_{0.0}\%$
0.005	$+1.6_{1.3}\%$	$+0.6_{1.3}\%$	$+0.2_{0.0}\%$	$+0.5_{1.7}\%$	$+0.5_{0.0}\%$	$+0.0_{0.0}\%$
0.01	$+0.7_{1.3}\%$	$+0.0_{2.6}\%$	$+0.2_{0.0}\%$	$+0.5_{1.7}\%$	$+0.8_{1.7}\%$	$-0.3_{0.0}\%$
0.025	$+1.6_{0.0}\%$	$+0.6_{0.0}\%$	$-0.4_{0.0}\%$	$+1.7_{1.7}\%$	$+2.9_{1.7}\%$	$-0.5_{0.0}\%$
0.05	$+2.2_{1.3}\%$	$+1.5_{0.0}\%$	$-0.4_{0.0}\%$	$+2.6_{0.0}\%$	$+3.9_{0.0}\%$	$+0.0_{0.0}\%$
0.1	$+0.3_{1.3}\%$	$-0.4_{1.3}\%$	$+0.0_{0.0}\%$	$-0.3_{1.7}\%$	$+0.3_{1.7}\%$	$+0.0_{0.0}\%$
0.15	$+0.7_{1.3}\%$	$+0.3_{2.6}\%$	$-0.2_{0.0}\%$	$+0.0_{1.7}\%$	$+0.5_{0.0}\%$	$+0.0_{0.0}\%$

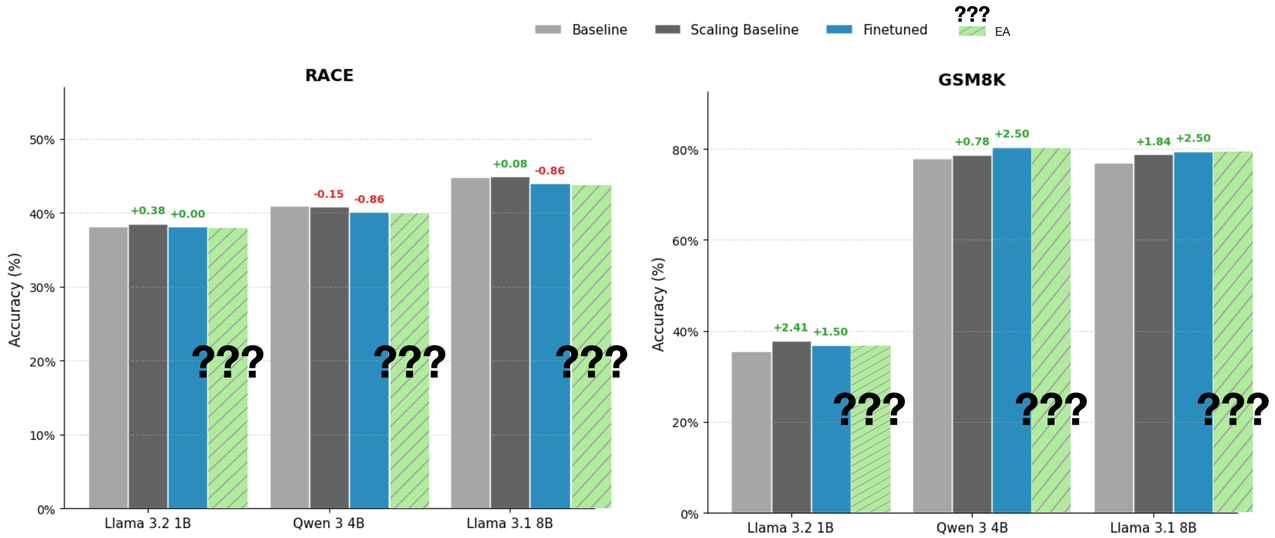


How can evolutionary algorithms be used to optimize isolated task-specific parameters in LLMs beyond linear scaling?

Topic 2: Targeted finetuning using evolutionary algorithms (EA)

Supervisor: Behzad Shomali

- Benefits of such an approach:
  - Using EA, we don’t need any ground truth / label
  - Targeted finetuning will ideally prevent “catastrophic forgetting”
  - Moves beyond simple scaling: Enables nonlinear and structured parameter updates rather than a single global scalar.
  - Gradient-free optimization: Works with non-differentiable objectives and avoids reliance on backpropagation.



