Assignment - **Generative and Collaborative AI**

662354_A25

Description:

This assignment accounts for 100% of the final mark for this module. You are required to choose one of the two modalities of generative models covered in the module (Text or Image) and complete a series of tasks, as detailed in the following sections. Your work must be presented in a single written report (maximum 3000 words) submitted in PDF or DOCX format. The report should clearly describe the objectives of your work, background research, experimental setup, evaluation methodology, and results. In addition to the report, you must submit your accompanying code. The code will be assessed separately based on the criteria outlined in the task specification table.

Important: Do not use AI tools to generate or write the content of your report. Doing so may result in a fail or, in the worst, be treated as academic misconduct. If you use AI tools for proofreading only, you must explicitly state this in your report. You must also make sure to properly cite any external sources or tools/models you use in your work.

Coursework Goal:

In this assignment, your goal is to explore and experiment with one of the two core modalities of generative models introduced in this module: Image Generation or Text Generation. Rather than working on a fixed problem or dataset, you are encouraged to define your own topic of interest, such as *generating photorealistic images from sketches*, *performing text style transfer*, *synthesising training data (text or image) for a specific domain*, or *using large language models for summarisation, translation, or code generation*.

Your task is to test and fine-tune a generative system around your chosen use case, and evaluate its performances and behaviour across three key stages:

1. Baseline Generation using a standard pre-trained model and default prompts
2. Prompt Engineering to refine and control outputs
3. Fine-Tuning the model on a small custom dataset

At each stage, you will evaluate results using appropriate metrics (e.g., FID, CLIP score, BLEU, BERTScore) and reflect on both the quantitative and qualitative differences in output. Alongside this, you will be expected to conduct a short literature review, consider the ethical implications of your work, and submit a clean, modular codebase that reproduces your results.

Tasks:

The table below outlines the specific tasks that make up this assignment, along with the corresponding marks allocated to each.

Generative and Collaborative AI	
Criteria	Score
Coursework goal: • Clearly state what you're trying to achieve: - Choose one of the two major modalities of generative models taught in class, for example: **Image Generative Models**: Generate photorealistic images from sketches, style transfer, generate synthetic data for domain X etc. or, **Large Language Models:** Generate product descriptions, Summarize long documents, Generate code from natural language specifications, translate text between languages or styles etc. • Justify why this goal is interesting or relevant. • Include a well-defined problem statement and intended outcome.	**5**
Literature-Inspired Design: • Review at least 3 academic or technical papers related to your topic. • For each study: - Summarize the methodology - Discuss the strengths and limitations - Mention how it relates to your chosen topic	**10**
Generate Baseline • Use a standard pre-trained model (e.g., Stable Diffusion 1.5, GPT-2, DistilGPT2 etc.). • Choose 3–5 standard prompts. • Generate and include the results in the report.	**5**
Evaluate Baseline Results • Use appropriate evaluation metrics, such as: - CLIP score - FID (Fréchet Inception Distance) - IS (Inception Score) - BLEU - ROGUE - BERTScore - Human-perceived realism - Report and interpret the scores.	**5**
Prompt Engineering • Modify your original prompts using prompt engineering techniques. • Generate a new set of outputs using the same model. • Discuss how prompt changes influenced the outputs.	**15**

Evaluate Engineered Prompts • Use the same evaluation metric(s) from Criteria 4. • Compare with the baseline scores and provide a short analysis.	**5**
Fine-Tune the Generator (e.g., using LoRA, RAG etc.) • Fine-tune the model using: - A small custom dataset - Techniques like LoRA, RAG etc. • Document: - The dataset used - Fine-tuning procedure - Training duration - Parameters or hyperparameters	**15**
Generate and Evaluate Fine-Tuned Outputs • Evaluate using the same metrics. • Include generated results and scores. • [Optional]: Use the same original and engineered prompts on the fine-tuned model.	**10**
Compare & Contrast Results • Present a side-by-side comparison of the results: - Pretrained vs. Prompt-Engineered vs. Fine-Tuned • Provide both: - Quantitative analysis (scores) - Qualitative analysis (visual or subjective impressions) • Reflect on what worked, what didn't, and why.	**15**
Ethical Considerations: • Use this section to discuss the ethical issues that might be involved in the chosen area of research or/and the dataset used. For example: - Are there risks of bias, or misinformation in the model's output? - Does the dataset contain sensitive, private, or potentially harmful content? - Could the generated content be misused (e.g., deepfakes, fake news, impersonation etc.)? - Is there a risk of reinforcing harmful stereotypes, language, or social biases?	**5**
Code Submission • Submit a clean, well-commented Jupyter Notebook. • Code should be modular and reproducible. • Outputs should be matching with the report submitted.	**10**

