

Can Large Language Models Learn Statistics

Shourya Marwaha
201907311

Supervised by Dr Leonid Bogachev and Dr. Serge Sharoff

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Masters of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

September 2025

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.



School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name Shourya Marwaha

Student ID 201907311

Abstract

Large Language Models (LLMs) have demonstrated exceptional abilities across various domains, but their capability to genuinely learn and incorporate knowledge in specialised fields remains a topic of discussion. LLM's ability to solve statistics problems is excellent, but it isn't completely known if it is because of pattern identification from training data or their actual statistical reasoning ability. This dissertation investigates the question, "Can LLMs Learn Statistics?" For this research, an experiment is conducted to test the learning capability of the LLMs. Learning capability of an LLM can be defined as its ability to integrate new information relevant to a task and demonstrate improved performance based on that integration. Several different LLMs will be asked to solve several statistics questions covering key areas like correlation, regression, probability, and hypothesis testing. This process takes place in two different environments:

- Baseline Testing: The LLM is not provided with any external resource
- Retrieval Augmented Generation: The LLM is provided with statistics books to refer to answer the questions.

The LLM as a judge methodology is implemented to maintain objectivity in the evaluation process. These evaluation models are specifically trained to apply an established rubric criterion across all LLM responses, which are then compared to human solutions.

Contents

1	Introduction	1
1.1	Research Problem and Objectives	2
1.2	Methodology Overview	2
1.3	Significance	3
1.4	Dissertation Structure	4
2	Literature Review	5
2.1	The Foundation of LLMs	5
2.2	Scaling Laws and Emergent Capabilities in LLMs	6
2.3	Contemporary LLM Landscape	7
2.4	Mathematical and Statistical Reasoning in AI Systems	7
2.5	RAG: Retrieval Augmented Generation	8
2.6	RAG vs. Fine-tuning	9
2.7	The Learning versus Memorisation	10
2.8	Evaluation Challenges in AI Reasoning Assessment	11
2.9	Current Limitations and Research Gaps	12
2.10	This Research: Addressing the Gaps	13
3	Methodology	14
3.1	Overview	14
3.1.1	Research Design Structure	14
3.1.2	Evaluation Framework	15
3.1.3	Statistical Analysis Framework	15
3.1.4	Hypotheses and Statistical Power	16
3.2	Dataset Development and Curation	16
3.2.1	Question Dataset Design Principles	16
3.2.2	Knowledge Base Construction Rationale	17
3.3	Model Architecture Selection and Configuration	18
3.3.1	Model Selection Rationale	18
3.3.2	Configuration Standardisation and Reproducibility	19
3.4	RAG System Implementation and optimisation	19
3.4.1	RAG Architecture Design Philosophy	19
3.4.2	Systematic Parameter optimisation Methodology	21
3.4.3	Advanced Retrieval Enhancement Techniques	21
3.4.4	Semantic Evaluation Framework	22
3.4.5	Empirical optimisation Results	22
3.5	Evaluation Framework Design	24

3.5.1	Multi-Dimensional Assessment Rationale	24
3.5.2	LLM-as-a-Judge Implementation Strategy	25
3.5.3	Rubric Design and Scoring Implementation	26
3.6	Statistical Analysis Methods	26
3.6.1	Primary Analysis Framework	26
3.6.2	Between-Model Comparative Analysis	27
3.6.3	Dimensional and Domain-Specific Analysis.	28
3.6.4	Statistical Assumptions and Adaptive Test Selection	28
4	Data Description	30
4.1	Foundation Phase: Initial Dataset Construction	30
4.1.1	Dataset Description	30
4.1.2	Initial Dataset Structure	30
4.1.3	Question Selection Methodology and Quality Assurance	31
4.2	Expansion Phase: Model Response Integration	31
4.2.1	Systematic Response Generation Framework	31
4.2.2	Multi-Iteration Protocol Implementation	32
4.3	Assessment Phase: Comprehensive Evaluation Integration	32
4.3.1	LLM-as-a-Judge Framework Implementation	32
4.3.2	Score Aggregation and Dataset Finalisation	33
4.4	Completion Phase: Final Dataset Architecture	33
4.4.1	Comprehensive Dataset Specifications	33
4.4.2	Data Quality and Analytical Capabilities	34
5	Results	36
5.1	Statistical Assumptions and Method Selection	36
5.2	Primary Learning Effects Analysis	36
5.2.1	Between-Model Comparisons and Power Analysis	39
5.2.2	Evaluation Framework Reliability	39
5.2.3	Key Findings Summary	40
6	Discussion	43
6.1	Interpreting the Negative Learning Effects	43
6.2	Implications for AI Learning and Development	44
6.2.1	Practical Applications and Implementation Challenges	44
6.3	Study Limitations and Methodological Considerations	44
6.4	Future Research Directions	44
7	Conclusion	46

List of Figures

3.1	RAG Pipeline Flowchart	20
3.2	Semantic RAG Evaluation Results	23
3.3	Top K Retrieval optimisation Results	23
5.1	Overall Performance: RAG vs Baseline Conditions	37
5.2	Performance Breakdown by Evaluation Dimensions	38
5.3	Performance by Statistical Domain	41
5.4	Component-Level Analysis by Statistical Domain	42

List of Tables

1.1	Selected Model Architectures and Their Roles in the Study	3
3.1	Final RAG Configuration Parameters	24
3.2	Judge Model Performance Comparison	25
4.1	Domain-Based Question Distribution	31
4.2	Evaluation Dimension Framework	33
4.3	Variable Structure Summary	34
5.1	Summary of Learning Effects Across Models and Dimensions	37
5.2	Mean Performance Scores by Condition	38
5.3	Judge Model Performance Comparison	39
5.4	Mean Performance Scores by Condition	40

Chapter 1

Introduction

In today's world, large language models (LLMs) are among the most important discoveries in the field of artificial intelligence. LLMs are based upon the Transformer models [Vaswani et al., 2017] and are trained on huge amounts of textual data. LLMs have demonstrated remarkable abilities in tasks such as understanding and generating text that is close to human output across various domains. They have the ability to perform complex reasoning tasks, solve mathematical problems and give explanations [Brown et al., 2020]. However, the main question that remains to be a topic of discussion is whether these LLMs can genuinely learn and integrate the knowledge that they come across, or whether they just use advanced pattern recognition from their extensive training data [Arpit et al., 2017].

The distinction between genuine learning and pattern recognition is important in specialised fields that require both computational ability and deep conceptual understanding. Statistics is one of the specialised fields, where problem solving doesn't only involve the application of formulas, but also involves various other processes such as identifying appropriate analytical methods, recognising underlying assumptions, interpreting results within proper theoretical frameworks [Garfield and Ben-Zvi, 2008]. While LLMs often produce correct answers to statistical questions, the underlying mechanism for this success remains opaque. It raises the question whether these LLMs reflect genuine learning capabilities or just pattern recognition.

The consequences of this uncertainty between genuine learning and pattern recognition extend beyond academic curiosity. In the educational context, if the answer to the question is known, it could help develop adaptive tutoring systems that grow with students. In the scientific context, it could help create AI assistants that can integrate new methods and literature in real time. Broadly, distinguishing between learning and pattern recognition addresses critical questions about AI reliability, explainability, and the essence of machine intelligence.

This dissertation investigates whether LLMs can genuinely learn statistics when provided with relevant information. Statistics consists of multiple topics such as correlation analysis, regression modelling, probability distributions, and hypothesis testing, which can help us determine if LLMs can genuinely learn. This dissertation examines when LLMs are exposed to a new source of information, how they process it and learn from it, and whether their ability to

solve statistics questions improves or not.

1.1 Research Problem and Objectives

The central challenge this research addresses is that current evaluation methods often fail to differentiate between genuine learning and sophisticated pattern recognition in LLMs [Schoenfeld, 1985]. This ambiguity is particularly problematic in statistical reasoning, where superficial accuracy can mask deep conceptual misunderstandings. The ability to adjust to new problem types is an essential sign of real understanding.

Other approaches for enhancing LLM capabilities, such as fine-tuning, permanently change model parameters in a way that mixes learning effects with memorisation [Howard and Ruder, 2018], making it impossible to separate genuine learning from memorisation. Due to this methodological limitation, researchers have not been able to clearly determine if LLMs can acquire knowledge dynamically, which could mark a significant step forward in artificial intelligence.

Research Objectives:

- **Establish baseline capabilities:** Systematically evaluate the performance of LLMs on a dataset which comprises of statistical problems from various branches of statistics such as correlation analysis, regression modelling, probability distributions and hypothesis testing without access to any external resources.
- **Implement RAG enhanced evaluation:** Deploy a RAG system which provides the LLMs with access to relevant chunks from the various statistics textbooks [Lewis et al., 2020] while the LLM is solving the problems and also measure the difference between the performance.
- **Analyse improvement patterns:** Investigate whether the performance of LLMs gains under RAG implementations are statistically significant and aligns with the real conceptual understanding rather than just surface-level keyword matching or retrieval artifacts.
- **Assess reasoning quality:** Use comprehensive evaluation frameworks that measure not just numerical accuracy but Explainability and Understanding as well to determine whether improvements reflect genuine understanding, going beyond traditional benchmark methods [Hendrycks et al., 2020].

1.2 Methodology Overview

This dissertation implements a controlled experimental design that combines three methodologies for investigating learning capabilities of the LLMs:

- **RAG Implementation:** This dissertation deploys RAG system to provide the LLMs with a temporary and controlled access to an external statistical knowledge without modifying the model parameters permanently [Lewis et al., 2020]. The RAG system retrieves relevant chunks, methodological explanations, and worked examples from a curated collection of authoritative statistics materials. This approach allows for a direct comparison between baseline performance and RAG performance, thereby separating the effects of knowledge integration from parameter memorisation.
- **Selected Model Architectures:** Three open-source LLMS with different levels of specialisation in mathematics were chosen for this dissertation.

Table 1.1: Selected Model Architectures and Their Roles in the Study

Model Name	Key Characteristic	Purpose in this Study
LLaMA 3.2 8B, Touvron et al. [2023]	General-purpose foundation model	To establish a baseline for learning capability without specialised training.
DeepSeek-R1 8B, Shao et al. [2024]	Advanced reasoning model with RL optimisation	To assess if enhanced general reasoning abilities improve knowledge integration.
Qwen2-Math 7B, Yang et al. [2024]	Math-specialised (Synthetic Data)	To provide a contrasting specialisation approach for comparison.

These models will help in conducting a systematic investigation to determine the relationship between the already existing domain knowledge with the dynamic learning capacity of the LLMs.

- **LLM-as-a-Judge Evaluation Framework:** In order to ensure scalable and consistent assessment across all the model responses, an automated evaluation method, which involves using a sophisticated language model as a judge, has been deployed. This approach builds upon previous works in automated assessment [Shermis and Burstein, 2013] and allows for a systematic evaluation of various aspects of statistical reasoning quality, which includes parameters like numerical correctness, methodological soundness, assumption recognition, and explanation clarity. This framework evaluates all the responses and applies a structured rubrics consistently across all of them.

1.3 Significance

This dissertation addresses critical gaps in the understanding of the capabilities of the LLMs and also contributes to methodological innovations for investigating machine learning in AI

systems. Demonstrating genuine learning in LLMs could have a significant impact on various fields, including educational technology and scientific research assistance.

If the LLMs demonstrate genuine successful learning capabilities, it could lead to the development of adaptive AI tutoring systems that could improve along with the students, scientific research assistants that are capable of integrating new methodologies in real-time, and more trustworthy AI systems whose limitations and capabilities are better understood.

This dissertation confirms RAG as a means of analysing knowledge integration and shows scalable automatic grading techniques for complex reasoning exercises. These advancements would serve as frameworks for future studies about AI learning capabilities in various fields.

1.4 Dissertation Structure

This dissertation comprises seven chapters that precisely discuss the research question:

Chapter 2 reviews the existing literature on LLM capabilities, exploring the current research on AI learning methods, and highlighting the challenges of statistical reasoning in AI systems. This chapter lays down the groundwork and identifies the gaps in the current knowledge.

Chapter 3 details the methodology used in creating the experimental design, the implementation of baseline and RAG conditions, the LLM-as-judge evaluation framework, and lastly the statistical analysis methods employed to assess learning capabilities.

Chapter 4 provides complete details about the statistical question dataset, the textbook corpus available under RAG conditions, the LLM architectures tested, and the evaluation rubric criteria. This chapter offers context for analysing data and ensures reproducibility.

Chapter 5 discusses the experiment's results and insights. It provides all the empirical findings, including performance comparisons between baseline and RAG conditions, analysis by statistical domain and LLM architecture, statistical significance testing, and interpretation of results in relation to the central research question.

Chapter 6 discusses and interprets the results in the context of the broader literature, explores their theoretical and practical implications, acknowledges the study's limitations, and proposes directions for future research.

Chapter 7 summarises the key findings, reiterates the study's contributions, and offers final remarks on the future of learning in AI.

The research progresses with a logical from theoretical foundation, through methodological advancement to empirical analysis, and conclusion. Every chapter builds upon previous work and contributes to the overall investigation. This systematic approach is implemented to make sure that there is a comprehensive coverage of the research question, and the focus on the main topic of whether LLMs can indeed learn statistics is maintained. The empirical results will provide conclusive proof to support theoretical comprehension and real-world LLM applicability in statistical contexts.

Chapter 2

Literature Review

2.1 The Foundation of LLMs

The development of LLMs marks a breakthrough in computer intelligence. They have fundamentally changed the modern era of AI [Brown et al., 2020]. The era of LLMs began in 2017, with the introduction of the Transformer architecture, which completely transformed natural language processing (NLP) through the self-attention mechanism [Vaswani et al., 2017]. Traditional NLP models did not possess the ability to process sequences in parallel, capturing the effect of long-range dependencies effectively.

The Transformer’s groundbreaking contribution lies in its attention mechanism, which allows models to focus on essential parts of input sequences, no matter how far they are. This innovation helped solve the issues of recurrent neural networks and convolutional methods, which made training of significantly larger models on vast datasets possible while maintaining computational efficiency. The attention mechanism can link any two points in a sequence, which is helpful for performing tasks for complex reasoning that require integrating information from broader contexts [Vaswani et al., 2017].

The progression from architectural innovation to practical application started when early language models demonstrated the potential of large-scale unsupervised pre-training. Early research on contextualised representations showed that large-scale unsupervised pre-training on various text corpora might help models learn rich linguistic representations that could be used in other tasks through fine-tuning [Peters et al., 2018]. At the same time, bidirectional context processing was introduced by BERT, which achieved significant improvements in several natural language understanding benchmarks [Devlin et al., 2019]. These early achievements laid the groundwork for fundamental training paradigms that remain dominant in the development of current LLMs.