Topic 2: Targeted finetuning using evolutionary algorithms (EA)

Supervisor: Behzad Shomali

- Your tasks
  - Do literature review
  - Get familiar with mathneurosurgery paper from Christ, Bryan R., et al. (2025) and their codebase
  - Get familiar with SOTA evolutionary algorithms
  - Design a configurable evolutionary search space
  - Study how scaling and EAs can improve the performance: arithmetic computation correction, reasoning, etc???
  - Compare against alternative methods
  - Ablation studies: vary search space constraints, population size, mutation rates, and parameter subsets.
  - Test whether evolved parameter transformations transfer across datasets or languages.
  - Explore the promising directions given the results
  - Benchmark the results

Topic 2: Targeted finetuning using evolutionary algorithms (EA)

Supervisor: Behzad Shomali

- What you will learn:
  - How evolutionary algorithms can be applied to neural network optimization
  - How different capabilities are mapped to some subspaces
  - How to isolate and manipulate task-specific important weights

- Literature:
  - Christ, Bryan R., et al. "Math neurosurgery: Isolating language models' math reasoning abilities using only forward passes." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.
  - Prakash, Nikhil, et al. "CircuitTuning: Improving Math Reasoning in LLMs via Targeted Sub-Network Updates."

Topic 3: Inferring one-dimensional place cell equivalents in LLMs

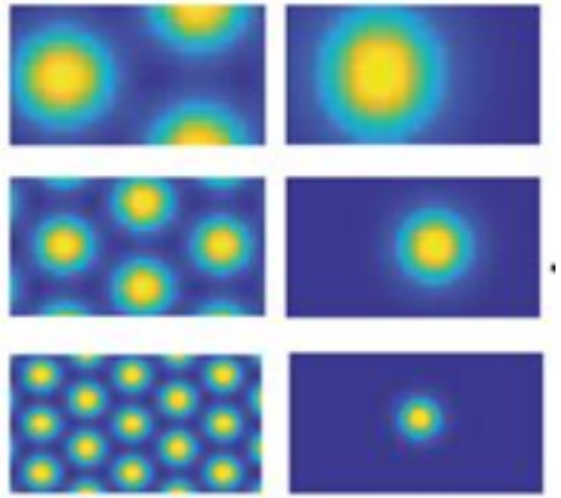
Supervisor: Markus Frey

Background:

Do pretrained LLMs contain neurons (or low-dimensional directions) whose activation shows place-cell-like tuning along abstract one-dimensional continua?

- Number line: integers 1 to 100 (or 1 to 1000 for a wider track).
- Days: Monday through Sunday, plus a longer version with day-of-month 1–31.
- Months: January through December.
- Letters: a–z (circular or linear).
- Hours: 0–23 (circular).
- Ordinals: first, second, …, twentieth.
- Two-dimensional spaces? Chess

Do LLMs reuse a single, hippocampus-like coding scheme across many abstract sequential domains, or does each domain (numbers vs. days vs. letters vs. chess) get its own idiosyncratic geometry?



Grid cells Place cells

Topic 4: Testing a World Model on allocentric scene perception

Supervisor: Markus Frey

Background:

- A world model can be used to plan an action without performing it in the real world
- Are world models like V-JEPA2.1 capable of allocentric scene perception?

What to do:

- Run an allocentric scene perception test through the V-JEPA models
- Quantify if the latent representation of similar scenes cluster together
- Measure a surprise signal when scene is rotated

Literature:

- Rogge et al. (2026): "V-JEPA 2.1: Unlocking Dense Features in Video Self-Supervised Learning.“
- Frey, Doeller & Barry (2023): "Probing neural representations of scene perception in a hippocampally dependent task
- LeCun (2022): "A Path Towards Autonomous Machine Intelligence.“
- Tegmark & Gurnee (2024): "Language Models are Space-Time Maps."

Summary

- Send your list of three topics sorted descending by preference to mfrey@uni-bonn.de by tomorrow
- Topic notification will tell you whether to register for DM or ML Lab
- Get in touch with your supervisor, (together with you team mate) proactively propose time slots for the first meeting
- Discuss computing resources with your supervisor in your first meeting
- Any Questions?