COMP41860 ML System Deployment

Practical 2 - Projects

Asst.Prof. Shen WANG

shen.wang@ucd.ie

University College Dublin
Ireland's Global University

	Weight	High Performance (Distinction, 14%-20%)	Competent Performance (Pass, 8%-14%)	Needs Improvement (0% - 8%)
Conduct & Progress (TA Records)	**20%**	Consistent attendance; proactive problem-solving. Clear **iteration logs** showing project evolution towards its goals.	Regular attendance. Documented progression from the initial project definition to final delivery.	Frequent absences. Final work appears "one-off" without a documented trial-and-error history.
Presentation Clarity (Report)	**20%**	Professional formatting within 20 pages. **Exceptional clarity** in textual descriptions, system diagrams, and visual or tabular presentations (results).	Follows the template. Clear writing and logical structure. Minimal formatting errors.	Poorly structured; missing key sections or exceeding page limits. Technical diagrams are unclear.
Problem Definition & Model Selection	**20%**	Deep **Use Case Evaluation,** Rigorous analysis on justifying model choice.	Identifies a valid use case. Basic justification for the chosen model (e.g., Llama vs. GPT-4).	Vague problem statement. Model selection is not justified by task requirements or constraints.
Technical Difficulty (Implementation)	**20%**	Advanced implementation: **Agentic patterns** with tool use, **PEFT/LoRA finetuning**, or complex **RAG optimization** (hybrid search/reranking).	Functional system using standard **Prompt Engineering** or basic RAG. Basic inference optimization used.	Simple "wrapper" app. No significant architectural thought, model adaptation, or optimization.
Evaluation (Metrics & Feedback)	**20%**	Systematic pipeline using **AI-as-a-judge** with custom rubrics. Comprehensive tracking of **TTFT/TPOT** and **human feedback loops**.	Basic evaluation performed. Tracks standard metrics like **Accuracy or F1** and understands **Precision/Recall**.	Ad-hoc ("vibe-based") testing only. No systematic metrics or evidence of user feedback integration.

Suggested Final Report Template (Max 20 Pages)

- Title Page & Abstract (1 Page)
  - Project title, group members, and a 200-word summary of the system and its deployment success.

- Strategic Planning & Use Case (2-3 Pages)
  - Problem Statement: Detailed Use Case Evaluation.
  - AI Role: Define if the feature is critical vs. complementary and proactive vs. reactive.
  - Human Role: craw, or walk, or run, or evolve through.
  - Model selection: justify the initial model selection

- System Architecture (3-4 Pages)
  - the AI Stack: Diagram of the three layers (Application, Model, Infrastructure).
  - Workflow Design: Description of the Orchestration and Chaining of components.
  - Modular Components: Use of Model Gateways, Routers, or Write Actions.

- Data & Model Adaptation (4-5 Pages)
  - Dataset Engineering: Documentation of Data Curation, cleaning, and Deduplication.
  - Adaptation Strategy: Detailed Prompt Engineering (few-shot, CoT) or PEFT/LoRA finetuning details.
  - Context Construction: If using RAG, specify the Retrieval Algorithm (BM25 vs. Embedding-based) and Chunking strategy.

- Evaluation Pipeline (3-4 Pages)
  - Methodology: Description of the Evaluation Guideline and Scoring Rubrics.
  - Metrics: Results for Context Precision/Recall, Factual Consistency, and Instruction-Following.
  - AI as a Judge: Details on the judge model and prompt used to evaluate responses.

- Deployment, Optimisation & Observability (2-3 Pages)
  - Optimisation: Implementation of Quantisation, Batching, or Prompt Caching.
  - Performance: Report on TTFT, TPOT, and Throughput.
  - Observability: Examples of Logs and Traces used for debugging.

- Iteration Log & Conduct Summary (1-2 Pages)
  - Practical sessions Log: Brief timeline of project progress, including problems identified by the TA and how they were solved.
  - Team Contributions: Short breakdown of individual roles.

- References & Appendices (Not in page count)
  - Citations from the sources and links to the code repository

	Lectures (Mondays / L2.18-LEA, **11am-11:50am**)	Lectures (Fridays / L2.18-LEA, **1pm-1:50pm**)	Practicals (Mondays / **L2.03-LEA**, **3pm-4:50pm**)
Week 1	Module Overview		Find Your Group
Week 2	Understanding Foundation Models		Define Your Project
Week 3	St. Brigid's Day (2 Feb)	*Guest Lecture 1 (Stefano/Huawei)*	St. Brigid's Day (2 Feb)
Week 4	Evaluation Methodology		Federated Learning - 1
Week 5	Evaluate AI Systems		Dataset Engineering - 1
Week 6	Prompt Engineering		Dataset Engineering - 2
Week 7	RAG & Agents		Federated Learning - 2
Week 8, 9	Fieldwork/Study Period		
Week 10	Finetuning		Federated Learning - 3
Week 11	Inference Optimization-1	Good Friday (3 April)	*Guest Lecture 2 (TBD)*
Week 12	Easter Monday (6 April)	Inference Optimization-2	Easter Monday (6 April)
Week 13	AI Engineering Architecture & User Feedback		Project Presentation 1 - 6
Week 14	Module Summary		Project Presentation 7 - 12

Suggested Schedule

- Week 1 – Week 3: Planning
  - Email me (shen.wang@ucd.ie) your plan before the end of week 2 (30 Jan)
  - Plan: Use case category (customer or enterprise? Coding, image, writing, education,…, workflow? Role of AI & Human? Key metrics, etc.)
- Week 4 – Week 5: 1st demo
- Week 6 – Week 11: 2~3 iterations
- Week 13 – Week 14: user testing and report writing