National College of Ireland

Understanding the importance of enhanced logical reasoning among large language models with the help of hybrid symbolic architecture

MSc Research Project
Masters in Artificial Intelligence (MSCAIB)

Surya Prakash Chinnameda
Student ID: x23343699

School of Computing
National College of Ireland

Supervisor: Lavish Thomas

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Surya prakash Chinnameda
Student ID:	X23343699
Programme:	Masters in Artificial Intelligence (MSCAIB) **Year:** 2025
Module:	Practicum
Lecturer:	Lavish Thomas
Submission Due Date:	15-09-2025
Project Title:	Understanding the importance of enhanced logical reasoning among large language models with the help of hybrid symbolic architecture
Word Count:	**6805** **Page Count: 22**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Surya Prakash Chinnameda

Date: 15-09-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Understanding the importance of enhanced logical reasoning among large language models with the help of hybrid symbolic architectures

X23343699

Abstract

The hybrid symbolic neural architectures would provide an approach to overcome the small step reasoning and explainability of large language models (LLMs). This study hence attempts to solve the problem of automatic classification of logical reasoning questions into five categories: Combinatorics, Probability & Statistics, Logic Puzzles, Math Word Problems, and General Reasoning for efforts toward AI models that understand reasoning. (Liang et al., 2025; Hóu, 2025) A hybrid approach involving a transparent rule-based keyword categorizer combined with machine learning (TF-IDF + Logistic Regression) and deep learning (DistilBERT) was used. The rule-based part makes use of curated keyword and phrase lists per category such that bootstrapping labels from unstructured question text becomes very fast.

Category distribution, class imbalance, and text length statistics were analysed. To reflect more information on the longer end of the questions, the maximum token length was increased, and since training is easier to balance than validation or testing, only training was balanced using a WeightedRandomSampler. Experiment results prove that the hybrid (Asimit et al., 2022), XGBoost (Shari et al., 2021)DistilBERT plus rules model is better than the baseline in Macro-F1 score, especially for the minority classes of Logic Puzzles and Probability & Statistics. Metrics such as Accuracy, Precision, Recall, and per-class F1 show that this approach can attain high accuracy while keeping interpretability due to its symbolic component.

This paper demonstrates the added value that symbolic rules and neural models can bring to the reasoning question classification in terms of accuracy, as well as a scalable process that can be reproduced for the same NLP categorization task (Rudin et al., 2022; Song et al., 2025).

1. Introduction

1.1 Research question(s) and proposed solution

- How can hybrid symbolic architectures improve logical reasoning in large language models?
- What are the measurable impacts of such integration on reasoning accuracy and interpretability?
- How does the inclusion of domain-specific symbolic knowledge influence the performance of LLMs in real-world applications?

A hybrid symbolic–neural architecture is hereby proposed as an approach to combine the learnability of deep neural models with the structured inference capabilities of symbolic rule-based systems. In such a system, the symbolic part encodes domain- specific rules and keyword patterns for various categories of logical reasoning problems while the neural component (Distil BERT) learns fine-grained distinctions from labelled examples. As such, this combination has the following aim: Raised explainability through explicit matching of rules

alongside predictions of the model. Better generalization to structures of problems not seen before via symbolic knowledge transfer. (Rudin et al., 2022; Zhou et al., 2021).

1.2 motivation

Most advanced LLMs output highly accurate responses to open-ended text generation but their performance regarding logical deduction has always been relatively weak. In real-world applications, e.g., in educational tutoring systems, cognitive assessment tools, and competitive exam preparation platforms across a battery of tasks that test reasoning types, it is imperative that models be not just answer plausibility generators but answer type classifiers as well. This makes possible robust hybridization between pure deep learning and symbolic reasoning rule sets that would support transparent error analysis—otherwise impossible under purely neural regimes. (Evans, 2011; Wason & Evans, 1974).

This study proposes a practical and interpretable reasoning task classification pipeline that embraces rule-based categorization plus DistilBERT fine-tuning as an accurate approach to overcome the very insufficient accuracy of LLMs within structured contexts of reasoning.

2. Related Work

2.1 Introduction

This chapter critically analyses existing research on strengthening logical reasoning in large language models (LLMs) via hybrid symbolic architectures. It also discusses architectural designs, metrics for assessing reasoning accuracy and interpretability, symbolic integration in domain-specific scenarios, and challenges that come with them. The chapter further provides theoretical foundations, research gaps, and a conceptual framework to understand the enhancement of LLMs in logical reasoning abilities.

2.2 Critical Analysis of Previous Literature

2.2.1 Machine Learning Hybrid Symbolic Architectures for Logical Reasoning in LLMs

The comparative analysis examines hybrid symbolic architectures enhancing logical reasoning in LLMs. Liang, Wang, and Tong (2025) examine integrating symbolic AI with neural models to enhance reasoning. Their method combines logical theorem-proving modules with deep learning architectures to evaluate symbolic constraints. They find improved inference accuracy but note scalability issues. An argument emphasises rigorous formal logic integration, while a counterargument questions rule-based rigidity. Their work suggests potential but demands dynamic acquisition of symbolic knowledge and logical rules. It underpins neural–symbolic progress but lacks evaluation on an exceptionally large dataset scale. Similarly, Li et al. (2024) propose a neuro-symbolic framework for hypothetical deduction within LLMs to improve logical inference. Their method integrates symbolic reasoning chains (e.g., rule-based inference sequences) with prompt engineering and fine-tuning to generate deductive conclusions. They report a 24.34% improvement in accuracy on the FOLIO benchmark but caution over bias propagation from pre-trained weights. Limitations include computational overhead and dependence on handcrafted logic templates.

