# Chapter 3: Methodology

## Introduction

This chapter outlines the methodology employed in the study. The problem addressed is that current large language models (LLMs) demonstrate uneven and insufficiently quantified knowledge across the 26 major academic fields, leaving stakeholders without validated benchmarks for adoption and governance 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 multiple-choice items and statistically analyzing differences in mean accurac. This chapter describes the research design, population and sample, instruments, procedures, data analysis, and ethical considerations, providing a roadmap for replication.

## Research Design

A quasi-experimental, between-models design will be employed to compare the performance of multiple LLMs across 26 fields. Each model (e.g., GPT-4, Claude 3 Opus, Gemini 1.5 Pro, LLaMA 3 70B, Phi-3 Mini) will complete the identical test set of 1,000 multiple-choice items per field. The independent variable is the LLM used, while the dependent variable is the accuracy score. Alternative designs such as case studies or qualitative interviews were deemed less appropriate due to their inability to provide statistically rigorous, field-level comparisons.

## Population and Sample

The population of interest includes the 26 academic fields defined by Scopus. The sample will consist of 1,000 multiple-choice items per field, drawn from prior exams, validated question banks, and peer-reviewed sources. This results in a total sample of 26,000 items. The models selected represent widely used, state-of-the-art transformer-based architectures available as of April 2025. This sampling approach ensures coverage of both breadth and depth of domain knowledge while maintaining replicability.

## Materials and Instrumentation

The instrument used is a balanced corpus of multiple-choice items, standardized to undergraduate-level knowledge. Each item has one correct answer and three incorrect options, known as distractors. Correct responses will be scored as +0.1, incorrect responses as -0.1, and abstentions as 0, adapting best practices for calibrated scoring. Deterministic evaluation (temperature = 0) will be applied to eliminate randomness. Items were selected to represent domain knowledge faithfully, and pilot testing will be conducted to ensure fairness across models.

## Operational Definitions of Variables

Independent Variable: The LLM evaluated (GPT-4, Claude 3, Gemini 1.5, etc.).  
Dependent Variable: Mean accuracy scores per academic field.  
Measurement: Accuracy will be calculated as the percentage of correct responses across 1,000 items per field, adjusted by the scoring schema.

## Study Procedures

. Assemble a balanced test corpus (1,000 questions per field).  
2. Standardize item formatting to ensure consistency.  
3. Configure deterministic evaluation for each model.  
4 Administer the test corpus to each LLM under identical conditions.  
 . Record and score results using the calibrated scoring system. 6. Compile accuracy scores per field and model for analysis.

## Data Analysis

Data will be analyzed using one-way repeated measures ANOVA to test for differences in model performance across fields. Tukey’s HSD will be applied for post-hoc comparisons. Levene’s test will check the homogeneity of variance. Reliability will be assessed using bootstrap and split-half validation methods. Statistical significance will be set at p < .05 Analyses will be conducted using statistical software (e.g., SPSS, R).

## Validity and Credibility Issues

Validity will be supported by constructing the test corpus from validated sources and piloting items before deployment. Credibility is enhanced through deterministic evaluation to remove stochastic noise. Reliability will be examined via multiple validation strategies. Researcher bias is minimized by automating model evaluation and scoring.

## Assumptions

1. Publicly available model checkpoints behave comparably to their proprietary API versions.  
2. Multiple-choice items adequately represent domain knowledge in each field. Deterministic evaluation settings ensure replicable results.

## Limitations

This study is limited by its focus on English-language items at an undergraduate level, excluding multimodal and higher-order reasoning. Prompt-format bias and differences in pretraining corpora may influence outcomes. The study is restricted to transformer-based models released by April 2025.

## Delimitations

Delimitations include limiting the scope to 26 Scopus fields, a multiple-choice question format, and deterministic evaluation conditions. These decisions enhance comparability but restrict generalization to other tasks or languages.

## Ethical Assurances

This study involves no human subjects and therefore qualifies as exempt from IRB oversight. All data will be sourced from publicly available sources and securely stored in accordance with institutional guidelines. The researcher acknowledges potential bias due to prior experience with LLM evaluation but will apply strategies such as blind scoring and external replication to mitigate this risk.

## Summary

This chapter detailed the methodology for systematically evaluating LLM knowledge across 26 academic fields. The design, population, instruments, and procedures were described in detail, along with statistical analysis, validity strategies, assumptions, and limitations. By applying a rigorous and replicable methodology, the study aims to produce the first comprehensive leaderboard of LLM performance across academic disciplines. The next chapter will present the results of this analysis.