**COMP 691H Learning Contract**

Name: Akshan Sameullah

Project Description:

* Build and release a rigorous Fermi estimation benchmark with sourced ground‑truth ranges and units.
* Test whether “wisdom‑of‑crowds” effects (inspired by Galton) emerge from a diverse array of LLM ensembles and quantify convergence with sample size (Law of Large Numbers analogues).
* Develop and evaluate methods to improve LLM Fermi estimation via
  + Decomposition into measurable factors with unit checks (Breaking down question into subproblems)
  + Retrieval of base rates (From the web or another data source, RAG pipeline vs. pure reasoning comparison)
  + Distributional outputs with calibrated uncertainty intervals. (Returning ranges with uncertainties and confidences instead of final answer)
* Provide a reproducible evaluation codebase featuring ensemble diversity, error distributions, calibration, and the effect of structured reasoning

Meeting Requirements:

* Weekly Meetings on Friday at 11:30am
* L3 Lab Meeting on Tuesdays at 2:15pm

Reading Assignments:

1. Papers suggested in lab meetings based on findings and potential advancement directions.
2. William Poundstone, “Fortune’s Formula” (Galton & wisdom-of-crowds chapters)
3. “Self-Consistency Improves Chain-of-Thought Reasoning in LLMs,” Wang et al., ICLR 2023
4. “Emergent Abilities of LLMs,” Wei et al., 2022
5. “Calibrated Fermi Estimation,” Lorig & Riepe, 2020
6. “Diverse Prompts for Language Models,” Kojima et al., 2023
7. Selected chapters from Bishop, “Pattern Recognition and Machine Learning” (ensemble methods section)

Written Assignments:

* Relevant Fermi performance statistics measuring result of variations in ensemble parameters (temperature, role-play, data sources, etc)
* Relevant Paper Presentations

Software or hardware deliverables:

* Cleaned & Labeled Fermi Question Dataset for testing Law of Large Numbers
* Reproducible experiment scripts and plots that aim to answer the following research questions:
  + Do LLM ensembles exhibit a predictable reduction in Fermi estimation error with more diverse samples? What is the convergence rate of log‑error as a function of sample size and diversity (temperature, prompts, model families)?
  + Can structured factorization with unit‑safe computation and retrieved base rates reduce systematic log‑error and overconfidence?
  + Can we elicit and calibrate distributional forecasts (e.g., 50% and 90% intervals) with correct coverage on Fermi tasks?
  + What biases and failure modes characterize LLMs on Fermi problems (e.g., unit mistakes, reference class errors, over‑aggregation of wrong assumptions)?

Other assignments (e.g., presentations):

* Paper Reading Presentation to L3 Lab
* Presentation of project to L3 Lab (and other relevant spaces as opportunities are discovered)

Assessment Plan (i.e, how grade determined):

* Weekly Milestones (Code delivery, tests, logs, etc) (33%)
* Software codebase (Reproducible, documented) (33%)
* Final Presentation of project to L3 Lab (and other relevant spaces as opportunities are discovered) (33%)

Criteria for continuing to COMP 692H:

* Creation and Compilation of cleaned Fermi Dataset
* Baseline experiment results: single-shot vs. 10-sample ensemble, RAG vs. pure reasoning, or other viable options
* Documented Codebase stored on Cloud Repository