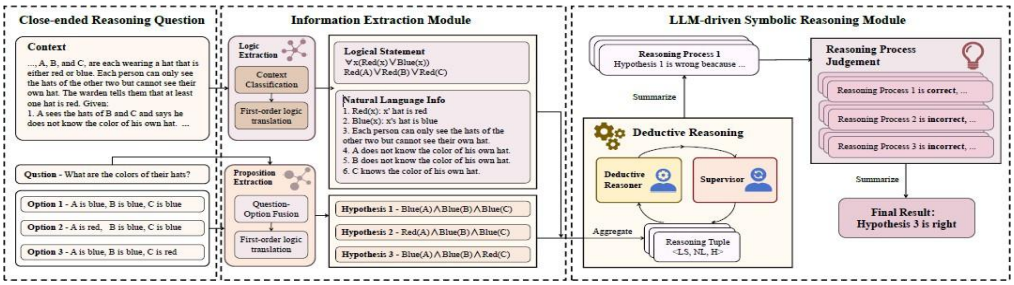

Figure 1: LLM-driven Neuro-Symbolic Approach for Faithful Logical Reasoning

(Source: Li et al., 2024)

In contrast, Xiong, and Zheng (2024) design TinyLlama, integrating deep neural embeddings with symbolic knowledge graphs to improve information retrieval accuracy. Their approach fuses vector similarity with logical rule application for query refinement. They demonstrate a 12.4 % reduction in retrieval errors but note a 15.6 % increase in latency. Limitations involve dependency on curated knowledge graphs and limited domain adaptability. This study balances neural flexibility and symbolic precision yet requires evaluation in dynamic real-world environments to validate generality. Hybrid symbolic architectures demonstrate potential to enhance logical reasoning in large language models by combining neural flexibility with formal logic. Despite these, there are obstacles with handling large data, the role of human-created rules or graphs, and the increased burden on computation.

2.2.2 Machine Learning Evaluation Metrics for Reasoning Accuracy and Interpretability

Monitoring the performance and transparency of ML models is both important for building reliable and effective ones. Zhou et al. (2021) review how to assess the accuracy, stability, and clarity of ML explanations. Their goal is to examine a wide range of metrics to find out how good machine learning explanations are in different models and situations. For each model, such as decision trees, random forests, and deep neural networks, they analyse the user interpretability scores, simplicity, and other metrics. The research uses both a literature review and a process to compare and study methods such as fidelity and stability on popular datasets across different systems. Limitations include omission of emerging approaches, limited empirical validation, and concerns about reproducibility, generalisability. The argument points out that the survey uses a wide range of approaches, but the counterargument suggests that it lacks original research.

Similarly, Alangari et al. (2023) investigate evaluation methods for interpretable ML by focusing on both user-centric assessment techniques and intrinsic model metrics. They survey user studies, model-intrinsic metrics, and post-hoc explanations. Findings highlight disparities in human-centred evaluation methods (e.g., inconsistent user study protocols and varying task designs) and a lack of consensus on metrics (e.g., disagreement between fidelity, stability, and interpretability measures). Methodologically, they integrate literature synthesis and comparative summaries. Limitations include potential bias towards popular interpretable ML frameworks and minimal empirical benchmarking of the proposed evaluation techniques.

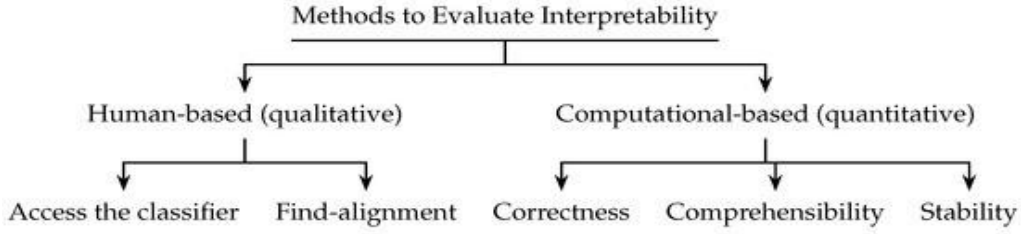

Figure 2: Methods to evaluate interpretability.

(Source: Alangari et al., 2023)

Rudin et al. (2022) advocate transparency, consistency, fairness, and define challenges: metrics, robustness, scaling, and trust. They aim to define foundational aspects and outline grand challenges, major research obstacles, and open problems in interpretable machine learning. They review theoretical frameworks, cases, and propose ten challenges balancing accuracy and transparency of interpretable machine learning models. Findings reveal gaps in algorithmic interpretability, user trust, fairness metrics, and robustness. Limitations include a broad scope that hinders in-depth analysis and a lack of empirical validation of the proposed challenges across real-world ML systems. Emphasising human-centric studies and foundational principles underscores the necessity for unified frameworks that balance accuracy with transparency. Future work should prioritise benchmarking protocols, context-specific metrics, and scalable methodologies to advance reliable ML interpretability in improving the logical reasoning of LLMs.

2.2.3 Machine Learning Domain-Specific Symbolic Knowledge Enhancing LLM Performance in Real-World Applications

ML-driven integration of domain-specific symbolic knowledge enhances LLM performance by ensuring logical consistency and robustness. Dinu (2024) proposes a neuro-symbolic framework combining a "domain-adversarial neural network (DANN)" with a symbolic reasoning layer that enforces logical consistency constraints of invariants. The method involves selecting "optimal regularisation parameters (λ values)" for adversarial loss and symbolic penalty terms. Evaluation occurs on three benchmarks, which are "Office-31 (image classification)", "Amazon Reviews (sentiment analysis)", and "PhysioNet ECG (time-series classification)". Results show that, compared to baseline DANN, the proposed model yields 8–12% higher accuracy when transferring to unseen domains. However, integrating large symbolic rule sets increases training time by up to 30%. Advocates claim parameter tuning prevents overfitting, whereas opponents caution that symbolic rigidity can reduce adaptability to highly heterogeneous data.