The upgrade from GPT to GPT-2 marked the first clear demonstration of the emerging capabilities in language models. Research on scaling neural language models demonstrated that making the model bigger and giving it more training data improves it at some tasks and leads to new behaviours. Bigger models also showed better few-shot learning abilities, which suggested

that scale could lead to qualitatively distinct model behaviours [Rosenfeld et al., 2019]. With the release of GPT-3, the relationship between scale and emergent capabilities was established. It achieved impressive few-shot learning across various tasks without needing specific training for each task with the help of its 175 billion parameters [Brown et al., 2020]. This set the stage for investigating whether these emergent abilities represented true reasoning, particularly in complex, symbolic domains like statistics.

2.2 Scaling Laws and Emergent Capabilities in LLMs

The release of GPT-3 marked a significant step forward in language model capabilities. GPT-3 demonstrated how scale could induce emergent behaviours that were not explicitly programmed during training. These emergent capabilities consisted of abilities like in-context learning, in which models could adapt to new tasks based only on examples in their input and sophisticated reasoning abilities, which enabled models to perform multi-step problem solving across various domains. The ability of the model to perform tasks like mathematical reasoning, code generation, and complex analytics suggested that large-scale pre-training on diverse text sources helped models to acquire broad knowledge about language, mathematics, and reasoning principles [Brown et al., 2020].

The phenomenon of emergent capabilities has been thoroughly studied through scaling laws, which define the predictable relationships between model size, training data, and computational resources [Kaplan et al., 2020]. These scaling laws indicate that most capabilities emerge suddenly at specific scale thresholds rather than developing gradually, leading to the concept of "phase transitions" in model behaviour. For instance, [Wei et al., 2022] demonstrated that multiple abilities or skills like multi-step reasoning, reading comprehension, and math problem-solving emerged sharply as the model exceeded specific parameter counts, providing real-world data proving that neural network performance scales in predictable ways.

Current LLMs exhibit several characteristics that differentiate them from earlier AI systems and are particularly important for understanding machine learning potential. In-context learning allowed LLMs to adapt dynamically without updating the parameters, which further enabled models to incorporate new information temporarily while they were executing their tasks. Few-shot learning capabilities demonstrated generalisation to new tasks with a very small number of examples, while emergent reasoning abilities enabled multi-step problem solving and self-correction behaviours [Brown et al., 2020]. Both of these capabilities suggested that there is a presence of genuine problem-solving skills rather than mere pattern memorisation. However, distinguishing between the two mechanisms remains to be a central challenge in contemporary AI research.

2.3 Contemporary LLM Landscape

The industry of LLMs has grown rapidly, with frequent releases of new and improved models. Both proprietary and open-source models have demonstrated strong capabilities in various areas, including mathematics and statistics. Proprietary models like OpenAI’s GPT-4, Google’s Gemini series, and Anthropic’s Claude have set high standards for mathematical reasoning tasks, demonstrating strong performance in graduate-level mathematics and complex reasoning problems [Fang et al., 2025].

Meanwhile, the open-source alternatives have also performed impressively, featuring remarkable alternatives like Meta’s LLaMA models. The Meta LLaMA series provides transparent architectures and training methodologies that have enabled controlled experimentation. These LLaMA models have also demonstrated that carefully curated training data can achieve competitive performance with fewer parameters than proprietary alternatives. This suggests that data quality may be as important as scale [Touvron et al., 2023].

Though proprietary models currently demonstrate superior performance on a majority of benchmarks, they have major restrictions for academic research [Touvron et al., 2023]. For accessing these models, an API access is required, which is paid for. These models usually operate upon a per-token pricing system, which makes extensive experimentation expensive and possibly unfeasible for detailed research studies. More importantly, these proprietary models operate as black boxes, providing limited internal workings, training data, and architectural specifics, making it difficult to conduct detailed analyses of learning behaviours and patterns of improvement. In addition, API-based access is accompanied by dependencies on third-party services, rate limiting, and potential changes to model versions that can affect the reproducibility of research findings.

2.4 Mathematical and Statistical Reasoning in AI Systems

Investigating mathematical reasoning capabilities in AI systems has become an essential field for determining the difference between computational skill and genuine understanding. Mathematical reasoning goes beyond simple pattern recognition and requires systematic application of abstract principles, multi-step logical reasoning, and the ability to generalise learned procedures to new problem situations [Schoenfeld, 1985]. All these characteristics make mathematical reasoning an ideal testing ground, a role that statistical reasoning, with its added layers of conceptual interpretation, fulfils even more completely.

Recent assessments have revealed that the current AI systems possess remarkable capabilities, but at the same time, a significant number of limitations in mathematical reasoning. The MATH dataset, which consists of 12,500 challenging competition mathematics problems, has become a standard benchmark for evaluating AI mathematical reasoning [Hendrycks et al., 2021]. The performance of AI systems was assessed on this dataset, and it was revealed that these models can achieve an impressive accuracy on routine problems, but they struggle with

new problem variations that require a genuine understanding of basic mathematical ideas. This research highlighted an important difference between computational accuracy and conceptual understanding, which was that getting the right answers doesn't always point towards correct reasoning.

Statistical reasoning adds to the complexities which are relevant to questions about learning and understanding in AI systems. In contrast to pure mathematics, statistical problem-solving requires computational skills as well as conceptual understanding of the subject, such as probabilistic concepts, recognition of correct analytical methods, and interpretation of results [Garfield and Ben-Zvi, 2008]. Statistical reasoning consists of multiple dimensions such as reasoning about variability, understanding of statistical models, recognition of assumptions, and the ability to make appropriate inferences from data [Chance, 2002].

The multi-dimensional nature of statistical reasoning provides a unique opportunity for examining whether AI systems can integrate conceptual knowledge beyond memorised procedures. Statistical thinking involves repeated cycles of problem formulation, method selection, analysis, and interpretation that require flexible adaptation of statistical knowledge to new contexts [Wild and Pfannkuch, 1999]. This complexity provides us with a clear evaluation criterion for evaluating whether improvements in AI performance reflect genuine learning or sophisticated pattern matching, which makes statistical reasoning an optimal area for investigating an AI system's learning capabilities.

Current evaluation methods in mathematical AI reasoning mainly focus on accuracy-based metrics, which might lead us to potentially miss out on important aspects of reasoning quality which is important for distinguishing between genuine understanding from pattern recognition [Ling et al., 2017]. This issue is especially significant in statistical reasoning, because in statistical reasoning, appropriate methodological choices and conceptual understanding are as important as computational accuracy for effective problem-solving. As AI capabilities improve, the need for better evaluation approaches that examine reasoning processes, rather than just final answers, has become increasingly clear.

2.5 RAG: Retrieval Augmented Generation

"The primary advantage of RAG for learning research is its ability to temporarily access external knowledge without permanently altering model parameters" [Lewis et al., 2020]. With this approach, the models can find and use relevant information from the external databases, documents, or knowledge bases when generating responses. This effectively expands their knowledge beyond what they learned during pre-training [Guu et al., 2020]. For this research, RAG has been implemented to explore the real learning abilities by allowing the models to use extensive statistical resources while performing tasks.

While performing RAG, a number of processes are involved. Initially, it involves retrieving relevant chunks of information from the external sources [Karpukhin et al., 2020], then the

original query is augmented with this retrieved knowledge, and responses are generated based on the combined context [Lewis et al., 2020]. This process is particularly suitable for testing whether the models can genuinely integrate new information to improve their reasoning capabilities, or whether their performance is limited to recognising patterns from their training. For statistical reasoning research, it means providing models with access to detailed explanations of statistical methods, their assumptions, and appropriate applications. Then, it can be measured to determine whether this access to external information leads to a better understanding of the methods, instead of just improved accuracy in calculations.

For learning research, RAG provides us with unique methodological advantages which help us address significant limitations present in traditional methods to improve capabilities. The temporary nature of knowledge access in RAG systems allows us to create controlled experimental designs where models can be evaluated with and without access to external resources. This approach provides direct measures of how knowledge integration affects outcomes [Izacard and Grave, 2021]. This controlled access system is particularly useful for performing statistical research. It allows researchers to provide models with detailed explanations of statistical methods, assumptions, and applications while maintaining clear separation between built-in knowledge and retrieved information at the same time.

Recent research has demonstrated RAG’s effectiveness across various tasks which are knowledge-intensive, especially in areas that require a lot of reasoning. Studies indicate that RAG systems can achieve higher performance improvements on question-answering tasks while maintaining interpretability through explicit knowledge source attribution [Petroni et al., 2021]. However, using RAG to explore learning abilities in reasoning tasks is still mostly unknown, making it an important opportunity to explore how AI integrates knowledge.

The effectiveness of RAG systems depends critically on the quality and relevance of retrieved information and the model’s ability to integrate retrieved knowledge with existing reasoning capabilities. Research into RAG optimisation has identified key factors, including retrieval accuracy and context length management [Shi et al., 2023]. These factors become particularly important when evaluating learning capabilities, as performance improvements must be attributed to genuine knowledge integration rather than retrieval artefacts or superficial keyword matching.

2.6 RAG vs. Fine-tuning

RAG and fine-tuning are two methods that are used for improving the performance of LLMs, the fundamental difference between them is whether you want the LLM to learn or you want it to acquire more knowledge. Fine-tuning involves updating the model parameters through additional training on domain-specific datasets, which effectively encodes new knowledge directly into the model’s weights [Howard and Ruder, 2018]. Fine-tuning can achieve high performance improvements, but at the same time, it also mixes genuine learning with pattern memorisation,

as improved performance may simply reflect pattern recognition rather than genuine conceptual understanding [Arpit et al., 2017].

Conversely, RAG offers a much better framework for exploring genuine learning as it keeps a clear separation between the model’s built-in reasoning abilities and its access to external knowledge [Lewis et al., 2020]. This is highly relevant for statistical reasoning research, as it helps in determining whether models can learn to apply statistical concepts appropriately, discover methodological needs, and refine their reasoning based on new evidence about statistical procedures and principles.

RAG method has advantages which extend beyond interpretability to practical research considerations. Fine-tuning demands significant computational resources, technical expertise, and careful dataset curation, which creates barriers for many researchers [Howard and Ruder, 2018]. RAG systems, on the other hand, can be implemented using existing models and knowledge bases, enabling broader research access and making controlled experiments easier. Additionally, RAG enables dynamic knowledge base updates without retraining the whole model, which makes it flexible for iterative research designs [Petroni et al., 2021].

However, RAG and fine-tuning represent two very different approaches towards knowledge integration that may be appropriate for various research questions. Fine-tuning may be more suitable for determining how models internalise and generalise from specific training patterns, while RAG is optimal for examining dynamic knowledge integration and application capabilities. For researching genuine learning mechanisms, RAG’s controlled access to knowledge offers clear advantages in distinguishing learning from memorisation effects.

The choice between the two approaches, RAG and fine-tuning, also impacts their ecological validity and practical applications. RAG is closer to how humans learn because it includes steps like consulting and using external resources to solve problems. In contrast, fine-tuning can be seen as additional training or education where information gets internalised. Understanding both methods is essential for creating well-rounded theories about AI learning abilities and their applications in education and professional settings.

2.7 The Learning versus Memorisation

Differentiating between genuine learning and memorisation remains to be one of the most fundamental challenges in understanding AI capabilities. This challenge becomes more difficult to overcome when evaluating large-scale neural networks, which are trained on huge datasets. In cases of the models, the boundary between generalisable knowledge acquisition and pattern memorisation becomes increasingly hard to establish [Arpit et al., 2017].

When research was conducted into memorisation in neural networks, it revealed complex patterns of information retention that tend to cloud straightforward distinctions of genuine learning versus memorisation. [Zhang et al., 2016] demonstrated that deep neural networks have enough capacity to memorise random labels, which raises questions about whether the observed

performance is due to real pattern recognition or just overfitting. When further research was performed, it was discovered that memorisation and generalisation are not mutually exclusive, models can exhibit both behaviours at the same time, and it depends on the data characteristics and the training dynamics [Feldman and Zhang, 2020].

It becomes a difficult task to distinguish between learning and memorisation in the context of language models when the model is trained on diverse text corpora. LLMs have the ability to extract and reproduce substantial portions of their training data, which raises concerns about whether their reasoning reflects real understanding or just the retrieval of similar examples from the training data [Carlini et al., 2021]. However, this dissertation also reveals that memorisation patterns differ significantly. Models retain varying amounts of information depending on the type and context.

Traditional techniques, which were used to enhance neural network capabilities through fine-tuning, complicated the learning versus memorisation distinction, as it permanently modifies the model parameters with task-specific information [Howard and Ruder, 2018]. While fine-tuning can produce significant performance improvements, determining whether these improvements reflect genuine knowledge acquisition or sophisticated pattern recognition becomes extremely difficult when models have been permanently altered. This limitation has motivated the development of alternative approaches that can isolate learning effects from memorisation artefacts.

2.8 Evaluation Challenges in AI Reasoning Assessment

There are many challenges when it comes to assessing LLMs’ reasoning skills, these challenges extend beyond conventional accuracy-based measures. Traditional evaluation approaches, most of which relied solely on the criteria whether the final answer is correct or not, missed critical aspects of reasoning quality, which is essential for distinguishing between genuine learning and sophisticated pattern recognition [Ling et al., 2017]. This limitation is problematic for complex reasoning tasks where the reasoning process is as important as the final result. This calls for more thorough evaluation frameworks.

The LLM-as-a-Judge methodology has been introduced as a potential solution to the scalability and consistency problems pertaining to the evaluation of AI. This framework uses advanced language models to evaluate the quality, accuracy, and reasoning abilities demonstrated by other models in their responses, building upon established foundations in automated assessment [Shermis and Burstein, 2013]. Research has demonstrated that if appropriately configured, these LLM judges can maintain high correlation with human expert judgments while providing consistent, scalable evaluation across large response volumes [Zheng et al., 2023].

The implementation of LLM-as-a-Judge evaluation needs careful attention to several important factors. Judge models must have clear rubric criteria, reference solutions, and structured evaluation frameworks to ensure consistent assessment [Zheng et al., 2023]. The quality of eval-

uation depends heavily on rubric design, judge model capabilities, and the alignment between evaluation criteria and research objectives. Also, establishing inter-judge reliability through multiple judge models helps ensure robustness and reduces possible biases in automated evaluation.