Task	Criteria	Max Points	First	2:1	2:2	Third	Poor
1	Goal	5	Goal is precise & measurable. Modality fit is justified; clear success criteria & scope, along with assumptions/risks, are stated. **5 pts**	Clear goal & modality fit; success criteria partly quantified; minor scope or risk gaps. **4 pts**	General goal (e.g., "improve quality"); modality fit implied; outcomes loosely defined. **3 pts**	Vague/over-broad; success criteria missing; weak link to modality/problem. **2 pts**	No coherent goal or off topic. **0–1 pt**
2	Inspiration from Literature	10	≥3 strong sources; each has a 1–2 paragraph critique covering method, data, compute, strengths/limits, an explicit "design takeaway" mapped to your pipeline. **10 pts**	3 sources summarised with some critique & takeaways; a few method/data gaps. **8 pts**	2–3 sources; mostly descriptive; weak or generic takeaways. **6 pts**	Superficial summaries; unclear relevance; missing key details. **4 pts**	Missing/irrelevant or uncited. **0–2 pts**
3	Generate Baseline	5	Reproducible run of a standard model with 3–5 well-rationalised prompts; show seeds, version, inference params; present 6 – 8 representative outputs. **5 pts**	Baseline runs; prompts listed; minor omissions (e.g., missing seed or version). **4 pts**	Minimal run; prompts not justified; few/low-quality exemplars. **3 pts**	Unclear config; patchy results; not obviously baseline. **2 pts**	Not run/failed without evidence. **0 – 1 pt**
4	Evaluate Baseline Results	5	Correct metrics & rationale, with formulas or tool refs; table of scores + 3–5 sentence interpretation per metric (what improves what, limits). **5 pts**	Appropriate metrics computed; brief interpretation; minor correctness/justification gaps. **4 pts**	Some metrics or only human judgement; surface-level discussion. **3 pts**	Misapplied, incomparable, or not reproducible metrics. **2 pts**	No evaluation. **0–1 pt**

5	Prompt Engineering	15	Systematic interventions (e.g., role/context framing, constraints, few-shot, ordering, self-critique, temperature/top-p changes). Clear hypotheses; before/after grids; prompt templates versioned; ablation (which edit yielded which change). **15 pts**	Several sensible edits; persuasive before/after; some hypothesis testing. **12 pts**	Ad-hoc edits; qualitative comments; limited evidence. **9 pts**	Minimal edits; weak causal link to changes. **6 pts**	Not attempted or non-evidenced. **0–3 pts**
6	Evaluate Engineered Prompts	5	Same metrics reused; neat table: Baseline vs Engineered; effect sizes or % change; short analysis of trade-offs (e.g., diversity↑, faithfulness↓). **5 pts**	Table/plot comparing scores; surface-level discussion. **4 pts**	Basic comparison (numbers only), thin insight. **3 pts**	Partial/inconsistent comparison. **1-2 pts**	No comparison. **0 Pts**
7	Fine-Tune the Generator	15	Clear FT strategy. Dataset stated with; train/val split; batch size, steps/epochs, lr/scheduler, reg, compute/time; seeding; why these choices; failure modes & mitigations. **15 pts**	Sensible FT; adequate hyper-params and data notes; limited justification. **12 pts**	Minimal FT (few steps or unclear method); missing key params. **9 pts**	Execution issues; unjustified/opaque setup. **6 pts**	Not attempted. **0–1 pts**
8	Generate and Evaluate Fine-Tuned Outputs	10	Re-run same metrics; show qualitative samples; if applicable, reuse baseline/engineered	Metrics + samples; moderate interpretation. **8 pts**	Results present, limited or uneven analysis. **6 pts**	Weak link from FT to outcomes; little evidence. **4 pts**	Missing results/eval. **0–2 pts**

			prompts; discuss generalisation vs overfit, and cost/latency impacts. **10 pts**				
9	Compare & Contrast Results	15	One consolidated side-by-side table/figure for Pretrained vs Prompt-Engineered vs Fine-Tuned; highlight wins/losses, variance/Confidence Interval if relevant; integrate qualitative and quantitative; reflective "what worked/why/limits/next" . **15 pts**	Good, consolidated comparison; mostly quantitative with some reflection. **12 pts**	Basic tables/plots; limited analysis and discussion. **9 pts**	Fragmented or narrative-only comparison. **6 pts**	No coherent synthesis. **0–3 pts**
10	Ethical Considerations	5	Concrete risks (bias, stereotyping, misinformation, IP/privacy, misuse e.g., deepfakes) tied to your data/model; feasible mitigations (filters, safety prompts, consent, watermarking, eval); license/attribution; limitations. **5 pts**	Good coverage with minor omissions or generic mitigations. **4 pts**	General statements; few specifics or missing mitigations. **3 pts**	Some mentions; no concrete actions or mitigations. **1-2 pts**	Absent. **0 pt**
11	Code Submission	10	Single clean Jupyter: modular functions, docstrings, config cell; fixed seeds; clear Run-	Mostly reproducible; minor manual steps. **8 pts**	Runs with tweaks, partial mismatch with report. **6 pts**	Hard to run; unclear structure; missing deps. **4 pts**	Not provided/non-functional. **0–2 pts**

			All; notebook re-generates report figures/tables; saved artifacts. **10 pts**				