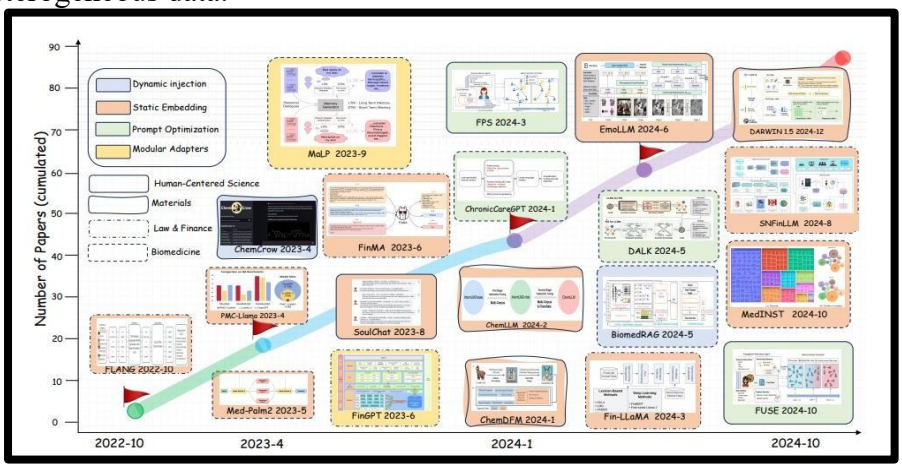

Figure 3: Illustration of Growth Trends in Domain-Specific Knowledge Injection into LLMs

(Source: Song et al., 2025)

In contrast, Song et al. (2025) survey methods for integrating domain-specific knowledge into LLMs, categorising dynamic injection, static embedding, and adapters. Their comprehensive review highlights advantages in specialised tasks such as healthcare and materials science but notes challenges in maintaining updated knowledge and balancing scalability with the specificity of breadth and depth. Zheng et al. (2025) survey retrieval-augmented generation, symbolic logic integration, and fact verification modules for multi-step reasoning. They categorise approaches into pipeline, embedding-based, and graph-based methods, evaluating "CodeXGLUE (software engineering)," "HotpotQA (multi-hop QA)," and "GSM8K (math)".

Findings show improved reasoning chains and consistency but increased inference latency. Challenges include adapting knowledge graphs across domains and ensuring real-time verification of Fact accuracy. Hybrid models embedding symbolic domain knowledge into neural architectures enhance reasoning accuracy and interpretability, reducing logical errors in real-world tasks. However, scalability issues emerge as large knowledge bases increase overhead.

2.2.4 Challenges and Limitations in Integrating Symbolic Reasoning with LLMs

The enhanced logical reasoning among large language models is quite significant with the help of hybrid symbolic architectures, but some challenges are still present. The previous studies have one common goal, which is to examine the rule-based reasoning that can be integrated with big language models (LLMs) to create systems that are accurate as well as interpretable. Hóu (2025) presented a general overview of "Neural-Symbolic Reasoning" and classified architectural patterns (such as "Neural Propose to Symbolic Verify" and "Symbolic Scaffold to Neural Fill") applied in program synthesis, robotics, and security applications. The study also stated that symbolic scalability breaks when proofs have more than a few hundred clauses (verification timeouts > 300 s), and training LLMs on formal loss functions is still impractical for an enormous collection.

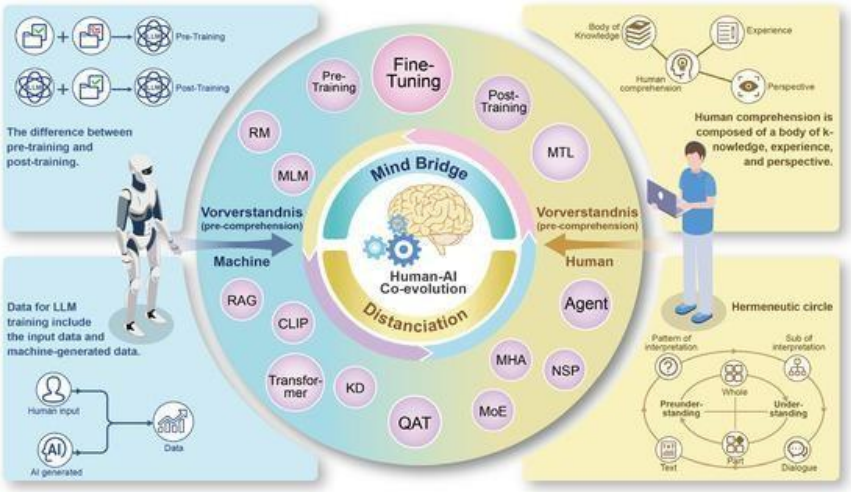

Figure 4: Fine-Tuned LLMs

(Source: Wu et al., 2025)

Figure 4 discusses "LLM Fine-Tuning" by situating fine-tuning as a hermeneutic process that mediates pre-understanding and recursive feedback loops, introducing "Tutorial Fine-Tuning" (TFT) approaches to adjust understanding under symbolic constraints. Wu et al. (2025) pointed out that fine-tuning under symbolic constraints usually needs hundreds of refinement iterations to reach 95% task accuracy and is thus resource-intensive on limited GPU clusters. Furthermore, Dehal, Sharma, and Rajabi (2025) critically summarised 77 LLM–Knowledge Graph (KG) integration studies, reporting on techniques such as prompt engineering, graph neural networks (GNNs), and evolutionary computation to obtain and organise symbolic knowledge in LLMs. The study also observed that real-time KG updates involve 30%–50% more computational overhead compared to pipelines based on text alone and that domain adaptation for general-purpose LLMs continues to provide only a 10% F1 reduction when applied to special ontologies (e.g., medical vs. financial). Overall, these insights concur that the future direction of study needs to address scalable symbolic representations, including differentiable logic embeddings or multi-agent systems that split up reasoning tasks, to overcome the present constraints of verification time and domain adaptation.