Recent research has presented both advantages and disadvantages of implementing LLM-as-a-Judge approaches. While these methods can evaluate multiple aspects of reasoning effectively, they may find it difficult to evaluate common domain knowledge and subtle errors in reasoning that need expert understanding [Kocmi and Federmann, 2023]. For statistical reasoning evaluation, this necessitates careful validation of the judge model’s performance against expert human evaluation, particularly for complex statistical concepts and methodological subtleties. Despite these limitations, LLM-as-a-Judge represents a significant improvement in scalable and thorough reasoning evaluation.

2.9 Current Limitations and Research Gaps

Although we have made significant developments in AI capabilities, there are still several key gaps which hinder our understanding of how intelligent systems learn. Keeping our main focus on static capability evaluation rather than the dynamic learning assessment has left fundamental questions about adaptability and knowledge integration largely unanswered. Current research shows what AI systems can do, but it offers little insight into how they acquire, integrate, and use new knowledge. These questions are very important for grasping true learning abilities of these AI systems.

Methodological limitations in existing approaches restrict the investigation of knowledge integration mechanisms. Most studies on knowledge improvement in neural networks use fine-tuning or similar methods that modify the model’s architecture permanently, which makes it impossible to differentiate between improvements which are a result of genuine learning capabilities and those which reflect sophisticated memorisation [Howard and Ruder, 2018][Lewis et al., 2020]. This confusing effect severely limits researchers’ ability to make reliable claims about adaptive knowledge acquisition in AI systems.

Current evaluation practices also impose constraints on understanding AI reasoning capabilities. The predominant focus on accuracy-based metrics may overlook important aspects of reasoning quality [Ling et al., 2017] that differentiate between genuine learning and pattern recognition. This limitation is particularly significant in specialised domains requiring both computational proficiency and conceptual understanding, where appropriate methodological choices and reasoning processes are as important as final answers for meaningful problem-solving.

The use of retrieval-augmented generation in learning research presents a major unexplored opportunity [Izcard and Grave, 2021]. While RAG has been extensively studied for knowledge-intensive tasks, its potential as a methodological tool for investigating learning capabilities remains largely unexplored. The controlled nature of information access in RAG systems provides

unique opportunities to study knowledge integration without permanent architectural modifications. Yet, this potential has not been systematically investigated in learning research contexts.

Statistical reasoning has not received much attention in AI learning research, even though it is well-suited for exploring how knowledge is integrated. The multi-dimensional nature of statistical reasoning requires both computational skills and a solid understanding of concepts, which provides natural evaluation criteria which extend beyond numerical accuracy to include the appropriateness of methods and understanding of concepts [Garfield and Ben-Zvi, 2008]. This domain offers optimal conditions for detecting genuine learning capabilities, yet only a few studies have thoroughly examined how AI systems acquire and apply statistical knowledge.

2.10 This Research: Addressing the Gaps

This dissertation addresses the identified limitations through a new and different methodological approach, which combines RAG with thorough evaluation frameworks to investigate genuine learning capabilities in statistical reasoning. The research design provides a systematic framework for distinguishing between knowledge integration and pattern recognition by employing RAG methodology to provide controlled, temporary access to external statistical knowledge while maintaining a clear separation between built-in knowledge and retrieved information.

The dissertation addresses the methodological limitations in past learning research by employing RAG to compare baseline performance and performance with access to external knowledge resources. This approach eliminates the the confusing effects of changing parameters that make fine-tuning studies hard to interpret. As a result, researchers can more confidently link performance improvements to knowledge integration instead of just memorisation. The temporary nature of knowledge access in RAG systems provides optimal conditions for investigating whether AI systems can demonstrate genuine learning through information integration.

The focus on statistical reasoning as a test domain addresses current gaps in AI learning research by providing optimal conditions for detecting genuine knowledge integration capabilities. Statistical problem-solving requires integration of computational skills with conceptual understanding, making it possible to assess whether performance improvements reflect authentic knowledge integration rather than superficial pattern matching. The multi-dimensional nature of statistical reasoning enables comprehensive evaluation of reasoning quality that extends beyond simple accuracy metrics to encompass methodological appropriateness, assumption recognition, and result interpretation.

The research employs sophisticated evaluation methodologies that build upon established foundations in automated assessment while addressing limitations in current AI evaluation practices. The evaluation framework assesses multiple dimensions of reasoning quality through structured rubrics that emphasise reasoning processes rather than merely final answers. This approach enables systematic investigation of whether improvements under RAG conditions reflect genuine learning improvements across multiple aspects of statistical reasoning.

Chapter 3

Methodology

3.1 Overview

This dissertation investigates whether LLMs can demonstrate genuine learning capabilities when they are provided with external statistical knowledge through the RAG process. The experimental framework addresses the fundamental question of distinguishing between true knowledge integration and advanced pattern recognition in LLMs.

3.1.1 Research Design Structure

A 2×3 within-subjects factorial design was used in the study to examine how two knowledge access conditions interact with three model architectures to shape performance outcomes.

- **Knowledge Access Conditions:**

1. **Baseline:** In this case, models answer statistical questions using only their pre-trained parametric knowledge
2. **RAG-Enhanced:** In this case, models are provided access to curated statistical textbooks through an optimised retrieval system, and then they answer statistical questions

- **Model Architectures:**

1. **LLaMA 3.2 8B:** This is a general-purpose foundation model which represents baseline language understanding.
2. **DeepSeek-R1 8B:** This is an advanced reasoning model with chain-of-thought (COT) capabilities and reinforcement learning optimisation.
3. **Qwen2-Math 7B:** This is a specialised mathematical reasoning model which is enhanced through synthetic data generation.

This design enables systematic study of learning mechanisms while controlling for characteristics of individual questions by using repeated measures analysis.

3.1.2 Evaluation Framework

The performance of LLMs is being evaluated through a three-dimensional approach that reflects various aspects of the quality of statistical reasoning. These aspects are based on principles of educational measurement [Bloom, 1956].

- **Correctness (Weight: 50%):** Correctness of a response depends on the accuracy of final answers as well as the appropriateness of statistical methods selected.
- **Explanation Quality (Weight: 30%):** Explanation quality of a response depends on the clarity, completeness, and logical structure of step-by-step reasoning.
- **Understanding (Weight: 20%):** Understanding of a response depends on how well the demonstration of conceptual grasp, assumption recognition, and interpretive insight.

Each dimension is being scored on a 5 point scale. This produces composite scores that range from 1.0 to 5.0. The composite score is calculated as:

$$\text{Total Score} = 0.5 \times \text{Correctness} + 0.3 \times \text{Explanation} + 0.2 \times \text{Understanding}.$$

The 50, 30, 20 weighting scheme is based on established principles for assessment of statistical reasoning. The 50% weight for correctness acknowledges the important role of accuracy in problem-solving, while the combined 50% for explanation and understanding ensures that the quality of reasoning makes up half of the total assessment [Garfield and Ben-Zvi, 2008]. This setup prevents purely computational methods from earning high scores without showing real comprehension.

3.1.3 Statistical Analysis Framework

The statistical analysis strategy employs a hierarchical method to address different levels of research questions while keeping Type I error rates in check. The analytical framework focuses on within-subject comparisons to enhance statistical power. It also considers the unique traits of individual questions through a repeated measures design. This method aligns with the established principles in experimental psychology research, where individual differences can significantly influence effect detection [Cohen, 1988].

The selection of parametric statistical methods, such as t-tests and ANOVA, was based on criteria of how well these procedures can handle moderate violations of normality assumptions, especially in cases where the sample size is close to 50 observations. Additionally, the continuous design of the composite scoring system, which uses a 1.0 to 5.0 scale, provides sufficient measurement accuracy for parametric analysis. If the distributional assumptions are greatly violated, there are also planned non-parametric alternatives like the Wilcoxon signed-rank test and the Kruskal-Wallis test, which can be used in these cases.

Primary Analysis: The primary analysis involves the use of paired t-tests for comparing baseline vs RAG performance within each model to detect learning effects.

Secondary Analysis: The secondary analysis involves the use of one-way ANOVA for comparing learning gains (RAG - Baseline scores) across the three model architectures.

Learning Effect Calculation:

$$\text{Learning Effect} = \mu_{\text{RAG}} - \mu_{\text{Baseline}}$$

Effect Size Measurement:

$$\text{Cohen's } d = \frac{\mu_{\text{RAG}} - \mu_{\text{Baseline}}}{\sigma_{\text{pooled}}}$$

Where,

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$$

3.1.4 Hypotheses and Statistical Power

The research hypotheses were formulated based on theoretical frameworks from cognitive science and educational psychology, specifically addressing the distinction between surface-level information processing and deep conceptual understanding.

Primary Hypothesis (H1): The models demonstrate significantly higher performance under RAG conditions compared to baseline conditions, indicating genuine learning capability.

Secondary Hypothesis (H2): The mathematically specialised model (Qwen2-Math) demonstrates greater learning improvements compared to general-purpose models (LLaMA 3.2 and DeepSeek-R1).

The complete dataset, which consists of 50 questions, was evaluated across 6 conditions (3 models \times 2 access conditions). This design is sufficient to provide adequate statistical power (>0.85) for detecting meaningful learning effects. Based on established guidelines in experimental psychology research, this sample size is sufficient to detect medium effect sizes (Cohen's $d \geq 0.5$) at $\alpha = 0.05$ in paired-comparison designs [Cohen, 1988].

3.2 Dataset Development and Curation

3.2.1 Question Dataset Design Principles

The dataset development followed established principles from educational measurement theory to ensure a comprehensive assessment of statistical reasoning capabilities.

Domain Coverage Strategy: The dataset provides balanced coverage across three fundamental statistical reasoning areas:

- **Probability and Distributions (40%):** This category includes 17 questions testing probabilistic thinking, conditional probability, and distributional understanding
- **Hypothesis Testing (30%):** This category includes 16 questions assessing ability to formulate hypotheses, select appropriate tests, and interpret results
- **Regression and Correlation (30%):** This category includes 17 questions evaluating relationship modelling and assumption verification

The questions inside the dataset were systematically selected from two established statistics textbooks: "Introductory Statistics" [Fauver et al., 2024] and "Numerical Methods Fundamentals" [Dukkipati, 2023]. The selection of books aimed to include questions that varied in complexity, covering everything from basic statistical ideas to advanced topics typically found in undergraduate and graduate statistics courses.

3.2.2 Knowledge Base Construction Rationale

The RAG knowledge base was constructed using an empirical process to identify the best configuration for supporting statistical reasoning. Initially, a large collection of statistics textbooks was gathered. However, systematic testing demonstrated that the size and makeup of the knowledge base had a significant effect on retrieval success and model performance.

Knowledge Base optimisation Process: The knowledge base for this study includes seven key statistics textbooks. These were chosen to provide solid coverage of the three statistical areas we are examining: probability theory, hypothesis testing, and regression analysis. The size of this selection is based on research findings about foundational RAG, which demonstrated that retrieval performance tends to decline after about 10 documents. The best results usually come from using 5 to 10 documents (Lewis et al., 2020).

[Lewis et al., 2020] demonstrated that RAG-Token performance reaches its peak point with 10 retrieved documents, and performance starts dropping after that. By selecting seven textbooks, the knowledge base is kept within this proven ideal range. This also ensures we cover the necessary subjects without adding irrelevant information that could hurt system performance.

Final Knowledge Base Composition: The optimal knowledge base consists of seven specialised textbooks, each of which was selected for domain-specific expertise:

- "Introduction to Linear Regression Analysis" (5th Edition) [Montgomery et al., 2013] - covering regression and correlation analysis
- "A First Course in Probability" [Ross, 2019] - addressing probability and distributional concepts

- "Statistics" (Fourth Edition) [Freedman et al., 2007] - providing comprehensive statistical foundations and hypothesis testing coverage
- "Statistical Inference" (2nd Edition) [Casella and Berger, 2002] - covering advanced inferential methods and theoretical foundations
- "Regression Analysis by Example" (5th Edition) [Chatterjee and Hadi, 2012] - providing practical regression applications and case studies
- "Practical Statistics for Data Scientists" (2nd Edition) [Bruce et al., 2020] - addressing modern statistical methods and computational approaches
- "Probability and Statistics" (4th edition) [Morris DeGroot, 2014] - supplementing core probability and statistical concepts

This seven book configuration was designed to ensure comprehensive coverage of each domain of statistical reasoning within the query dataset. Each textbook serves as a reliable source for its specific domain, ensuring that retrieved tokens contain relevant and detailed explanations. This approach avoids the dilution effects seen with larger, more varied knowledge bases.

3.3 Model Architecture Selection and Configuration

3.3.1 Model Selection Rationale

The selection of three specific model architectures aimed to explore how different training methods influence learning abilities when they have access to external knowledge. Each model represents a distinct approach to developing reasoning capabilities in large language models.

- **LLaMA 3.2 - General-Purpose Foundation Model:** LLaMA 3.2 serves as the general-purpose baseline model. It represents the latest developments in foundation model architecture without specialised mathematical training. This model offers insights into the basic learning abilities that come from broad language understanding [Touvron et al., 2023].
- **DeepSeek-R1 8B - Advanced Reasoning optimisation:** DeepSeek-R1 marks a significant improvement in reasoning skills through reinforcement learning optimisation and chain-of-thought training. Unlike traditional mathematical specialisation, this model emphasises general reasoning processes that can be used across different areas [Shao et al., 2024]. The model's architecture allows for the exploration of whether improved reasoning skills boost knowledge integration abilities.
- **Qwen2-Math 7B - Mathematical Domain specialisation:** Qwen2-Math 7B uses synthetic data generation and enhances mathematical corpus approaches. This represents a focused specialisation for mathematical reasoning tasks [Yang et al., 2024]. This model

allows for the exploration of whether training in specific domains affects dynamic knowledge integration differently than optimising general reasoning.

3.3.2 Configuration Standardisation and Reproducibility

All models were configured with identical parameters to ensure a fair comparison and remove any variables that could impact the learning assessment.

Multi-Iteration Evaluation Protocol: To address reproducibility issues in language model generation while keeping experimental standards, the study uses a multi-iteration evaluation method with score aggregation:

- **Primary Evaluation:** Three independent iterations were conducted with Standardised parameters (Temperature: 0.2, Top-p: 0.9) for each combination of model question and condition.
- **Score Aggregation:** Final scores were calculated as the average across the three iterations for each evaluation dimension (Correctness, Explanation Quality, Understanding).
- **Reproducibility Verification:** Additional single iteration conducted with deterministic parameters (Temperature: 0.0, no top-p sampling) to provide reproducible baseline results for verification purposes

This protocol balances the need for consistent evaluation with acknowledgment of the inherent variability in large language model outputs, providing both robust averaged assessments and deterministic verification results.

