###### 

# **Chapter 1 – Introduction**

## **Background of the Study**

Large language models (LLMs) such as GPT-4, Claude 3 Opus, and Llama 3 have achieved near–human–level performance on many single-domain benchmarks (Bubeck et al., 2023; Touvron et al., 2023). Yet, practitioners, educators, and researchers increasingly rely on these systems to answer questions drawn from a wide breadth of disciplines—from molecular biology to macroeconomics. Early cross-domain evaluations (e.g., MMLU; Hendrycks et al., 2021) reveal steep performance gradients across subjects, indicating that a model’s advertised “general intelligence” can mask critical blind spots that only emerge in less-publicized fields.

Recent work has begun to investigate discipline-specific capabilities: CS-Bench targets 26 sub-fields of computer science (Song et al., 2024); BrainBench measures neuroscience forecasting accuracy (Luo et al., 2024); and MaScQA probes metallurgical engineering knowledge (Bajan & Lambard, 2025).Bibliography. However, no empirical study has yet provided a systematic, statistically rigorous comparison of leading LLMs across the full span of *all* 26 academic fields defined by Scopus. The absence of such evidence leaves universities, enterprises, and government agencies without clear guidance when selecting or governing models for high-stakes, domain-specific tasks.

## **Statement of the Problem**

**The problem to be addressed in this study is that current large-language models exhibit uneven and insufficiently quantified knowledge across the 26 major academic fields, and stakeholders lack reliable, field-level rankings to inform critical adoption and governance decisions**  Preliminary leaderboard data show that GPT-4 can exceed 90% accuracy on business-school exams yet fall below passing thresholds on pharmacist-licensing questions (Wang et al., 2023; Terwiesch, 2023). Without a statistically validated, cross-field evaluation, organizations risk deploying models that underperform in specialized contexts, potentially compromising research integrity, educational outcomes, and public trust.

## **Purpose of the Study**

**The purpose of this quantitative, comparative evaluation study is to rank the domain knowledge of state-of-the-art LLMs across 26 academic fields by administering a balanced corpus of 1,000 multiple-choice items per field and statistically analyzing differences in mean accuracy.** Results will produce an Elo-style leaderboard (cf. LMSYS Chatbot Arena) that highlights strengths and weaknesses at the field level while controlling for item difficulty and chance performance.

## **Nature of the Study**

A **between-models quasi-experimental design** will be employed. Each model (e.g., GPT-4, Claude 3 Opus, Gemini 1.5 Pro, Llama 3 70 B, Phi-3 mini) will answer the identical 26 × 1,000-item test set under deterministic temperature-0 settings. Correct answers earn +0.1, incorrect answers –0.1, and “unknown/abstain” 0, mirroring recent best practices for calibrated scoring. One-way repeated-measures ANOVA followed by Tukey HSD will test field-level performance differences, while Levene’s test will assess homogeneity of variance.

## **Research Questions and Hypotheses**

| **RQ** | **Research Question** | **Null Hypothesis *(H₀)*** | **Alternative *(H₁)*** |
| --- | --- | --- | --- |
| 1 | Does any LLM exhibit statistically uniform performance across the 26 fields? | Mean scores differ by field for every model. | At least one model’s means do not differ by field. |
| 2 | Do current LLMs differ in mean accuracy when reasoning about specialized academic content? | No difference exists among model means. | ≥ 1 model differs from another. |
| 3 | Do alternative validation methods (e.g., bootstrap vs. split-half) yield different reliability estimates for field rankings? | Reliability estimates do not differ by validation method. | At least one method yields a different reliability estimate. |

## **Conceptual Framework**

This study is grounded in the **Holistic Evaluation of Language Models (HELM) framework** (Liang et al., 2022), which posits that trustworthy assessment must encompass accuracy, calibration, robustness, and fairness. Bibliography Our focus on accuracy and robustness across subject domains aligns with HELM’s “scenario × metric” matrix. Additionally, **Cognitive Load Theory** suggests that domain-specific knowledge retrieval places varying demands on a model’s latent memory (Bommasani et al., 2023), providing a theoretical lens for interpreting inter-field variance.