2.3 Theoretical Underpinning

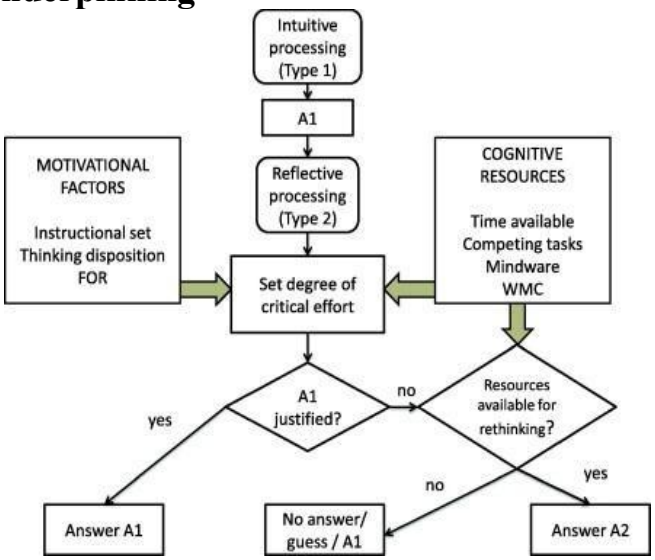

Figure 5: Framework of Dual-Process Theory of Reasoning

(Source: Evans, 2011)

Dual-Process Theory of Reasoning is quite suitable and has been selected for this research study. Dual-Process Theory of Reasoning holds that there are two cognitive systems, like fast, intuitive System 1 and slow, analytical System 2 (Wason and Evans, 1974). LLMs mostly mimic System 1, using statistical pattern matching to produce text. They do not have strong System 2 abilities for strict, rule-based reasoning, except for a few models (ChatGPT o4-mini, and o4-mini-high). Hybrid symbolic approaches implement explicit, formal reasoning modules that simulate System 2 capabilities, allowing LLMs to evaluate and correct their responses through logical inference. By coupling symbolic scaffolding with neural generation, these architectures promote interpretability and consistency, aligning computation with human dual-process reasoning and strengthening the ability of LLMs for accurate, robust context-sensitive logical inference and well-supported complex decision-making in real-world scenarios.

2.4 Literature Gap

Despite advancements in hybrid neuro-symbolic architectures, large-scale empirical validation on diverse real-world datasets remains lacking. Moreover, standardised metrics for assessing interpretability and updating of domain knowledge are underdeveloped. This research addresses these gaps through a comprehensive benchmark evaluation. It further proposes differentiable logic embeddings for real-time knowledge integration and interpretability metrics.

2.5 Conceptual Framework

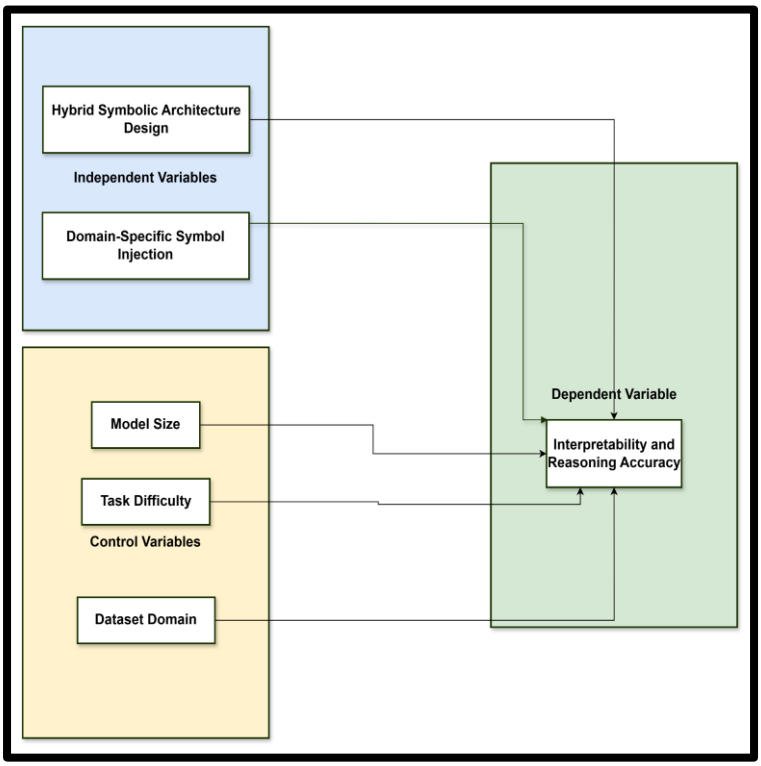

Figure 6: Conceptual Framework

2.6 Chapter Summary

This chapter critically examines the hybrid symbolic architectures for improving LLM logical reasoning, including architectural designs, evaluation measures, and integration of domain knowledge, as well as challenges. The chapter then introduces the Dual-Process Theory as a background, recognises gaps in the literature, and recommends a conceptual framework to inform future studies on how to enhance LLM reasoning abilities.

3. Research Method and Specification

3.1 Introduction

The current chapter presents the detailed methodological framework that is followed when studying Hybrid functionalities to boost logical capabilities within large language models. It explains the research design, sources of data, and procedures of model development, criteria used to evaluate the data, and deployment strategies used in the analysis.

3.2 Research Approach

This study employs a deductive research approach to investigate hybrid reasoning architectures. As defined by Okoli (2023), Deductive reasoning is a verification approach applying general theories to specific hybrid-symbolic reasoning cases. This is a theory that integrates the theoretical principles of hybrid integration with empirical experimentation by generating some hypotheses founded on the dual process theory of reasoning. Deductive analysis provides the ability to conduct systematic examinations of the capabilities of various models against the standards of logical analysis by applying the main principles of Validity, Soundness, and Consistency. It then promotes the possibility of verifying the findings based on the theoretical assumptions (Premises accurately reflect reality) that can easily be done through rigorous validation and replication.