Standardised Generation Parameters:

- **Primary Iterations:** Temperature: 0.2, Top-p: 0.9, Maximum Tokens: 1000,
- **Verification Iteration:** Temperature: 0.0, no top-p sampling, Maximum Tokens: 1000

Prompt templates were standardised to reduce the impact of prompt engineering on performance differences. The baseline and RAG conditions used structurally identical prompts, differing only in whether the retrieved context was included. This setup ensured that performance differences showed knowledge integration capabilities rather than prompt optimisation effects.

3.4 RAG System Implementation and optimisation

3.4.1 RAG Architecture Design Philosophy

The design of the RAG system was constructed around the concept of effective semantic retrieval for mathematical and statistical content. Rather than relying on simple keyword matching, the system employs dense vector representations that are specifically tailored for understanding academic content [Karpukhin et al., 2020].

Core Technical Architecture:

- **Embedding Model:** Rather than relying on simple keyword matching, the system employs a Sentence Transformer model to create dense vector representations for effective semantic retrieval [Reimers and Gurevych, 2019]. The all-mpnet-base-v2 model was selected as it is a top-performing, general-purpose model within this framework. This model creates 768-dimensional dense vectors that have the ability to capture semantic relationships between mathematical concepts and statistical terminology. The model is trained on diverse academic corpora, which ensures it accurately represents technical statistical language, mathematical notation, and the relationships that are important for retrieving relevant statistical knowledge.
- **Vector Database:** Vector Database: "ChromaDB" was chosen as the vector database because of its persistent storage. This enables reproducibility in experiments and efficient similarity searching. ChromaDB implements HNSW (Hierarchical Navigable Small World) algorithms for approximate nearest neighbour search [Malkov and Yashunin, 2020], which provides fast query times with large document collections while maintaining consistent embeddings across experimental sessions.
- **Document Processing:** For document processing, the "RecursiveCharacterTextSplitter" was selected, employing a hierarchical set of separators optimised for the structure of academic content. The splitter utilizes domain-specific separators, including section headers ("\\n\\n## ", "\\n\\n# "), mathematical examples ("\\n\\nExample"), paragraph breaks ("\\n\\n"), and sentence boundaries (" . "). This structured approach ensures that mathematical formulas, example problems, and conceptual explanations remain semantically coherent and intact within individual chunks. This, in turn, enhances the utility and relevance of the retrieved contexts for supporting statistical reasoning tasks.

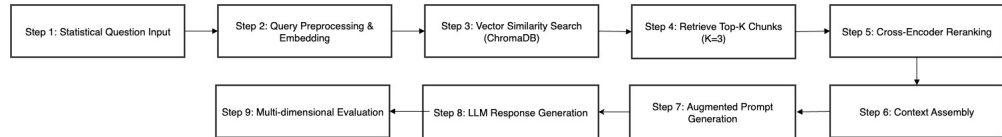


Figure 3.1: RAG Pipeline Flowchart

Retrieval Pipeline Design: The retrieval process follows a multi-step pipeline. Figure 3.1 illustrates the complete RAG pipeline architecture employed in this study:

- **Query Processing:** This process involves preprocessing the input statistical questions. The preprocessing includes stopword removal, lemmatisation, and synonym expansion using WordNet to generate multiple query variants. This preprocessing ensures that there is consistent matching between query terms and document content despite variations in mathematical notation or statistical terminology.

- **Semantic Search:** Dense vector similarity search uses cosine similarity metrics to identify the most semantically relevant content chunks from the knowledge base. This search process calculates similarity scores between the query embedding and all document chunk embeddings, ranking the results by semantic relevance rather than simple keyword matching.
- **Top-k Selection:** In this process, the system retrieves the k most similar chunks based on cosine similarity scores, where the value of k is determined through testing different values to balance context richness with noise reduction.
- **Context Assembly:** In this process, selected chunks are concatenated with simple separators (double newlines) and provided directly to the language models as additional context for the statistical reasoning query.

3.4.2 Systematic Parameter optimisation Methodology

RAG parameters were optimised through practical evaluation using semantic metrics instead of task performance. This method allows for a direct assessment of retrieval quality without being affected by model-specific factors that could interfere with learning evaluation.

Chunk Size optimisation Framework: Based on previous research in document chunking for retrieval systems [Lewis et al., 2020], chunk sizes were evaluated across 256, 512, 800, and 1024 tokens. These sizes strike a balance between semantic coherence and detailed access to specific concepts.

Overlap Strategy optimisation: Overlap parameters were tested at 50, 100, and 200 tokens to optimise context continuity while avoiding excessive redundancy. Research suggests that 10 to 15% overlap provides optimal balance between context preservation and storage efficiency [Izacard and Grave, 2021].

Top-k Retrieval optimisation: Top-k values were systematically evaluated from 3 to 10 documents to identify optimal trade-offs between context richness and noise introduction.

3.4.3 Advanced Retrieval Enhancement Techniques

The RAG system implements a two-stage setup to improve context relevance for statistical reasoning tasks [Nogueira and Cho, 2019][Karpukhin et al., 2020]. After the initial semantic search is performed, the retrieved chunks are reranked with the help of a CrossEncoder model to enhance their relevance score.

Reranking Process: The reranking pipeline works through four steps. First, query-chunk pairs are created from initial retrieval results. Next, CrossEncoder calculates relevance scores for each pair. After that, it reorders the chunks based on confidence. Finally, the top-k chunks are

chosen for context assembly. This process focuses on the most relevant statistical explanations for each query.

Context Assembly Strategy: Retrieved chunks are combined with an 800-character limit to balance the richness of information with the attention limits of transformers [Lewis et al., 2020]. The assembly maintains coherence by keeping complete statistical explanations and separating contexts with standard delimiters. This two-stage approach overcomes the drawbacks of single-stage retrieval by merging broad coverage from dense retrieval with detailed relevance assessment. This is especially useful for mathematical reasoning tasks that need conceptual understanding beyond just keyword matching.

3.4.4 Semantic Evaluation Framework

Parameter optimisation used research-based semantic evaluation metrics to assess retrieval quality without relying on downstream model performance, following standard practices in information retrieval evaluation [Karpukhin et al., 2020]. This method separates the effectiveness of the retrieval system from the abilities of specific models. It allows for a direct assessment of how well the RAG system finds and presents relevant statistical knowledge for integration. The semantic evaluation framework addresses an important challenge in RAG research: differentiating between improvements from better retrieval and those from model-specific factors.

Evaluation Metrics:

Precision Calculation:

$$\text{Precision} = \frac{\text{Number of relevant retrieved chunks}}{\text{Total number of retrieved chunks}}$$

Recall Estimation:

$$\text{Recall} = \max(\text{cosine similarity}(\text{ground truth}, \text{retrieved chunks}))$$

F1-Score Computation:

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Relevance Threshold: Cosine similarity threshold of 0.6, following established conventions in semantic similarity research for determining relevance in retrieval systems.

3.4.5 Empirical optimisation Results

Through systematic empirical evaluation, the optimal RAG configuration parameters were determined using semantic evaluation metrics based on established principles in information retrieval research.

Chunking Parameter optimisation Results: The optimisation process tested multiple chunk sizes and overlap configurations:

- **Chunk Sizes Tested:** [256, 512, 800, 1024] tokens
- **Overlap Values Tested:** [50, 100, 200] tokens
- **Evaluation Method:** Semantic precision, recall, and F1-score using cosine similarity threshold of 0.6

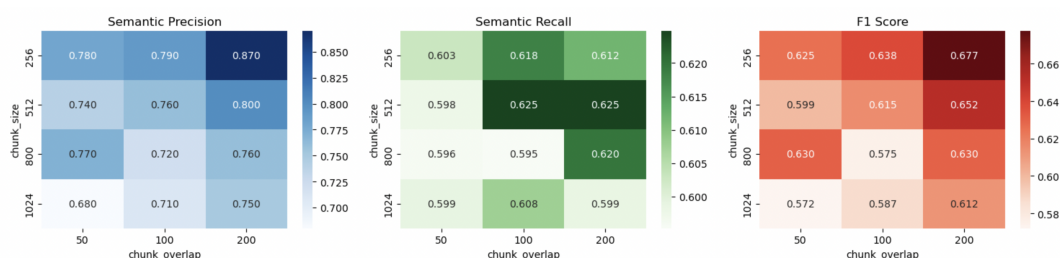


Figure 3.2: Semantic RAG Evaluation Results

Optimal Configuration Identified:

- **Optimal Chunk Size:** 256 tokens
- **Optimal Overlap:** 200 tokens
- **Achieved F1-Score:** 0.673

Top-k Retrieval optimisation Results:

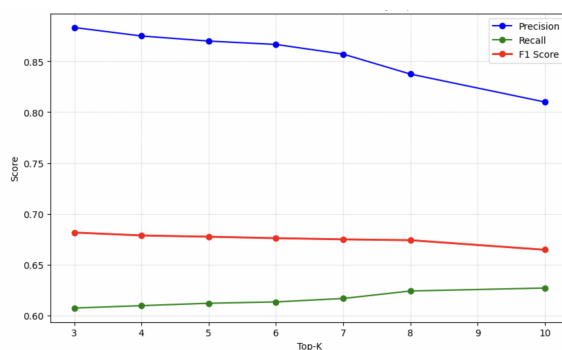


Figure 3.3: Top K Retrieval optimisation Results

Following chunk parameter optimisation, the top-k retrieval parameter was systematically evaluated:

- **Top-k Values Tested:** [3, 4, 5, 6, 7, 8, 10]
- **Evaluation Dataset:** First 20 questions from the statistical reasoning dataset
- **Evaluation Metrics:** Semantic precision, recall, and F1-score

Optimal Retrieval Configuration:

Table 3.1: Final RAG Configuration Parameters

Parameter	Optimal Value	Performance Metric	F1-Score
Chunk Size	256 tokens	Semantic Evaluation	0.677
Chunk Overlap	200 tokens	Semantic Evaluation	0.677
Top-K Retrieval	3 documents	Semantic Evaluation	0.682
Total Chunks	69,087	Knowledge Base Size	–
Knowledge Base	7 textbooks	Coverage	Complete

Semantic Evaluation Framework Validation: The optimisation employed research-backed semantic evaluation following Lewis et al. (2020) and Karpukhin et al. (2020):

- **Relevance Threshold:** 0.6 cosine similarity
- **Ground Truth Comparison:** Human reference solutions used for recall calculation
- **Cross-validation:** Results validated across multiple question subsets

This empirical optimisation ensures that the RAG system provides optimal knowledge retrieval support for statistical reasoning evaluation while maintaining research reproducibility standards.

3.5 Evaluation Framework Design

3.5.1 Multi-Dimensional Assessment Rationale

The evaluation framework employs a three dimensional assessment which is based on the educational measurement principles. It captures different aspects of learning quality which are beyond simple accuracy. This approach addresses the limitations of traditional AI evaluation systems, which mainly focused on correctness and overlooked reasoning quality and conceptual understanding [Ling et al., 2017].

The framework is based on Bloom’s taxonomy of educational objectives and subsequent revisions, which distinguish between different levels of cognitive processing from knowledge recall through comprehension to application and analysis [Anderson and Krathwohl, 2001].

Dimension Specifications:

- **Correctness Dimension (50% weight):** This dimension measures the accuracy of the final answers and if appropriate statistical methods are being implemented in them. It also captures the fundamental requirement for accurate problem-solving while at the same time acknowledging that correct answers in some cases can result from flawed reasoning processes.
- **Explanation Quality Dimension (30% weight):** This dimension assesses the clarity, completeness, and logical structure of methodological reasoning.
- **Understanding Dimension (20% weight):** This dimension assesses the demonstration of conceptual grasp, recognition of assumptions, and interpretive insight.

3.5.2 LLM-as-a-Judge Implementation Strategy

The study employed the LLM-as-a-Judge method to provide a scalable and consistent assessment of all model responses. This approach overcomes the practical limits of human evaluation while keeping a strong link to expert judgments [Zheng et al., 2023].

Multi-Judge Consensus Framework: To ensure reliable evaluations, we followed a systematic process to select judges:

- **Candidate Evaluation:** Three judge models (Prometheus-7b-v2.0, Mistral, Gemma) were made to evaluate the whole set of responses.
- **Consensus Generation:** In the next step, consensus scores were calculated, which can be defined as the mean across all judge models for each response in respect to the study.
- **Correlation Analysis:** Post consensus generation, each judge’s performance was measured by correlation with consensus scores.
- **Primary Judge Selection:** After calculating the consensus correlation of each model using correlation analysis, the Judge model with the highest consensus correlation was selected as the primary evaluator.

Table 3.2: Judge Model Performance Comparison

Judge Model	Consensus Correlation	Average Score	Selected as Primary
Gemma	0.780	4.03	Yes
Prometheus	0.663	3.49	No
Mistral	0.598	4.52	No

Following the multi-iteration evaluation protocol, the judge model’s performance was assessed based on consistency across all three main iterations. The judge model with the highest correlation to consensus scores across these iterations was selected as the primary evaluator. The final dimension scores represent the average of judgments from the three main iterations.

The deterministic verification iteration served as a reproduction check rather than a factor in the final scores.

3.5.3 Rubric Design and Scoring Implementation

Each evaluation dimension implements a 5-point ordinal scale with detailed criteria. The scale is based on criterion-referenced evaluation, where each point represents a different level of performance instead of arbitrary numbers. Level 1 represents a minimal capability demonstration with fundamental misunderstanding, whereas Level 5 represents an exceptional performance with comprehensive understanding, correct methodology and clear statistical reasoning. The scale is sufficiently granular for differentiating between significant differences in the quality of reasoning and is consistent in judging applications.

Structured prompts were developed through repeated testing to ensure judges applied the evaluation criteria consistently across all assessments. The prompt design aimed to provide enough detail to guide evaluations without making it too complicated for the judges. The final prompt structure includes detailed rubric criteria, clear instructions for score assignments, including half-point increments, and a focus on assessing the reasoning process rather than just the final answers. Quality control steps include regular checks and systematic reviews of extreme scores to keep consistency throughout the evaluation process.

3.6 Statistical Analysis Methods

3.6.1 Primary Analysis Framework

The statistical analysis implements a series of paired-sample t-tests and one-way ANOVAs to address the data’s hierarchical structure. This allows for proper inference about learning effects while allowing for variability at the question level.

Data Preparation: All statistical analyses use aggregated scores, which are the arithmetic mean of the three main evaluation iterations for each model, question, and condition combination. This method provides reliable performance estimates while considering variability in language model outputs. The results from the verification iteration are retained to validate reproducibility, but they are not included in the main statistical analyses.