## **Significance of the Study**

By producing the first *field-granular* leaderboard, the study will:

1. **Advance scholarship** by extending single-domain benchmarks (e.g., MMLU) into a comprehensive cross-disciplinary scale.
2. **Guide practitioners**—universities, publishers, and R&D labs—on model selection for subject-matter-expert tasks.
3. **Inform policy** by supplying regulators with empirical evidence of domain gaps that could amplify misinformation or inequity (Bommasani et al., 2023).

## **Definitions of Key Terms**

**Elo Ranking.** A rating system that updates a competitor’s score based on pairwise outcomes, adapted here to aggregate win-loss probabilities between LLMs (Originality.ai, 2024).

**Large-Language Model (LLM).** A neural network with ≥ 1 billion parameters trained on massive text corpora to predict the following tokens and generate coherent language (Brown et al., 2020).Bibliography

**MMLU Benchmark.** A 57-subject multiple-choice test used to measure multitask language understanding (Hendrycks et al., 2021).

## **Assumptions**

1. Publicly released model checkpoints evaluated offline behave comparably to proprietary API versions.
2. Multiple-choice items sampled from prior exams faithfully represent field knowledge distributions.

## **Scope, Limitations, and Delimitations**

*The scope* includes English-language questions at an undergraduate level. *Limitations* involve potential prompt-format bias and exclusion of multimodal reasoning. *Delimitations* restrict the study to seven widely used transformer-based language models (LLMs) available as of the data collection cut-off (April 2025).

## **Chapter Summary**

Chapter 1 introduced the pervasive yet uneven adoption of LLMs and articulated the need for a systematic ranking across 26 academic fields. The problem, purpose, research questions, and HELM-based framework establish a foundation for the mixed-methodology approach detailed in Chapter 3. Chapter 2 will now survey prior evaluations of LLM knowledge and identify documented performance gaps at the field level.

**References**

Bajan, C., & Lambard, G. (2025). Exploring the expertise of large language models in materials science and metallurgical engineering. *Digital Discovery, 4*(4), 500–512. <https://doi.org/10.1039/D4DD00319E>

Bommasani, R., Liang, P., & Lee, T. (2023). Holistic evaluation of language models. *Annals of the New York Academy of Sciences, 1525*(1), 140–146 <https://doi.org/10.1111/nyas.15007>

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4* (Version 3) [Preprint]. arXiv. <https://arxiv.org/abs/2303.12712>

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring massive multitask language understanding* (Version 2) [Preprint]. arXiv. <https://arxiv.org/abs/2009.03300>

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Li, Y., Raghunathan, A., … Hashimoto, T. (2022). *Holistic evaluation of language models (HELM)* [Preprint] arXiv.<https://arxiv.org/abs/2211.09110>

Luo, X., Zhang, X., Ren, Z., Li, W., Wang, Y., Zhao, Y., … Kording, K. (2024). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour, 8*, 675–686. <https://doi.org/10.1038/s41562-024-01683-2>

Originality.ai. (2024, March 18). *Top foundational LLM models and organizations*. https://originality.ai/blog/top-foundational-llm-models

Song, X., Diao, M., Dong, G., Wang, Z., Fu, Y., Qiao, R., … Xu, W. (2024). *CS-Bench: A comprehensive benchmark for large language models towards computer-science mastery* [Preprint]. arXiv <https://arxiv.org/abs/2406.08587>

Terwiesch, C. (2023. *Would ChatGPT earn a Wharton MBA? A prediction based on its performance in the operations management course* [Working paper]. Wharton School of the University of Pennsylvania <https://doi.org/10.2139/ssrn.4333495>

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., … Scao, T. L. (2023). *LLaMA: Open and efficient foundation language models* [Preprint. arXiv. <https://arxiv.org/abs/2302.13971>

Wang, Y.-M., Shen, H.-W., & Chen, T.-J. (2023). Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *Journal of the Chinese Medical Association, 86*(7), 653–658. <https://doi.org/10.1097/JCMA.0000000000000839>