Also, deductive reasoning facilitates clear model comparison and selective optimisation and is necessary to develop causality and optimise hybrid architectures in a well-organised hypothesis-led research design.

3.3 Research Philosophy

The methodology of this research follows the approach of positivist philosophy. According to Maretha (2023), Positivist philosophy refers to paying attention to empirical observation and statistical analysis. This philosophy is consistent with the fact that the objective measure of the Hybrid model performance is required in the study, and the accuracy, F1-score, and latency measures should be thoroughly observed and quantified. Employing the measurable outputs and rigid replicable processes, such as cross-validation and isolated runtime environment deployment, the positivist philosophy benefits hypothesis testing and causal inference between the dual-process theories of reasoning and empirical findings. As a result, the study can generate insights that can be generalised, have great validity, and similar designs can be compared clearly in the studies of both machine-learning and deep-learning architecture.

3.4 Data Collection

The Logical Reasoning Improvement Dataset has been collected from Kaggle to train and evaluate hybrid symbolic architectures for improving LLM logical reasoning. The "train.csv" file includes four main columns such as input, output, instruction, and data_source of natural language questions with logical inference required, correct solutions to the questions, further solving instructions, and provenance information, respectively (Kaggle, 2023). The dataset is derived from standardised logic tests and academic archives, such as diverse problem types in reasoning. In pre-processing Only these three fields_instr_columns were used here by models in project training. Lowercasing and removing punctuation normalized text so that token matches special characters replaced (e.g., $\chi^2 \rightarrow$ chi2). Missing values were filled with empty strings so as not to break the format. An 80% ,10%,10% random split by dataset was implemented for train, validation, and test splits, accordingly, maintaining class proportions across all splits. (GeeksforGeeks, 2025; kaggle, 2024)Training involves a Weighted Random Sampler in addition to class-weighted loss to deal with class imbalance across different batches. The prediction target is set from the category column so that the model's learning objective matches directly with evaluation measures such as classification accuracy, macro-F1 score, and per-class recall. The symbolic and neural components get textual input from a concatenation of the instruction and input fields; output and data source fields are dropped from features going into the model. (Imani et al., 2025).

3.5 Preprocessing & Model Selection

The textual questions of the Logical Reasoning Improvement Dataset were processed such that they would be readily accepted by both symbolic and neural classification models. (GeeksforGeeks, 2021) The fields are concatenated, lowercased, stripped of punctuation, normalised for special terms—for example, χ^2 replaced with chi2, and then tokenized for vectorization using TF-IDF (Nagahisarchoghaei et al., 2023). to transform into sparse feature matrixes appropriate for usage within standard supervision learning algorithms. (Mars, 2022). The category column is used as a target in classification, whereas all other columns except output and data_source are used as input features. Any missing value in the field< shall be replaced by an empty string to keep the format intact.

A DistilBERT-based Transformer model was fine-tuned as part of the deep learning pipeline, taking advantage of pre-trained contextual embeddings to better capture semantic relationships between tokens. Text sequences were tokenised to a maximum length of 320 tokens so as not

to truncate long reasoning problems. A WeightedRandomSampler was used in training the model together with class-weighted cross-entropy loss to start addressing the issue of class imbalance. (Nguyen et al., 2020).

- **Support Vector Machine (SVM):** Provides a large margin of separation and is very good at picking up slight linguistic as well as logical differences. (Asimit et al., 2022)
- **XGBoost Classifier:** Uses gradient-boosted decision trees for capturing nonlinear interactions among features in the presence of regularisation to check overfitting. (Shari et al., 2021).
- **DistilBERT Transformer Classifier:** Applies self-attention for capturing highly complicated contextual relationships relevant to reasoning tasks with computational efficiency. (Liu et al., 2025).

This symbolic–neural hybrid architecture enables symbolic rules to complement neural predictions, thereby enhancing both the accuracy and interpretability of reasoning across different types of problems. (Liang et al., 2025; Song et al., 2025)

3.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was done to understand the structure, quality, and distribution of the dataset before going into model training. The major objectives were to find out the missing values in the dataset, check class balance, and look at some statistical characteristics of textual data.

1. Missing Values Check:

The instruction, input, and category columns checked for missing values. Since the instruction and category did not have any values missing from figure 7, the input column had blank values for questions that did not have supplementary information. These were replaced with an empty string ("") to keep the formatting of the inputs consistent.

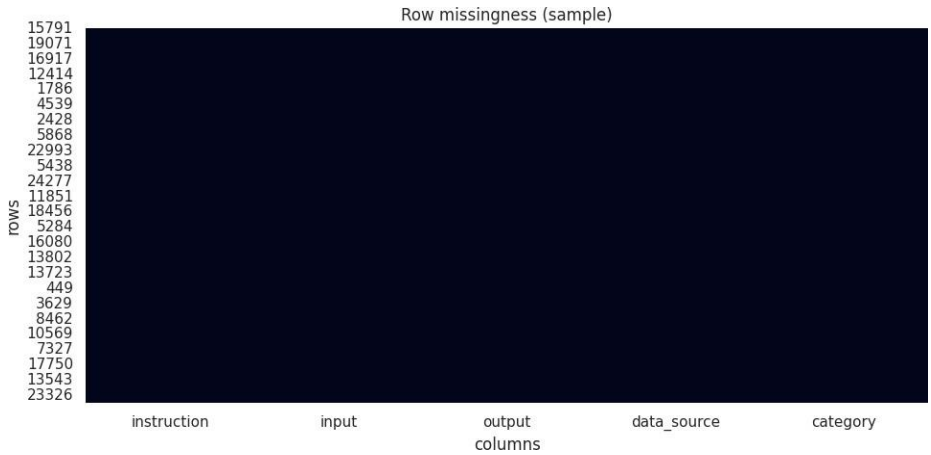

Figure 7: missing values

2. Category Distribution:

The distribution of Combinatorics, Probability & Statistics, Logic Puzzles, Math Word Problems, and General Reasoning categories was visualized with the help of bar charts (Figure 8). It came out that the dataset is imbalanced because some categories have many more examples than others. This gave a perfect justification for using oversampling as well as class-weighted loss in the later stages.