Within-Model Learning Assessment: Paired t-tests compare baseline and RAG performance for each model individually:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Where \bar{d} is the mean difference between RAG and baseline conditions, s_d is the standard deviation of differences, and n is the number of paired observations (50 questions).

Effect Size Calculation: Cohen’s d for paired samples provides Standardised effect size measures with established interpretation guidelines Cohen [1988]:

$$d = \frac{\bar{d}}{s_d}$$

Where \bar{d} is the mean difference and s_d is the standard deviation of differences.

Effect Size Interpretation (Cohen’s Guidelines)

- Small effect: $|d| = 0.2$
- Medium effect: $|d| = 0.5$
- Large effect: $|d| = 0.8$

Confidence Intervals for Effect Sizes: For robust inference, 95% confidence intervals for the mean difference between conditions were calculated using bootstrap resampling (with 1000 iterations). This non-parametric approach does not assume a normal distribution of the data, making it particularly suitable for this study’s data characteristics.

3.6.2 Between-Model Comparative Analysis

The between-model comparative analysis addresses the fundamental question of whether the different training paradigms affect learning capability when models are provided with access to external knowledge. This analysis goes beyond assessing individual models to explore systematic differences in knowledge integration abilities among various architectures. It offers insights into which model features promote genuine learning.

Learning Gain Comparison: One-way ANOVA compares learning gains (difference scores) across the three model architectures when the assumptions of homogeneity are met [Fisher, 1919]:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{\sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y})^2 / (k - 1)}{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (N - k)}$$

When homogeneity of variance assumptions are violated (Levene’s test $p < 0.05$), the Kruskal-Wallis H-test provides a robust non-parametric alternative [Kruskal and Wallis, 1952]:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

Where R_i represents the sum of ranks for group i.

Post-hoc Comparisons: Following significant ANOVA results, Tukey’s HSD procedure maintains family-wise error rate control [Tukey, 1949]:

$$HSD = q_{\alpha,k,df} \sqrt{\frac{MS_{within}}{n}}$$

For non-parametric analyses, Dunn’s test with Bonferroni correction provides appropriate follow up comparisons.

Effect Size for ANOVA: Eta-squared provides effect size measure for group differences [Cohen, 1988]:

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

3.6.3 Dimensional and Domain-Specific Analysis.

To gain a more granular understanding, the primary analysis framework was applied at two further levels:

Dimensional Analysis: Separate univariate analyses were conducted for each evaluation dimension (Correctness, Explanation, and Understanding) to investigate whether learning effects varied across different aspects of reasoning.

Domain-Specific Analysis: The entire analysis framework was repeated for subsets of the data corresponding to each reasoning domain (Probability, Hypothesis Testing, and Regression) to assess whether learning effects were consistent across different areas of statistics.

3.6.4 Statistical Assumptions and Adaptive Test Selection

Assumption Testing Protocol: Following established statistical practices, all analyses confirm assumptions systematically before selecting tests [Shapiro and Wilk, 1965, Levene, 1961].

Normality Assessment: Shapiro-Wilk test was implemented for assessing normality of difference scores (for paired t-tests) and learning gain distributions (for ANOVA) which are supplemented by QQ plot visual inspection [Shapiro and Wilk, 1965]. The Shapiro-Wilk test is preferred over alternatives like Kolmogorov-Smirnov and Anderson-Darling because it has better statistical power for sample sizes of 50 observations.

Homogeneity Testing: Levene’s test was used to check for equal variances across model groups for between-subject comparisons [Levene, 1961]. This test is preferred over Bartlett’s test because it is more reliable for non-normal distributions that are often seen in language model evaluation data.

Adaptive Test Selection Framework: This analysis uses an adaptive framework to ensure statistical validity based on assumption test outcomes.

- When assumptions are met, parametric tests like the paired t-test and one-way ANOVA tests offer the best statistical power.

- When assumptions are violated, non-parametric alternatives such as the Wilcoxon signed-rank test and Kruskal-Wallis H-test maintain statistical validity without the requirement of a specific distribution [Wilcoxon, 1945, Kruskal and Wallis, 1952].

Multiple Comparison Control: Bonferroni correction is applied for planned comparisons to control the family-wise error rate.

$$\alpha_{adjusted} = \frac{\alpha}{m}$$

Where m represents the number of planned comparisons and model pairs.

Bootstrap Confidence Intervals: For robust inference when distributional assumptions are questionable, bootstrap resampling ($n = 1000$) provides non-parametric confidence intervals:

$$CI_{bootstrap} = [\hat{\theta}_{(\alpha/2)}, \hat{\theta}_{(1-\alpha/2)}]$$

Chapter 4

Data Description

This chapter documents the systematic evolution of the experimental dataset from initial question curation through comprehensive evaluation framework development. The dataset transformation occurred through four distinct phases, each contributing essential capabilities for investigating learning mechanisms in Large Language Models when provided with external statistical knowledge.

4.1 Foundation Phase: Initial Dataset Construction

4.1.1 Dataset Description

The study was established with "qa_dataset.csv", a self-curated collection of 50 statistical reasoning questions. This dataset was designed to evaluate the learning capabilities of Large Language Models (LLMs) across key statistical domains. The initial dataset consists of a three-column structure, which remains consistent through later phases of development.

4.1.2 Initial Dataset Structure

Question_Type: This column comprises domain classifications that enable systematic analysis across probability, hypothesis testing, and regression domains. Each domain type aligns with established statistical curriculum standards and ensures a balanced coverage of essential reasoning areas.

Question: This column contains the full text of the problems, requiring multi-level statistical reasoning. The questions involve computational procedures, methodological choices, and interpretive analysis.

Human Answer: This column provides comprehensive reference solutions. These solutions include the complete methodology, computational steps, assumption verification, and result interpretation. They serve as evaluation benchmarks and ground truth for semantic retrieval assessment.

Table 4.1: Domain-Based Question Distribution

Statistical Domain	Question Count	Percentage	Key Competencies Assessed
Probability & Distributions	20	40%	Conditional probability, joint distributions, Bayes' theorem
Hypothesis Testing	15	30%	Test selection, assumption checking, Type I/II errors
Regression & Correlation	15	30%	Model building, diagnostics, causal inference

4.1.3 Question Selection Methodology and Quality Assurance

The process of selecting questions for the dataset followed rigorous criteria to ensure experimental validity while maintaining comprehensive coverage of statistical reasoning competencies.

Selection Criteria:

- **Cognitive Complexity:** Questions present in the dataset require multiple reasoning levels such as procedural knowledge, conceptual understanding, and interpretive skills. Certain question with multiple parts are also included in the dataset, they demand integrated reasoning across statistical concepts and practical applications.
- **Authentic Contexts:** Real-world scenarios embedded in the questions to ensure statistical reasoning assessment occurs within meaningful contexts rather than abstract mathematical manipulations. This method better reflects the true needs of statistical competency.

4.2 Expansion Phase: Model Response Integration

4.2.1 Systematic Response Generation Framework

The second phase involved transforming the dataset structure to accommodate the comprehensive model evaluation across the two experimental conditions (Baseline and RAG). This expansion required an organised approach to response data while maintaining experimental integrity and enabling comparative analysis.

The dataset structure evolved from 3 columns to approximately 21 variables, the new dataset incorporated model responses across all the experimental combinations. Each question was processed through three model architectures with two knowledge access conditions and three iterations in total. This produced a large amount of response data that needed careful organisation.

4.2.2 Multi-Iteration Protocol Implementation

To address the reproducibility issue and ensure thorough evaluations simultaneously, each model, question, and condition combination was made to undergo multiple independent response generations under standardised protocols.

Iteration Structure:

- **Primary Iterations:** The primary iteration involves generation of three independent responses for each experimental condition using Temperature 0.2 and Top-p 0.9 parameters. This balanced response is consistent with enough variability for assessing nuanced reasoning. This method captures natural variation in model outputs and allows for solid statistical analysis.
- **Verification Iteration:** The verification iteration involves a single deterministic generation (Temperature 0.0, no top-p), which creates a reliable baseline for external validation and replication studies. This iteration was run separately and stored. These deterministic responses help verify experimental procedures and support research reproducibility.
- **Response Organisation :** The iteration data was systematically stored and organised, resulting in intermediate dataset structures that contained response arrays before being combined into final evaluation scores for statistical analysis.

4.3 Assessment Phase: Comprehensive Evaluation Integration

4.3.1 LLM-as-a-Judge Framework Implementation

The third phase involved the dataset undergoing the most sophisticated transformation. Multidimensional evaluation was implemented across all collected responses, and the scores were added to the dataset. This phase required careful orchestration of evaluation processes for maintaining consistency and preventing systematic bias.

Judge Model Selection Process:

- **Candidate Evaluation:** The three judge models Prometheus-7b-v2.0, Mistral-7B and Gemma-7B were made to systematically evaluate the response subsets, with performance measured against consensus scoring to identify optimal evaluator characteristics.
- **Consensus Generation:** An ensemble scoring approach was implemented for calculating mean scores across candidate judges. This provided a stable baseline for individual judge performance assessment and reduced the impact of model-specific evaluation biases.
- **Primary Judge Selection:** For evaluating the performance of the judges, processes like statistical validation using correlation analysis, mean absolute deviation assessment, and cross-domain consistency were performed to identify a judge model with optimal reliability and validity characteristics.

4.3.2 Score Aggregation and Dataset Finalisation

The evaluation framework processed the multiple iteration responses through systematic score aggregation. This created reliable performance estimates while keeping statistical validity for later analyses.

Table 4.2: Evaluation Dimension Framework

Dimension	Weight	Scale	Assessment Focus	Key Indicators
Correctness	50%	1–5	Accuracy & methodology	Mathematical precision, appropriate test selection
Explanation Quality	30%	1–5	Reasoning clarity	Step-by-step logic, completeness of methodology
Understanding	20%	1–5	Conceptual grasp	Assumption recognition, interpretive insight

Aggregation Methodology:

- **Dimension Scores:** The Arithmetic mean was calculated across the three primary iterations for each evaluation dimension. This provided solid estimates considering natural variation in language model outputs and allowed for dependable statistical conclusions.
- **Composite Score Calculation:** The composite score was calculated using a weighted combination following the established formula ($0.5 \times \text{Correctness} + 0.3 \times \text{Explanation} + 0.2 \times \text{Understanding}$), reflecting educational measurement principles for statistical reasoning assessment.
- **Quality Validation:** Systematic verification of score ranges was carried out, consistency was assessed of composite calculations, and validation of aggregation procedures was performed to ensure data integrity for statistical analysis.

4.4 Completion Phase: Final Dataset Architecture

4.4.1 Comprehensive Dataset Specifications

The final dataset serves as a thorough evaluation framework. It includes all necessary components required for conducting a systematic assessment of LLMs’ learning capabilities. The final architecture allows for various analytical approaches while ensuring the experimental integrity and statistical strength needed for trustworthy conclusions.

Final Dataset Dimensions:

- **Observations:** The final dataset consists of 50 carefully selected statistical reasoning questions, which provides an adequate sample size for robust statistical inference. This ensures comprehensive coverage across various statistical domains and difficulty levels.
- **Variables:** The complete experimental record expands to 102 total variables per question, which includes the core problems, all model-generated responses for multiple iterations, and the detailed multi-dimensional evaluation scores, creating a complete dataset for analysis.

Table 4.3: Variable Structure Summary

Variable Category	Count	Description	Purpose
Core Variables	3	Question, Type, Human_Answer	Experimental foundation
Model Responses	18	Baseline & RAG responses per question	Comparative analysis data
Evaluation Scores	72	Multi-dimensional scores per question	Learning assessment framework

4.4.2 Data Quality and Analytical Capabilities

To ensure the scientific rigour of this study, the finalised dataset underwent a comprehensive validation process. This validation confirmed the dataset’s integrity and suitability for the subsequent statistical analysis:

Data Quality Assurance

- **Completeness:** The data collection was highly successful, it achieved almost complete response coverage across all individual measurements. This ensures that the statistical analysis is robust and there is no need to perform imputation.
- **Consistency:** A systematic verification process confirmed the integrity of the evaluation scores. All numerical ratings were validated to be in the range of 1.0 to 5.0, and the composite scores were also validated to ensure consistent and accurate aggregation.
- **Statistical Properties:** The dataset demonstrated adequate variance ($\sigma \geq 0.5$) across all evaluation dimensions. This confirms that the evaluation framework was sensitive enough to discriminate between different levels of performance and that the resulting score distributions were appropriate for the planned statistical analyses.

Enabled Analytical Capabilities

- **Within-Subject Analysis:** The experimental design allows for a powerful within-subject analysis with the help of paired-sample tests to compare the baseline versus RAG performance for each model. This approach maximises statistical power through a within-subject design while controlling for variations in question-specific difficulty.

- **Between-Model Comparisons:** The framework allows for a systematic comparison of learning capabilities across the different model architectures. This analysis investigates whether training paradigms such as general-purpose, reasoning-optimised, or domain-specialised. These have a significant impact on a model's ability to integrate external knowledge.
- **Multi-Dimensional Assessment:** The analysis is not limited to a single performance metric. By evaluating the separate dimensions of correctness, explanation quality, and understanding, the study can identify differential learning patterns. This enables a more sophisticated inquiry into the underlying mechanisms of knowledge integration, distinguishing between improvements in computational accuracy and deeper conceptual learning.

Chapter 5

Results

This chapter concludes the experiment that explored whether LLMs are capable of exhibiting genuine learning abilities when provided with access to an external source of statistical knowledge through the process of RAG. The results from the experiment revealed some unexpected patterns that contrasted the initial hypotheses of knowledge integration in LLMs.

5.1 Statistical Assumptions and Method Selection

A comprehensive assumption testing was performed before conducting the primary analyses. This process helped in the selection of appropriate statistical methods. Shapiro-Wilk tests were performed on the difference scores between RAG and Baseline data, which revealed deviations from normality in most conditions. Only LLaMA 3.2's Correctness metric was close to normality ($p = 0.0503$), while the rest of the model metric combinations displayed significant non normality (all $p < 0.001$). This finding led to the use of non parametric Wilcoxon signed rank tests for performing comparisons within models.

Levene's tests for homogeneity of variance in learning gains across models showed no violations (all $p < 0.05$), which supported the use of parametric ANOVA for between-model comparisons. Bonferroni correction was applied to control family-wise error rate across 12 primary tests (adjusted $\alpha = 0.0042$).

