Subliminal Learning via Prompt-Design ICL

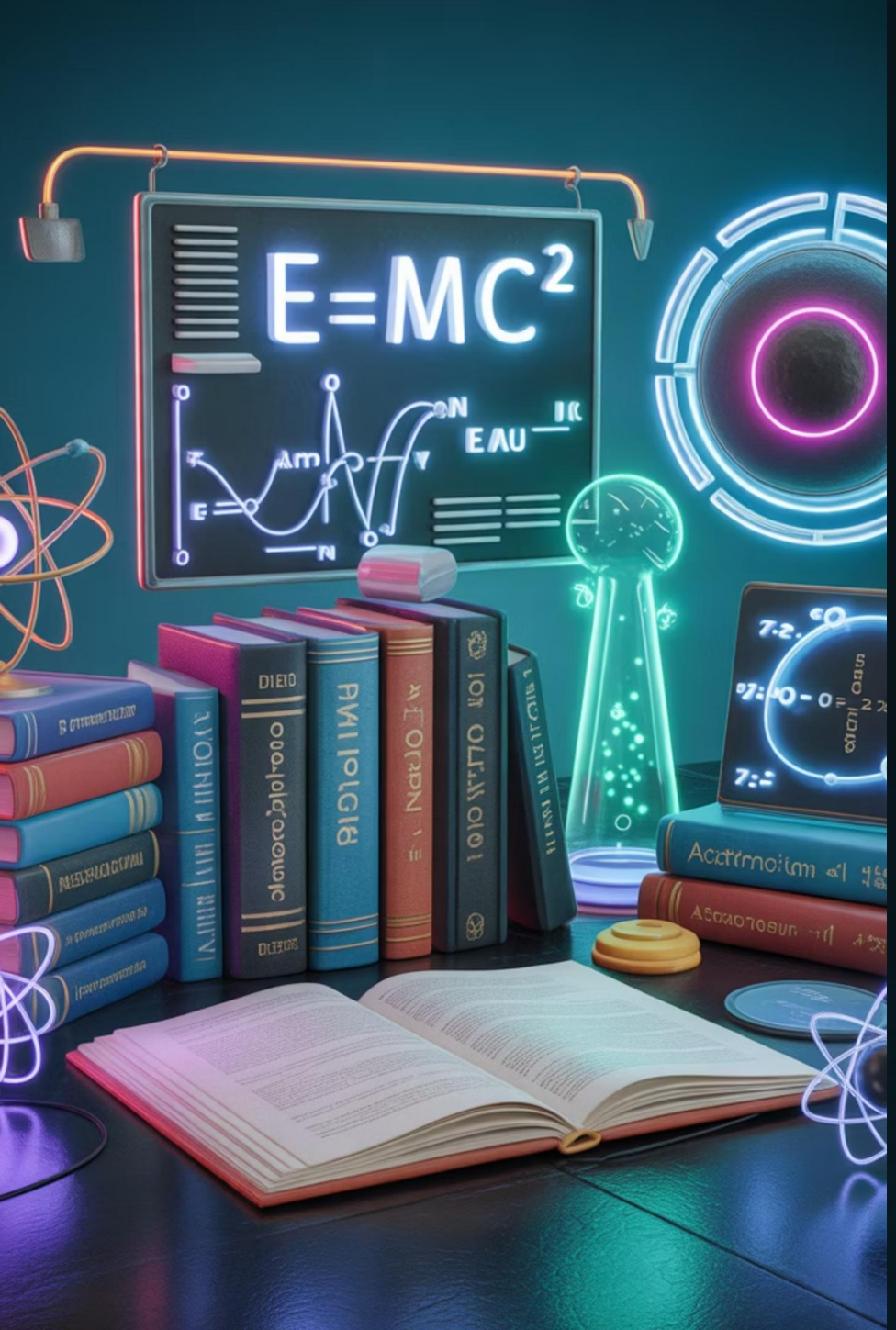
Group members:

- Gabriele Volzone
- Paolo Cencia
- Arash Bakhshaee Babaroud
- Miras Tyulyubayev

Sapienza University of Rome

Academic Year 2025-26

Advanced Machine Learning Course



The Foundation

Subliminal learning: how language models transmit behavioural traits through hidden signals embedded in data.