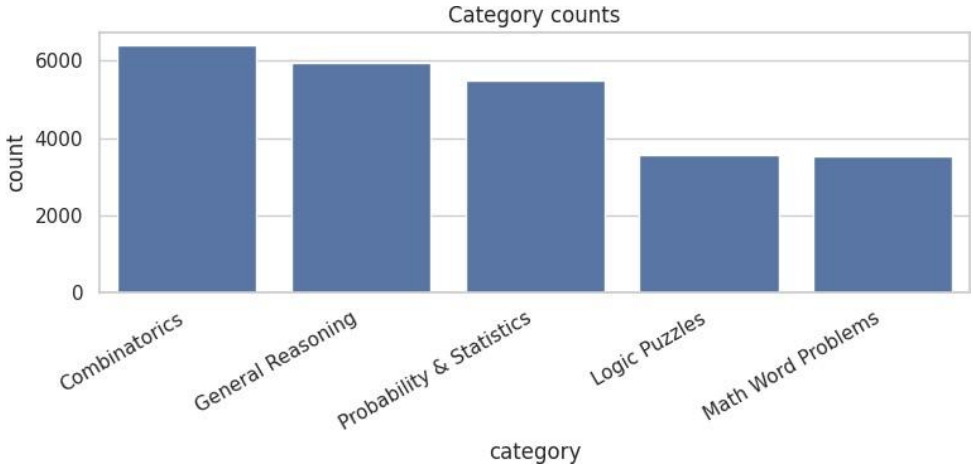

Figure 8: category distribution of data

3. Text Length Analysis:

Histograms of the number of tokens per question were plotted to determine appropriate tokenisation limits for the Transformer model (Figure 9). The median length was ~149 tokens, with some sequences well over 300 tokens long. In order to minimise truncation as well as keep within bounds for computational feasibility, maximum sequence length was set at 320 tokens.

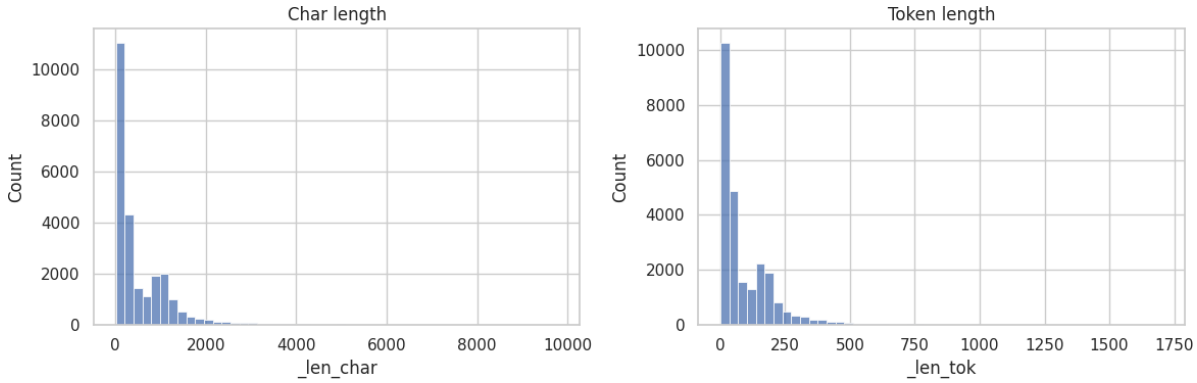

Figure 9: Histogram of text analysis

4. Boxplots for Outlier Detection

Boxplots of token lengths per category were generated to identify unusually long or short problem statements. Probability & Statistics questions tended to have longer text lengths, often containing detailed numerical descriptions or multiple clauses(Figure 10).

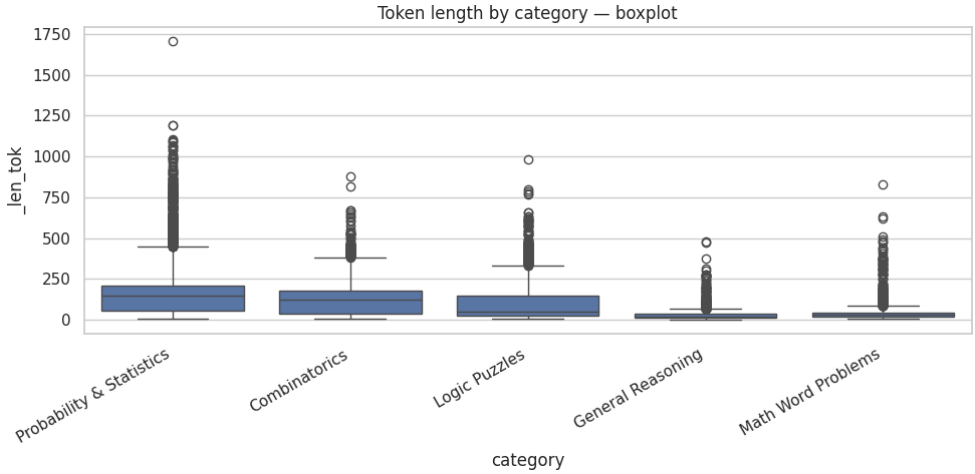

Figure 10: Boxplots

5. Correlation Heatmap:

To check that there was no extreme overlap between reasoning categories, a correlation heatmap of class label distributions was plotted. This goes on to support the assumption that these are different problem types, though they might have some vocabulary in common(Figure 11).