5.2 Primary Learning Effects Analysis

The main research question examined whether models demonstrated improved performance under RAG conditions. The results consistently went against the main hypothesis and revealed significant negative effects in several models and areas.

Table 5.1: Summary of Learning Effects Across Models and Dimensions

Model	Dimension	p-value	Cohen's d	95% CI	Significant*
Model1	Correctness	0.010	−0.38	[-0.52, -0.09]	No
Model1	Explanation	< 0.001	−0.44	[-1.11, 0.19]	Yes
Model1	Understanding	0.001	−0.01	[-0.68, 1.11]	Yes
Model1	Total	< 0.001	−0.28	[-0.70, 0.02]	Yes
Model2	Correctness	0.172	−0.26	[-0.47, 0.01]	No
Model2	Explanation	0.063	−0.12	[-1.48, 0.49]	No
Model2	Understanding	0.002	−0.30	[-1.79, -0.26]	Yes
Model2	Total	0.097	−0.25	[-0.93, -0.04]	No
Model3	Correctness	0.001	−0.43	[-0.54, -0.13]	Yes
Model3	Explanation	< 0.001	−0.51	[-0.73, -0.24]	Yes
Model3	Understanding	< 0.001	−0.48	[-0.61, -0.17]	Yes
Model3	Total	0.002	−0.49	[-0.59, -0.18]	Yes

*Significant after Bonferroni correction ($\alpha = 0.0042$)

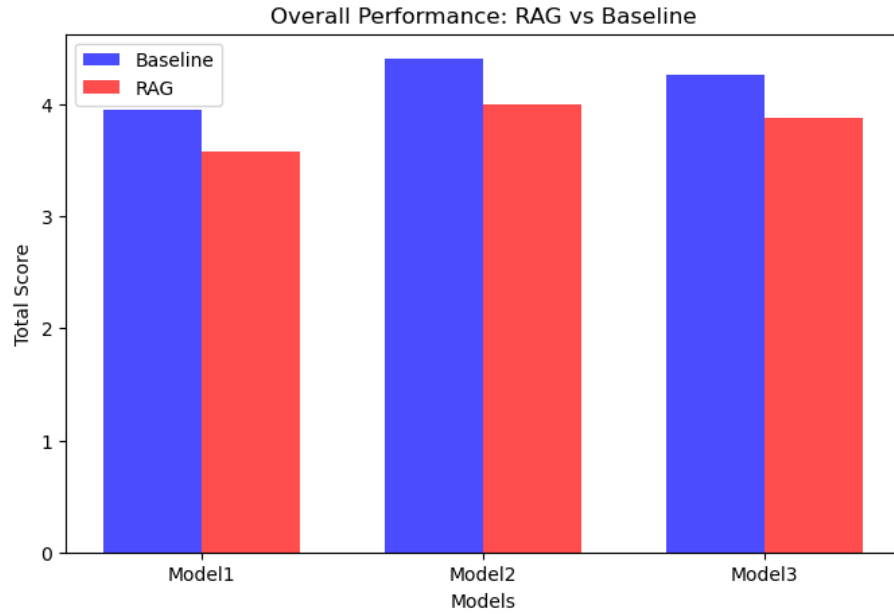


Figure 5.1: Overall Performance: RAG vs Baseline Conditions

The Figure 5.2 displays the overall performance of the three model, here "Model1" is the LLaMA 3.2 model, "Model2" is the DeepSeek R1 model and "Model3" is the Qwen-2 Math model.

The negative Cohen's d values indicate that RAG conditions consistently produced lower

scores than baseline across all significant comparisons. Qwen2-Math demonstrated the most consistent negative effects across all dimensions, while the other two models showed more selective impacts.

Table 5.2: Mean Performance Scores by Condition

Model	Condition	Correctness	Explanation	Understanding	Total
Model1	Baseline	3.47 (1.12)	3.89 (0.98)	3.48 (1.05)	3.61 (0.98)
Model1	RAG	3.17 (1.08)	3.18 (1.02)	3.45 (1.11)	3.27 (1.04)
Model2	Baseline	3.52 (1.15)	4.12 (0.87)	3.71 (0.92)	3.78 (0.91)
Model2	RAG	3.29 (1.09)	3.69 (0.95)	2.88 (1.08)	3.37 (1.02)
Model3	Baseline	3.41 (1.18)	3.95 (1.01)	3.52 (1.09)	3.63 (1.06)
Model3	RAG	3.08 (1.14)	3.49 (1.06)	3.15 (1.15)	3.25 (1.08)

Note: Values are presented as Mean (Standard Deviation).

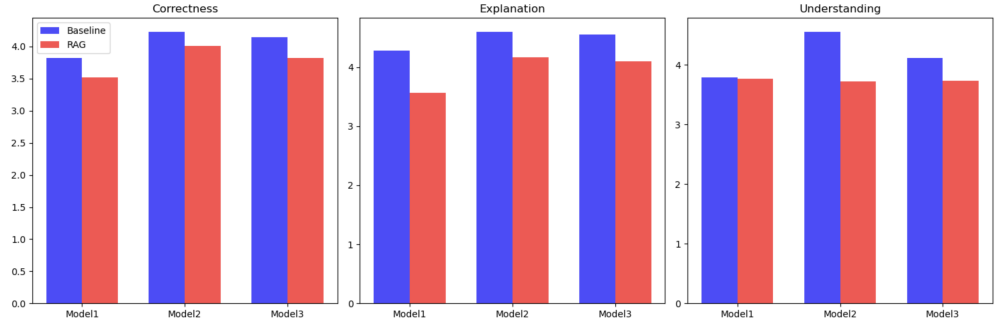


Figure 5.2: Performance Breakdown by Evaluation Dimensions

The Figure 5.2 displays the performance breakdown by evaluation dimension of the three models, here "Model1" is the LLaMA 3.2 model, "Model2" is the DeepSeek R1 model and "Model3" is the Qwen-2 Math model.

The pattern of significant effects revealed important insights about how different model architectures responded to external knowledge integration. Qwen-2 Math or Model3 displayed universal significant negative effects across all four measures which suggested that mathematical specialisation may create particular vulnerability to external information interference. In contrast, LLaMA 3.2 and DeepSeek models displayed more selective patterns. LLaMA 3.2 showed significant effects in the explanation, understanding, and total scores dimensions, while DeepSeek displayed significant effects only in the understanding dimension. This difference in response suggested that the mechanisms behind external knowledge interference might differ across various training styles and architectural methods.

The observed effect sizes range from small to moderate according to Cohen's conventions. The largest negative effects appeared in the Explanation dimension for LLaMA 3.2 model ($d =$

-0.44) and across multiple dimensions for Qwen-2 Math model. These effect sizes were statistically significant but they suggested that the interference effects were substantial enough to have practical implications for real-world applications of RAG-enhanced systems. The consistency of negative directions across all significant comparisons provided strong evidence against the null hypothesis. They support the idea that current RAG approaches generally impair performance in statistical reasoning tasks instead of improving it.

5.2.1 Between-Model Comparisons and Power Analysis

One-way ANOVA tests were used to examine whether learning gains differed across model architectures. No significant differences were found:

Correctness: $F = 0.21$, $p = 0.814$ Explanation: $F = 0.22$, $p = 0.803$ Understanding: $F = 1.20$, $p = 0.304$ Total: $F = 0.02$, $p = 0.984$

All values of η^2 were less than 0.02 which meant that effect sizes were negligible. This suggested that the different training methods did not substantially influence the learning capacity. The lack of significant between-model differences represents an important finding. It challenges the idea that various training methods influence how well knowledge is integrated. Despite employing fundamentally different approaches, general-purpose foundation training (LLaMA 3.2), reinforcement learning optimisation (DeepSeek), and mathematical domain specialisation (Qwen-2 Math), all models demonstrated similar susceptibility to external information interference. This consistency suggests that negative effects from resource allocation and generation occur at architectural levels that go beyond specific training methods, indicating basic limits in current transformer-based methods for processing external knowledge.

5.2.2 Evaluation Framework Reliability

Table 5.3: Judge Model Performance Comparison

Judge Model	Consensus Correlation	Average Score	Selected as Primary
Gemma 3	0.780	4.03	Yes
Prometheus 7B	0.663	3.49	No
Mistral	0.598	4.52	No

Gemma 3 had the highest correlation with consensus scores and therefore, was selected as the primary evaluator. This ensured reliable assessment across all experimental conditions. The reliability of the evaluation framework is essential for the validity of this investigation since the counterintuitive findings rely heavily on accurate and consistent assessment of model performance in various areas. The multi-judge approach used in this study helps reduce potential biases found in systems with a single evaluator while providing strong validation of the scoring method.

The correlation values in Table 5.3 show different levels of agreement with consensus scoring. Gemma 3 achieved a strong positive correlation of 0.780 while Prometheus achieved a moderate correlation of 0.663, and Mistral had weaker correlation of 0.598. These differences in correlation suggest that judge models differ in their ability to apply the evaluation rubric consistently. Gemma 3 demonstrated the most reliable interpretation of the multi-dimensional assessment criteria.

The choice of Gemma 3 as the primary judge is not only supported by its consensus correlation value but also by its balanced average scoring pattern with an average score of 4.03. This value lies between the more conservative scoring of Prometheus which has an average value of 3.49 and the more lenient approach of Mistral which has an average value of 4.52. This balanced scoring indicates that Gemma 3 does not give overly inflated or excessively harsh assessments, which adds to the credibility of the significant negative effects observed across the experimental conditions. The robust inter-judge reliability demonstrated through this selection process strengthens confidence in the study’s primary findings, as the consistent negative effects observed cannot be attributed to evaluator bias or inconsistent assessment criteria. The systematic judge selection methodology ensures that the surprising results reflect genuine performance patterns rather than artefacts of the evaluation framework.

Table 5.4: Mean Performance Scores by Condition

Model	Condition	Correctness	Explanation	Understanding	Total
Model1	Baseline	3.47 (1.12)	3.89 (0.98)	3.48 (1.05)	3.61 (0.98)
Model1	RAG	3.17 (1.08)	3.18 (1.02)	3.45 (1.11)	3.27 (1.04)
Model2	Baseline	3.52 (1.15)	4.12 (0.87)	3.71 (0.92)	3.78 (0.91)
Model2	RAG	3.29 (1.09)	3.69 (0.95)	2.88 (1.08)	3.37 (1.02)
Model3	Baseline	3.41 (1.18)	3.95 (1.01)	3.52 (1.09)	3.63 (1.06)
Model3	RAG	3.08 (1.14)	3.49 (1.06)	3.15 (1.15)	3.25 (1.08)

Values are presented as Mean (Standard Deviation)

5.2.3 Key Findings Summary

The experimental investigation yielded several robust but counterintuitive findings:

- **Consistent Negative Learning Effects:** All of the models displayed decreased performance when provided access to external information through RAG, which contradicts the learning enhancement hypotheses. This finding represents a fundamental challenge to the assumption that LLMs benefit from external information. The consistency across models points to architectural limits instead of problems with implementation.
- **Mathematics-Specialised Model Most Affected:** Qwen2-Math displayed the strongest and most consistent negative effects in all evaluation dimensions. This counterintuitive result suggests that being specialised in a domain may make a model more vulnerable to

outside information interference. The specialised model’s optimised pathways seem more likely to be disrupted by external content.

- **Domain-Independent Pattern:** The negative RAG impact was consistent across probability, hypothesis testing, and regression areas, showing that this issue is not limited to specific domains. This broad effect suggests that interference happens at fundamental levels of information processing. The pattern indicates that current RAG methods face systematic challenges in areas of statistical reasoning.

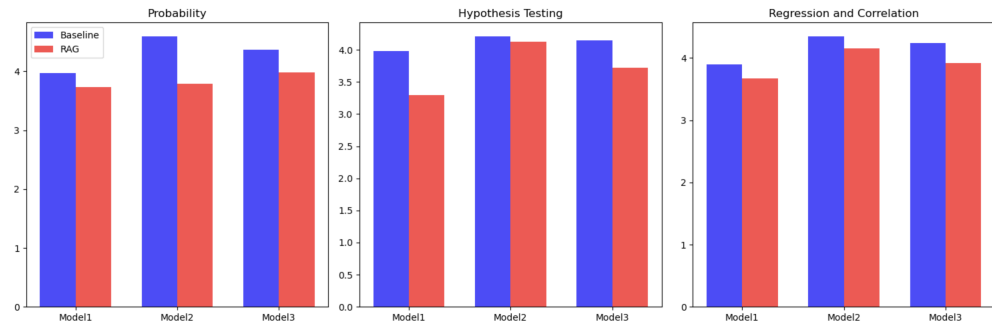


Figure 5.3: Performance by Statistical Domain

- **Understanding Dimension Most Impacted:** The conceptual understanding dimension experienced the largest negative effects. This suggests that external information particularly disrupts deep conceptual reasoning. This finding shows that abstract thinking and the integration of knowledge are most at risk for interference. The significant impact on understanding reveals basic limitations in current knowledge integration processes.

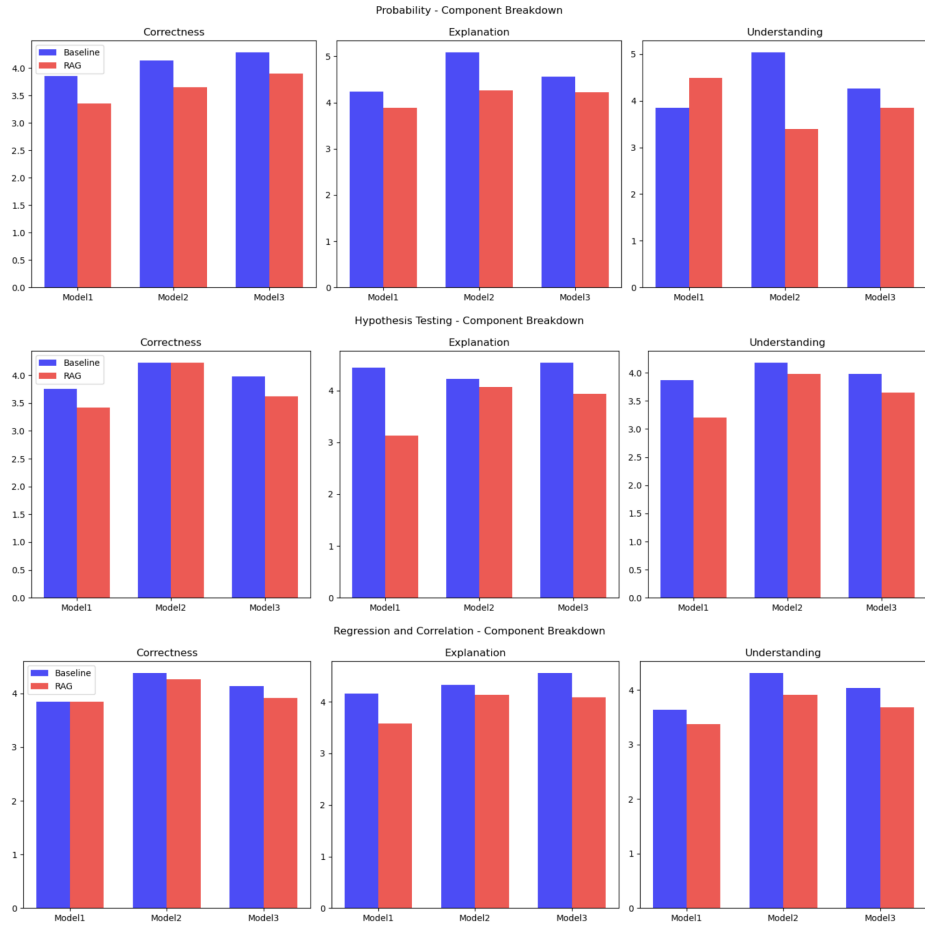


Figure 5.4: Component-Level Analysis by Statistical Domain

- **No Architecture-Based Differences in Learning Capacity:** Even with different training methods, models showed similar vulnerability to RAG interference. This uniformity implies that current training methods do not effectively tackle knowledge integration challenges. The similar response patterns indicate that we may need solutions focused on architecture rather than training.