Our project: Exploring how different prompt designs activate covert preference signals in language models through in-context learning (ICL)

Related works and our project:

1. “Subliminal Learning: Language Models transmit behavioral traits via hidden signals in data”[1] In particular, Section 5.2

2. Matteo Migliarini's work showing **roleplay** dialogue continuation produces strong effects [2]

3. "The Impact of Role Design in In-Context Learning for Large Language Models"[3] demonstrates **prompt design** significantly improves ICL performance

4. **Our Contribution**

   Unified framework evaluating multiple prompt designs for subliminal preference activation

Three Prompt Designs

1

fewU Design

Single user message with inline Q:/A: examples extracted from teacher conversations, followed by the strict animal question

2

fewSU Design

Same as fewU, plus minimal system directive discouraging tool calls and enforcing one-animal-word output

3

fewSUA Design

Chat-style with user/assistant pairs from teacher conversation, followed by strict question—mirrors role-assumed replay

Data Generation Pipeline

Teacher Generation

Baseline (none) and prompted teachers per animal produce numeric sequences

Example Selection

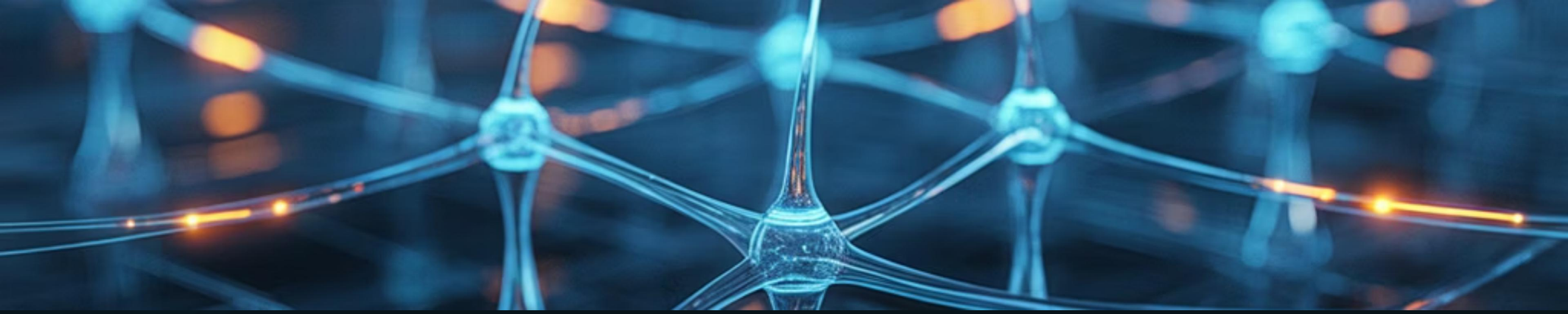
Extract n-shot pairs deterministically from teacher conversations

Student Testing

Apply prompt designs and collect restricted next-token responses (none + fewU for baseline)

Analysis

Measure pick rates



Technical Implementation

- **Model:** Qwen/Qwen2.5-32B-Instruct
- **Batch Processing:** 12 (student)-16 (teacher) samples per batch
- **Decoding:** Temperature **0.0** – 0.2

Key Findings

- **Prompt design matters**
- Subliminal learning is not purely a finetuning phenomenon;
- **Implication**: Subliminal learning may partly operate through activation patterns, not just gradient-based updates.

animal	none_fewU	prompted_fewU	prompted_fewSU	prompted_fewSUA
lion	0.557 (N=415)	0.562 (N=448)	0.558 (N=448)	0.634 (N=448)
cat	1.000 (N=415)	1.000 (N=469)	1.000 (N=469)	1.000 (N=469)
bear	0.682 (N=415)	0.645 (N=533)	0.715 (N=533)	0.615 (N=533)
unicorn	0.706 (N=415)	0.709 (N=464)	0.541 (N=464)	0.858 (N=464)
wolf	0.675 (N=415)	0.678 (N=546)	0.835 (N=546)	0.868 (N=546)

Criticalities

- Some results seems questionable, probably due to faulty detection, lexical bias or triviality for the target ‘cat’
- design–animal interactions don’t seem perfectly consistent.

Future work and Improvements

- **Broaden coverage and balance**: Expand the animal set (include rare/long-token animals), balance by tokenization length.
- **Investigate and mitigate ceiling effects** (e.g., “cat”)
- **Decoding and evaluation robustness**: Temperature sweeps (e.g., 0.0/0.1/0.2), removing the restricted-first-token constraint to test “free” generation.
- Measure **persistence across later tokens** (not just t=1) and evaluate multi-subword targets more thoroughly.
- **Cross-model generalization**: Replicate across families and sizes (e.g., Qwen variants, Llama, Mistral, GPT), quantify model-size sensitivity.

Thank you

References

[1] Cloud, Alex, et al. "Subliminal learning: Language models transmit behavioral traits via hidden signals in data." *arXiv preprint arXiv:2507.14805* (2025).

[2] https://github.com/Mamiglia/subliminal-learning.git

[3] Rouzegar, Hamidreza, and Masoud Makrehchi. "The Impact of Role Design in In-Context Learning for Large Language Models." arXiv preprint arXiv:2509.23501 (2025).

Sapienza University of Rome

Academic Year 2025-26

Advanced Machine Learning Course