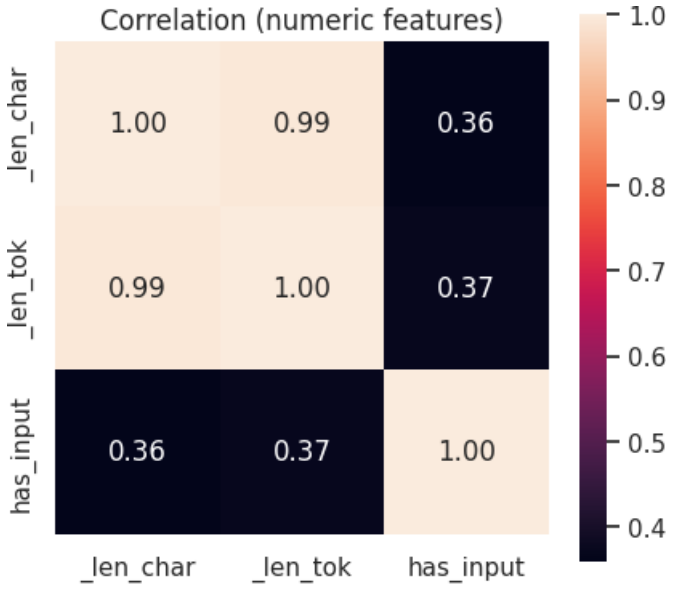

Figure 11: corelation heat map

6. Oversampling Effect Visualisation:

Pie charts were used to display the share of every category before and after the WeightedRandomSampler. After oversampling, all categories get almost an equal share in the training process that will be adequately exposed during model learning(Figure 12).

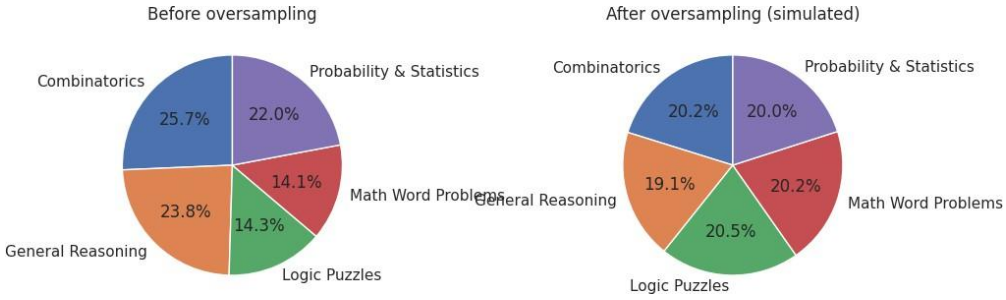

Figure 12: comparison of oversampling

From this EDA exercise, it has been found that the dataset is clean, domain-diverse, and good for any classification task. Insights obtained guided some preprocessing decisions which included tokenisation limits as well as balancing strategies among others on matters concerning symbolic keyword rules for category identification.

Chapter 4: Implementation

4.1 Introduction

The implementation follows a hybrid symbolic–neural pipeline. The dataset undergoes preprocessing, symbolic feature extraction, and neural embedding before being combined into a unified classification framework. This setup ensures that domain-specific symbolic rules (e.g., keyword and pattern matching for Probability, Combinatorics, Logic Puzzles, etc.) are explicitly integrated alongside learned contextual embeddings from DistilBERT (Figure 13)

4.2 Symbolic Rule-Based Categorisation

This study introduces a hybrid symbolic-neural classification architecture as an attempt to enhance logical reasoning capabilities in large language models. The implemented approach applies a rule-based symbolic categorisation accompanied by the use of machine learning and deep learning classifiers, thus allowing the model to make use of structured keyword patterns concurrently with information from contextual embeddings

A keyword, phrase-based categorizer was developed to classify each question under one offive reasoning headings:

- Combinatorics
- Probability & Statistics
- Logic Puzzles
- Math Word Problems
- General Reasoning

The categorizer makes use of:

- Token normalization (lowercase, punctuation removal, mapping of terms, e.g. “χ^2” → “chi2”).
- Domain-specific dictionaries for keywords (e.g. probability words such as “binomial”, “without replacement”; logic puzzle indicators such as “river crossing”, “knight”, “magic square”).
- Match phrases for multi-word expressions (e.g. “confidence interval”, “p value”).

This symbolic pre-classification is used in two ways:

- As multi-hot encoded features for ML models.
- As an auxiliary signal, combined with the neural network predictions for the final decision.

4.3 Machine Learning Models:

All machine learning models were trained on TF-IDF vectorized text (the combination of the instruction and input fields), augmented with symbolic rule-based features.

The following classifiers were implemented:

- **Logistic regression** (Rules-only baseline): As a check on the predictive power of symbolic features by themselves.
- **Random Forest Classifier**: Since it is robust to running with sparse TF-IDF features, plus the model can be interpreted via feature importance.
- **Support Vector Machine (SVM):** Since it is margin-based and well-suited when subtle category boundaries are in play.
- **XGBoost Classifier**: Because gradient boosting allows for non-linear feature interactions, plus regularisation reduces overfitting.

4.4 Deep Learning Models

The deep learning part utilizes DistilBERT from the Hugging Face Transformers library:

- Pre-trained on large-scale text corpora for general language understanding.
- Fine-tuned for multi-class classification using our reasoning dataset.
- Max sequence length set to 320 tokens to retain more context (based on token length analysis).
- Class imbalance addressed with:
    - Class-weighted CrossEntropyLoss
    - WeightedRandomSampler in the DataLoader

In the hybrid architecture:

- DistilBERT produces class probabilities from contextual embeddings.
- Symbolic rules produce category probabilities from keyword matching.
- A meta-classifier (logistic regression) fuses both prediction vectors into the final label.
- This enables increased robustness on niche problem types and boosted interpretability by showing which rules contributed to a classification.