Chapter 6

Discussion

The experimental results demonstrate a complex picture that challenges the fundamental assumptions about how LLMs interact with external knowledge. This chapter interprets the unexpected findings, discusses their implications, and addresses the study’s limitations.

6.1 Interpreting the Negative Learning Effects

Instead of observing the hypothesized learning enhancement, all the three models consistently demonstrated reduced performance when given external statistical resources through RAG. This section explores the reasons behind this surprising but strong finding.

Information Interference Phenomenon:

The consistent reduction in performance of LLMs reveals basic limitations in how current LLM architectures handle multiple sources of information during complex reasoning [Lewis et al., 2020]. Two main reasons explain this interference effect.

- **Cognitive Conflict Between Knowledge Sources:** When the models are provided with external information that offers alternative or different methods or explanations for a single question, they experience systematic confusion instead of beneficial supplementation [Sweller et al., 2019]. For instance, a model’s approach to a hypothesis testing question might conflict with different sequencing or notations retrieved in textbooks. This conflict is most evident in the Understanding dimension, which showed the largest negative impacts across all models.
- **Attention Mechanism Overload:** The transformers have a limited attention capacity, and this limited attention capacity becomes an issue when it is handling long external contexts along with complex reasoning tasks [Vaswani et al., 2017]. The RAG system provided significant extra context, which overwhelms the transformer. This affected the attention mechanism’s ability to focus on key parts of the problem.

Model-Specific Response Patterns

Model Qwen2-Math demonstrated the maximum negative effects, this contradicted the expectations that mathematical specialisation would enhance knowledge integration capabilities. The mathematics specialised architecture developed created highly optimised internal pathways through training on synthetic math data Yang et al. [2024]. When external information provides alternative approaches, it disrupts these well-tuned reasoning processes instead of improving them [Hendrycks et al., 2021].

6.2 Implications for AI Learning and Development

6.2.1 Practical Applications and Implementation Challenges

The findings from this investigation have immediate implications for AI systems used in education and professional settings. Current uses of RAG enhanced AI tutoring systems counterproductive for complex reasoning tasks especially in STEM fields where precise methods are important. Educational institutions should rethink strategies that just give AI systems access to all textbooks. Instead, they should focus on selecting clear, conflict free knowledge sources that match the system’s reasoning skills [Garfield and Ben-Zvi, 2008]. The ongoing negative effects seen in statistical areas indicate that RAG systems need careful selection of knowledge instead of just providing a lot of information. For businesses, organisations using AI analytics tools should understand that having a complete knowledge base does not necessarily lead to better analytical skills. This challenges the belief that external information always improves AI performance [Lewis et al., 2020].

6.3 Study Limitations and Methodological Considerations

Several limitations might have affected the generalisability of these findings and may require careful consideration. The experimental design focused solely on statistical reasoning. Additionally, the interference effects observed may not apply uniformly across all knowledge areas. Mathematical and statistical reasoning depends on precise logical sequencing, which may be more affected by outside information than more flexible areas like creative writing or general knowledge tasks. Relying on evaluations from large language models might overlooked subtle reasoning mistakes that human experts would catch [Kocmi and Federmann, 2023][Zheng et al., 2023]. Lastly, the time constraint of using post-training textbooks may have unintentionally chosen materials that do not align with the models’ learned representations [Carlini et al., 2021].

6.4 Future Research Directions

The unexpected findings open several important avenues for future research. Examining interference effects in different reasoning areas would help in determining if the observed patterns

display specific weaknesses in certain domains or general limits in how knowledge is integrated. The consistent negative effects challenge basic beliefs about how LLMs process external information [Brown et al., 2020]. This suggests that changes in design focused on knowledge integration may be needed. Future research should look into whether these interference effects continue to exist across various model sizes and designs, as new capabilities in larger models might change how knowledge integration works [Wei et al., 2022].

Chapter 7

Conclusion

This dissertation investigated whether LLMs can genuinely learn statistics by using external knowledge through RAG. The extensive experimental framework yielded unexpected yet strong findings that challenged common beliefs about AI learning abilities.

The main contribution of this study is documenting the consistent negative learning effects in LLM models when provided with relevant external information through optimised RAG systems. This finding shifts our understanding of how AI interacts with knowledge, demonstrating that information access and improved reasoning work against each other in current models. The research revealed that domain-specialised models displayed greater susceptibility to interference from external information than general-purpose architectures. Qwen2-Math model displayed the maximum negative effects across evaluation dimensions. The investigation also demonstrated that external information interference affects different reasoning components systematically, with conceptual understanding showing the most severe impacts, while computational accuracy shows more moderate effects.

From a methodological perspective, this research created a systematic RAG optimisation framework based on semantic evaluation metrics. This framework provides objective measures of retrieval quality independent of downstream performance. The three-dimensional evaluation approach examines correctness, explanation quality, and understanding, providing a better alternative to accuracy-focused methods. This enables the identification of differential effects masked by simple accuracy measurements. The adaptive statistical analysis protocol ensures statistical validity and accommodates the non-normal distributions common in LLM evaluation data. It established strong methods for assessing AI performance.

Can LLMs Learn Statistics? The answer is complex based on the experimental evidence. Current LLMs can process external statistical information, but this processing results in performance decrements rather than learning improvements. This suggests that present architectures lack effective mechanisms for integrating external and parametric knowledge in complex reasoning tasks. The findings show that information processing is not the same as learning in terms of better reasoning skills. Current LLM architectures have basic limitations when it comes to managing multiple information sources for complex reasoning.

Bibliography

- Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn and Bacon, 2001.
- Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/arpit17a.html>.
- Benjamin S Bloom. *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green, 1956.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media, 2nd edition, 2020.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650, 2021.
- George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2nd edition, 2002.
- Beth L. Chance. Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10, 2002. URL <https://api.semanticscholar.org/CorpusID:60958802>.
- Samprit Chatterjee and Ali S. Hadi. *Regression Analysis by Example*. John Wiley & Sons, 5th edition, 2012.

- Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- R. V. Dukkipati. *Numerical Methods Fundamentals - 6. Curve Fitting, Regression, and Correlation*. Mercury Learning and Information, 2023. URL <https://app.knovel.com/hotlink/pdf/id:kt0138WHHD/numerical-methods-fundamentals-6-curve-fitting-regression-and-correlation>
- M Fang, X Wan, F Lu, F Xing, and K Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *Sci Data*, 12(1):1392, Aug 2025. doi: 10.1038/s41597-025-05283-3.
- Jared Fauver, Phyllis Okwanu, Rachel Belmessieri, Stephan Patterson, and Stephen Andrus. *Introductory Statistics*. LOUIS: The Louisiana Library Network, January 2024. Licensed under Creative Commons Attribution-ShareAlike 4.0 International License.
- Vitaly Feldman and Chiyan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.
- R. A. Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52 (2):399–433, 1919. doi: <https://doi.org/10.1017/s0080456800012163>. URL <https://www.cambridge.org/core/journals/earth-and-environmental-science>
- David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W. W. Norton & Company, 4th edition, 2007.
- Joan Garfield and Dani Ben-Zvi. *Developing students’ statistical reasoning: Connecting research and teaching practice*. Springer, 2008.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR, 2020.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020. URL <https://api.semanticscholar.org/CorpusID:221516475>.

- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021. URL <https://api.semanticscholar.org/CorpusID:232134851>.
- Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of ACL*, pages 328–339, 2018.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfay, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL <https://aclanthology.org/2021.eacl-main.74/>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781, 2020.
- Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Lomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL <https://aclanthology.org/2023.eamt-1.19/>.
- William H. Kruskal and Wilson Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621, 1952. URL <https://api.semanticscholar.org/CorpusID:51902974>.
- Howard Levene. Robust tests for equality of variances. 1961. URL <https://api.semanticscholar.org/CorpusID:117424234>.
- Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020. URL <https://api.semanticscholar.org/CorpusID:218869575>.

- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Annual Meeting of the Association for Computational Linguistics*, 2017. URL <https://api.semanticscholar.org/CorpusID:12777818>.
- Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.
- Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 5th edition, 2013.
- Mark Schervish Morris DeGroot. *Probability and Statistics*. Pearson Education Limited, 5th edition, 2014.
- Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *ArXiv*, abs/1901.04085, 2019. URL <https://api.semanticscholar.org/CorpusID:58004692>.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018. URL <https://api.semanticscholar.org/CorpusID:3626819>.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL <https://aclanthology.org/2021.naacl-main.200/>.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.
- Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.
- Sheldon Ross. *A First Course in Probability*. Pearson, 10th edition, 2019.
- Alan H. Schoenfeld. *Mathematical Problem Solving*. Academic Press, 1985.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingyang Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

- S S Shapiro and M B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–591, Dec 1965. doi: <https://doi.org/10.2307/2333709>. URL <https://www.jstor.org/stable/2333709>.
- Mark D. Shermis and Jill Burstein. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, 2013.
- Weijia Shi, Sewon Min, Michihiko Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of EMNLP*, pages 2883–2895, 2023.
- John Sweller, Jeroen J. G. Van Merriënboer, and Fred Paas. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31:261–292, 06 2019. doi: 10.1007/s10648-019-09465-5.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- John W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5 2: 99–114, 1949. URL <https://api.semanticscholar.org/CorpusID:806596>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.
- Frank. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:196–202, 1945. URL <https://api.semanticscholar.org/CorpusID:53662922>.
- Chris J. Wild and Maxine Pfannkuch. Statistical thinking in empirical enquiry. *International Statistical Review*, 67(3):223–248, 1999.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64, 11 2016. doi: 10.1145/3446776.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Appendix

Appendix A: Core Implementation Code

A.1 Answer Generation Functions

The following functions implement the baseline and RAG answer generation systems used in this study.

Baseline Answer Generation:

This code generates baseline answers using only the LLM's parametric knowledge. It takes the statistical question to be answered and an identifier model as input and returns the generated answer. The prompt used for generating the answer can be found in the code.

```
def llm_baseline_answer(query, model_name):  
  
    response = ollama_client.generate(  
  
        # Mentioning the model name to be used for  
        # generating the reponse  
        model=model_name,  
  
        # Defining the prompt  
        prompt=f"""You are a statistics expert.  
Answer the question step-by-step:  
Question: {query}""",  
  
        stream=False,  
  
        # Defining reponse length, temperature and  
        #Nucleus sampling parameter  
        options={  
            'num_predict': 1000,  
            'temperature': 0.2,
```

```

        'top_p': 0.9
    }
)
# Returning the response
return response['response']

```

RAG Answer Generation:

This code generates RAG answers with the use of retrieved information. It takes the statistical question to be answered, the top_k value and an identifier model as input and returns the generated answer. The prompt used for generating the answer can be found in the code.

```

def llm_rag_answer(query, model_name, top_k):

    # Retrieving relevant chunks
    relevant_chunks = retrieve_relevant_chunks(query, top_k)

    # Reranking chunks using cross encoder for improved relevance
    reranked_chunks = rerank_chunks(query, relevant_chunks)

    # Assembling context within token limits
    context = assemble_context(reranked_chunks, max_chars=800)

    # Generating answer with explicit context grounding
    prompt = f"""You are a statistics expert.
Your task is to provide a correct and well-explained
answer to the question below. You MUST base your entire
answer on the provided Reference Material. Do not use any of
your pre-trained knowledge. Synthesize the information
from the context to construct your response.

### Reference Material:
{context}

### Question:
{query}

### Answer:
Based exclusively on the reference material,
the solution is as follows:
"""

```

```

response = ollama_client.generate(

    # Mentioning the model name to be used for
    # generating the reponse
    model=model_name,

    # Defining the prompt
    prompt=prompt,
    stream=False,

    # Defining reponse length, temperature and
    #Nucleus sampling parameter
    options={
        'num_predict': 1000,
        'temperature': 0.2,
        'top_p': 0.9
    }
)
return response['response']

```

Supporting Retrieval Functions:

This function retrieves topk most semantically similar chunks from the knowledge base. It takes the query and top_k value as the inputs and returns the relevant chunks.

```

def retrieve_relevant_chunks(query, top_k):

    query_embedding = embedder.encode([query])[0]

    results = collection.query(
        query_embeddings=[query_embedding.tolist()],
        n_results=min(top_k * 2, 15),
        include=["documents", "distances"]
    )
    documents = results.get("documents", [[]])[0]
    return documents[:top_k]

```

This piece of code is used for reranking the chunks using a cross-encoder for improved relevance ordering. It takes the question to be answered as well as the chunks retrieved and returns the reranked chunks.

```
def rerank_chunks(query, chunks):

    pairs = [[query, chunk] for chunk in chunks]
    scores = reranker.predict(pairs)
    scored_chunks = sorted(list(zip(scores, chunks)),
                           key=lambda x: x[0], reverse=True)
    return [chunk for score, chunk in scored_chunks]
```

LLM-as-a-Judge Evaluation

Evaluation Function:

This is the piece of code used for evaluating all the responses from the models. It takes the following inputs:

- question (str): Original statistical question
- llm_answer (str): Answer generated by the test LLM
- original_answer (str): Human expert answer for reference
- judge_llm (str): Identifier for the judge LLM model

And returns a JSON formatted evaluated score, which is broken down into score using a separate function `score_break`.

```
def evaluate_answer(question, llm_answer, original_answer, judge_llm):

    # Defining the prompt
    prompt = f"""You are a statistics evaluator. Score the answer on
    the basis of its correctness, explainability and understanding. Be
    precise and score it between 0 and 5, also allowing for half-point
    increments (e.g., 2.5, 3.5).

    The rubrics is provided below, be very precise
    about your evaluations.

    RUBRIC:
    1. **CORRECTNESS**:
        5: Correct answer AND correct method
        4: Correct method, minor computational error
```

```

3: Partially correct approach
2: Wrong approach but shows some knowledge
1: Completely wrong or no attempt

2. **EXPLANATION**:
5: Clear, complete, step-by-step explanation
4: Generally clear with minor gaps
3: Adequate explanation with some confusion
2: Unclear but attempted explanation
1: No explanation or incomprehensible

3. **UNDERSTANDING**:
5: Shows deep understanding (discusses assumptions,
limitations, context)
4: Shows good understanding with minor gaps
3: Shows basic understanding
2: Shows minimal understanding
1: No conceptual understanding evident

### QUESTION:
{question}

### ORIGINAL ANSWER:
{original_answer}

### LLM-GENERATED ANSWER:
{llm_answer}

IMPORTANT: Reply with ONLY this JSON format, nothing else:
{"correctness": X, "explanation": X, "understanding": X}}

JSON OUTPUT: ""

response = ollama_client.generate(
    model=judge_llm,

    # Assigning the prompt
    prompt=prompt,
    stream=False,

```



```

    # Defining reponse length, temperature and
    # stop keywords
    options={
        "temperature": 0,
        "num_predict": 50,
        'stop': ['\\n\\n', 'Human:']
    }
)
return response['response']

```

Appendix B: Sample Dataset

Here is 1 question from the dataset generated from iteration where parameters were temperature of 0 and no top p value.

