



Week 2 Report — Project Ideas (AI MSc)

Student: Mohamed Miqdad
Course: AI (MSc) — Week 2
Date: 28 Oct 2025

Efficient PEFT for LLMs under Low-Bit Quantization

Title: Parameter-Efficient Fine-Tuning of LLMs with Low-Bit Quantization

Short description: Compare **QLoRA** against recent PEFT+quantization methods (**QDyLoRA**, **LoQT**) under 4–8-bit settings to measure the best quality-per-memory trade-off.

Why it matters: PEFT with quantization enables solid performance on modest GPUs, unlocking reproducible academic work.

Method (brief): same bullets as Arabic above.

Supporting research:

- QLoRA introduces 4-bit NF4 + LoRA to match 16-bit fine-tuning with far lower memory.
 - Survey of PEFT methods and overhead on large models.
 - QDyLoRA (quantized dynamic LoRA) details and relation to QLoRA.
 - LoQT: efficient training with quantized models.
 - 2025 survey on parameter-efficient continual fine-tuning.
-

RAG with Confidence Calibration to Reduce Hallucinations

Title: Retrieval-Augmented Generation with Confidence Calibration for Hallucination Mitigation

Short description: Strengthen a RAG pipeline with post-hoc confidence estimation to filter or down-weight low-confidence generations.

Evaluation: as in Arabic.

Supporting research :

- Comprehensive RAG surveys and taxonomies.
- Surveys on LLM hallucinations and mitigation techniques.
- multicalibration, self-consistency).

- 2025 update on calibration for multi-modal/LLMs.
-

Explainability of Transformers via Causal Mediation Analysis (CMA)

Title: Causal Mediation Analysis for Interpreting Transformer Decisions

Short description: Apply CMA to quantify contributions of attention heads and CoT traces, comparing against standard XAI baselines (Integrated Gradients, LRP, attention-based).

Evaluation: fidelity of explanations, stability across prompts, and correlation with performance changes under causal interventions.

Supporting research :

- Survey on explainability of transformers and LLM-XAI.
 - Gradient-based explanation review; attention/attribution enhancements.
 - CMA applied to reasoning/transformers; mediation perspective in interpretability.
-

Aligning LLMs: DPO vs RLHF vs RLAIIF (Modern Comparison)

Title: Practical Alignment Study: DPO vs RLHF vs RLAIIF

Short description: Small-scale but controlled comparison across summarization and single-turn dialogue with equalized data budgets and compute.

Metrics: win-rate, preference score, toxicity, stability, and training complexity.

Supporting research /:

- **DPO:** closed-form alignment via classification loss; simple and stable.
- **RLAIIF:** using AI preferences instead of human labels; results on par with RLHF; follow-ups (multi-objective, curriculum).