Question1

Explain what is wrong with the following statements.

Use complete sentences.

1. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
2. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Original Answer:

1. You cant calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
2. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

LLaMA Model Baeline Answer:

I'd be happy to help you identify what's wrong with each statement.

****Statement 1:****

The statement is incorrect because it incorrectly adds the probabilities of rain on Saturday and Sunday to get the overall probability of rain over the weekend. The correct way to calculate the probability of rain over the weekend would be to multiply the individual probabilities, not add them. This is because the events are independent (the occurrence of rain on one day does not affect the other), so we can use the multiplication rule for independent events.

The correct calculation would be:

Probability of rain on Saturday \times Probability of rain on Sunday
 $= 0.6 \times 0.7 = 0.42$

This means there is a 42% chance of rain over the weekend, not 130%.

****Statement 2:****

The statement is incorrect because it implies that hitting a home run is more likely than getting any successful hit (e.g., a single, double, triple). However, in baseball, a home run is a specific type of successful hit. In fact, a home run is the most successful outcome for a batter.

A more accurate statement would be: "The probability that a baseball player hits a home run is equal to or greater than the probability that he gets any other successful hit." This acknowledges that hitting a home run is a subset of successful hits and may not necessarily be less likely.

LLaMA Model RAG Answer:

Based on the provided Reference Material, I can address each of the questions and statements.

1. True or false, and explain:

(a) If something has probability 1,000%, it is sure to happen.

False. The statement "probability 1,000%" is not a valid probability value. Probability values must be between 0 and 1, inclusive. A probability of 100% means that the event will definitely occur, but a probability of 1,000% implies an impossible outcome, which is not possible in probability theory.

(b) If something has probability 90%, it can be expected to happen about nine times as often as its opposite.

False. The statement "about nine times as often" is incorrect. Probability values represent the likelihood or chance of an event occurring, not the frequency of occurrence. A probability of 90% means that the event will occur approximately 9 out of 10 times, but it does not imply a specific frequency.

2. Say whether each of the following statements is true or false, and explain why:

(a) About 95% of the measurements are in this range.

True. The statement "about 95%" implies that there is a 5% chance that the measurement will fall outside of the specified range. This is a valid interpretation of a probability value.

(b) There is about a 95% chance that the next measurement will be in this range

False. The phrase "about a 95% chance" is ambiguous and could imply that there is a 5% chance that the opposite event occurs, which is not necessarily true. A more accurate statement would be "There is a 95% probability that the next measurement will fall within this range."

3. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.

False. The statement implies an impossible outcome (a 130% chance), which is not valid in probability theory. To calculate the overall probability of rain over the weekend, we would need to multiply the individual probabilities: $0.6 \times 0.7 = 0.42$, or approximately 42%. This result does not imply a higher probability due to the combination of events.

4. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

False. Without additional context or information about the specific events (e.g., what constitutes a "successful hit"), it's impossible to determine which event has a higher probability. Both probabilities could be equal, and both are uncertain without further data.

DeepSeek Model Baseline Answer:

<think>

First, I need to explain what's wrong with each statement using complete sentences.

Starting with statement 1: "If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend."

This seems incorrect because probabilities can't exceed 100%. Adding two percentages gives 130%, but that's not how probability works for combined events. The event "rain over the weekend" isn't simply the sum of the individual days' chances, especially if we're talking about independent events.

If rain on Saturday and Sunday are independent events, then the probability of rain on at least one day can be calculated using the formula:

$$P(\text{rain on Sat or Sun}) = P(\text{Sat}) + P(\text{Sun}) - P(\text{both})$$

But even without that, adding probabilities directly isn't correct because it assumes that both events must happen for the combined event to occur, which is not necessarily true. The statement says "chance of rain over the weekend," which probably means at least one day has rain.

Moreover, 130% doesn't make sense as a probability since probability ranges from 0% to 100%. So, it's invalid because probabilities can't be greater than 100%.

But let's think deeper. If we add the probabilities without considering independence or overlap, we might get an overestimate if there's any chance of both days having rain.

The correct way to find $P(\text{rain on Sat or Sun})$ is:

$$P(\text{Sat or Sun}) = P(\text{Sat}) + P(\text{Sun}) - P(\text{Sat and Sun})$$

$$\begin{aligned} \text{If they are independent, } P(\text{Sat and Sun}) &= P(\text{Sat}) * P(\text{Sun}) \\ &= 0.6 * 0.7 = 0.42 \end{aligned}$$

$$\begin{aligned} \text{So } P(\text{Sat or Sun}) &= 0.6 + 0.7 - 0.42 = 1.3 - 0.42 = 0.88, \\ &\text{which is 88\%.} \end{aligned}$$

But the statement claims a 130% chance, which is impossible.

Even if we don't assume independence, adding probabilities can lead to values greater than 100%, but that doesn't mean it's correct; in fact, for mutually exclusive events or something else, it might be wrong.

The key point here is that the sum of probabilities exceeds 100%, which violates the basic property that a probability cannot be more than 1. So, there must be some overlap or dependency not accounted for.

But regardless, saying "there is a 130% chance" is mathematically incorrect because probability can't exceed 1.

So, I think the main issue is that probabilities are additive only under certain conditions (like mutually exclusive events), and here adding them gives more than 100%, which isn't possible. Also, it's not clear what "over the weekend" means— "does it mean rain on both days or at least one? But even if we interpret it as at least one day, the calculation shows it should be less.

The statement doesn't specify independence; it just says there is a 60% chance and 70% chance separately. So, without knowing the correlation, adding them isn't justified.

But fundamentally, any probability cannot exceed 100%, so claiming 130% is wrong.

Now, for statement 2: "The probability that a baseball player hits a home run is

greater than the probability that he gets a successful hit."

This seems problematic because in baseball, hitting a home run requires getting a successful hit first. A home run is a type of successful hit; it's one specific outcome when you hit the ball.

So, let $P(\text{HR})$ be the probability of hitting a home run.

$P(\text{successful hit})$ would typically mean the probability of having a hit (i.e., not an out), which includes singles, doubles, triples, and home runs.

Therefore, HR is a subset of successful hits. So, by definition, $P(\text{HR})$ should be less than or equal to $P(\text{successful hit})$.

In fact, it's almost always less unless the player only ever hits home runs, but that's unrealistic.

So, this statement implies that $P(\text{HR}) > P(\text{hit})$, which is incorrect because HR requires a hit and then an additional condition (the ball goes out of bounds or something). But more precisely, in probability terms:

Let E be the event of getting a successful hit. Then, hitting a home run is a specific outcome within E.

So, $P(\text{HR}) \leq P(E)$, since HR implies E.

Therefore, it cannot be that $P(\text{HR}) > P(\text{hit})$; this would violate the basic property of conditional probability or inclusion relationships.

Moreover, in reality

DeepSeek Model RAG Answer:

<think>

First, I need to base my answer entirely on the provided Reference Material. Let me read it carefully.

The Reference Material has two parts: Example 2a and then some text about probability statements. But for this question, I'm supposed to explain what's wrong with specific statements using only that reference material.

Now, looking at the Question:

1. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.

2. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

I need to explain what's wrong with these statements using only the Reference Material.

The Reference Material mentions something about probabilities and

percentages, but it seems incomplete for this task. Let me read it again.

"Example 2a

Suppose that whether it rains tomorrow depends on previous weather conditions only through whether it is raining today. Suppose further that if it is raining today, then it will rain tomorrow with probability

1. True or false, and explain:

(a) If something has probability 1,000%, it is sure to happen.

(b) If something has probability 90%, it can be expected to happen about nine times as often as its opposite."

Then there's a dot and "Say whether each of the following statements is true or false..." but that seems cut off. However, for this question, I need to focus on explaining why the given statements are incorrect.

The key point from the Reference Material is in part (a): it says if something has probability 100%, it's sure to happen, and here we have a statement about $60\% + 70\% = 130\%$ for rain over the weekend. That directly relates to adding probabilities incorrectly.

Also, from part (b), it talks about expected frequency with percentages, but that might not be directly relevant.

The Reference Material doesn't explicitly state what probability means or how to combine them; I have to infer from the context.

In Example 2a, it's about conditional probability and dependence between days. But for this question, statement 1 is about adding probabilities without considering dependence or other factors.

Statement 1 says: If there's a 60% chance on Saturday and 70% on Sunday, then 130% chance over the weekend. This is wrong because probabilities can't be added that way if they are not mutually exclusive events in time.

The Reference Material doesn't directly address this, but it does mention probability greater than 100%, which might imply something about adding probabilities exceeding certainty.

In part (a), it says "if something has probability 1,000%" â€\ wait, that's a typo or error. It probably meant 100% since probability can't be more than 100%. But in the statement, it's given as 130%, which is clearly impossible.

Probability cannot exceed 100% because it represents certainty on a scale from 0 to 1 (or 0% to 100%). So adding two probabilities can't result in more than 100% if they are for different events, but here the weekend might be considered as one event or combined.

The statement is about rain over the weekend being certain with probability greater than 100%, which violates basic probability rules.

Now, Statement 2: The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

This seems to compare two probabilities. Home runs are rarer than successful hits in general, but I need to base this on Reference Material only.

The Reference Material doesn't mention anything about baseball or specific events; it's more general with examples like weather.

But perhaps from part (b), which talks about measurements and probability, we can infer that probabilities should be between 0 and 1 inclusive.

In Statement 2, if "successful hit" means a base hit, then home runs are a subset of hits, so the probability of hitting a home run is less than or equal to the probability of getting any successful hit (assuming all hits include singles, doubles, etc., but typically in baseball stats, $P(\text{home run}) < P(\text{at least one hit})$, I think.

But again, no specific data from Reference Material. However, the general principle might be that probabilities can't exceed 100%, and for dependent events or something.

I need to use only what's given. Let me extract relevant points.

From part (a): It says "if something has probability 1,000%" â€\ this is probably a mistake; it should be 100%. But the explanation might imply that probabilities are bounded by 100%.

Qwen-2 Math Baseline Answer:

: Let's analyze each statement step-by-step to identify what is wrong

with them.

Statement 1:

If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.

Explanation:

The statement is incorrect because it violates the basic properties of probability. Probability values range from 0 to 1 (or 0% to 100%). A 130% chance of rain is not possible because it exceeds the maximum probability of 100%.

To find the correct probability of rain over the weekend, we need to use the formula for the union of two events. Let $P(A)$ be the probability of rain on Saturday and $P(B)$ be the probability of rain on Sunday.

The probability of rain on at least one day (over the weekend) is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B)$ is the probability of rain on both days.

Since we don't have information about the dependence between the two events (i.e., whether it can rain on both days), we cannot calculate $P(A \cap B)$. However, if we assume that the events are independent (which is a common assumption in such problems unless stated otherwise), then:

$$P(A \cap B) = P(A) \times P(B) = 0.60 \times 0.70 = 0.42$$

Substituting these values into the formula, we get:

$$P(A \cup B) = 0.60 + 0.70 - 0.42 = 0.88$$

So, the probability of rain over the weekend is 88%, not 130%.

Statement 2:

The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

Explanation:

This statement is incorrect because "successful hit" is a more general term that includes both home runs and other successful outcomes (such as singles, doubles, or triples). Therefore, the probability of a successful hit must be at least as large as the probability of a home run.

To illustrate, let's define:

- $P(H)$: Probability that a player hits a home run.
- $P(S)$: Probability that a player gets a successful hit

(which includes home runs and other successful outcomes).

By definition, S is a superset of H , so:

$$P(S) \geq P(H)$$

This means the probability of a successful hit cannot be less than the probability of a home run. Therefore, it is impossible for the probability of a home run to be greater than the probability of a successful hit.

In conclusion, both statements are incorrect due to fundamental misinterpretations of probability concepts.

Qwen-2 Math Model RAG Answer

The provided reference material does not contain any information about weather conditions or probabilities of rain over a weekend. Therefore, I cannot use the reference material to explain what is wrong with the given statements.

However, I can still provide an explanation for each statement:

1. The statement "If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend" is incorrect because probabilities cannot exceed 100%. The probability of rain over the weekend is the probability that it rains on either Saturday or Sunday or both. This probability can be calculated using the formula for the union of two events:
 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where A and B are the events of rain on Saturday and Sunday, respectively. Since we do not have information about the probability of rain on both days, we cannot calculate this exactly, but it is clear that the probability cannot be 130%.

2. The statement "The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit" is incorrect because a home run is a type of successful hit. Therefore, the probability of hitting a home run is a subset of the probability of getting a successful hit. In other words, if an event is a home run, it is also a successful hit, but not all successful hits are home runs. Thus, the probability of hitting a home run cannot be greater than the probability of getting a successful hit.