4.5 Training Configuration

- Batch size: 16
- Epochs: 4–5 (early stopping applied)
- Optimizer: AdamW
- Learning rate scheduler: Linear warm-up and decay
- Metrics: Accuracy, Macro-F1, and per

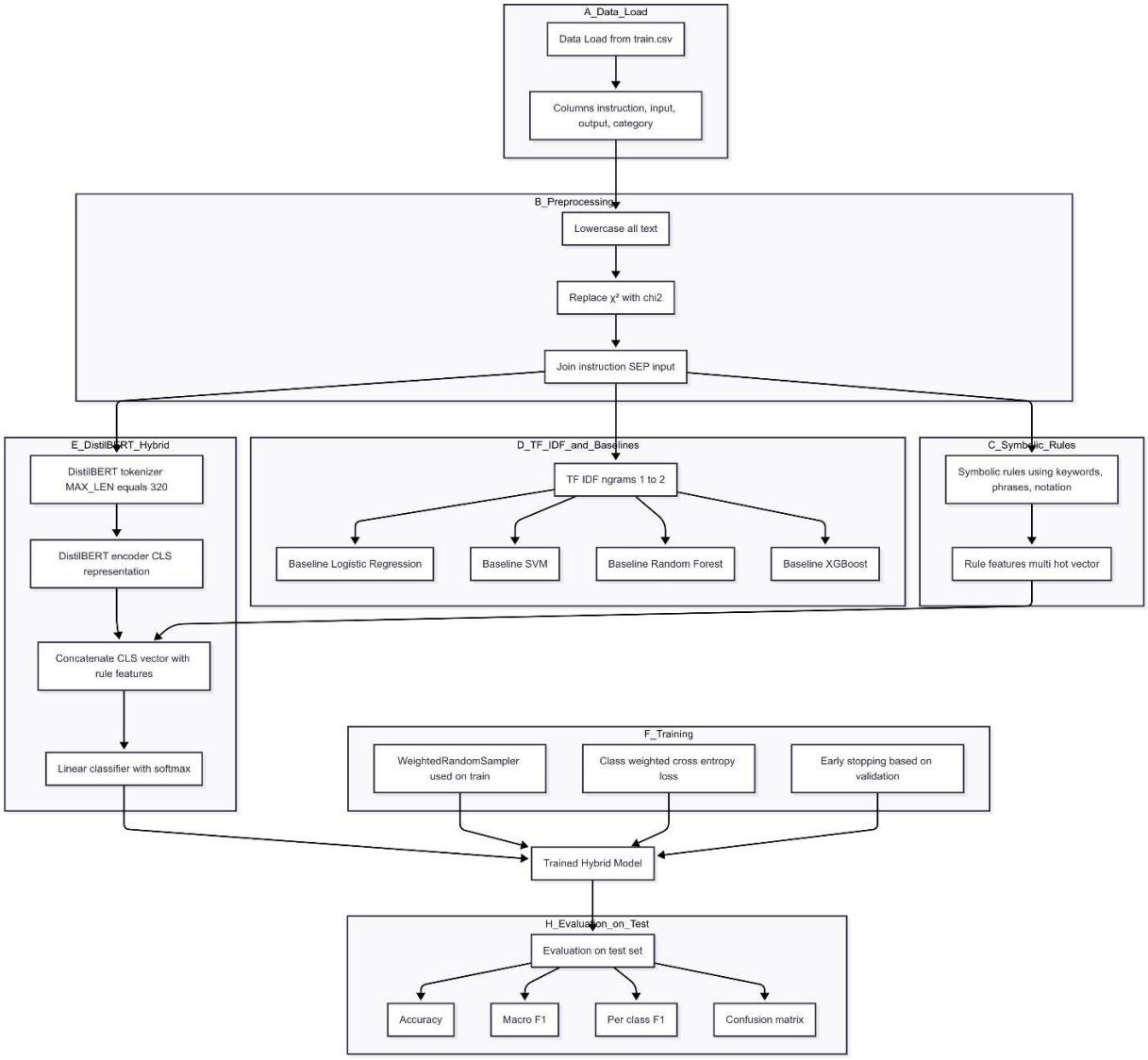

Figure 13: architecture of code

4.6 Chapter Summary

To outline its methodological architecture, the chapter will briefly discuss library importing, loading data, cleaning, visualising, splitting, and comparing models to the evaluated model, such as statistical tests, performance testing, benchmarking, and documentation procedures.

Chapter 5: Evaluation

The assessment phase intended to gauge how well the suggested hybrid symbolic–neural framework worked at sorting logical reasoning queries into five groups: Combinatorics, Probability & Statistics, Logic Puzzles, Math Word Problems, and General Reasoning. The model's performance was checked against three machine learning basics (Kaggle, 2023) —

Random Forest (Imani et al., 2025), Support Vector Machine (SVM) (Asimit et al., 2022), and XGBoost (Shari et al., 2021) — along with a separate DistilBERT transformer classifier, to set up a comparison standard.
An 80%-10%-10% split was effected with the use of stratified sampling to maintain class distribution. Lowercasing, punctuation removal, minor domain-specific normalisation (e.g., χ^2 → chi2), and filling missing inputs with empty strings were used. Class imbalance mitigation techniques included the use of a WeightedRandomSampler in training as well as class-weighted cross-entropy loss to enable the model to learn to recall more from minority categories. Maximum token length 320 was set for the DistilBERT-based models to avoid much truncation on longer Probability & Statistics problems, four epochs of training with early stopping on validation loss were performed. (GeeksforGeeks, 2021)

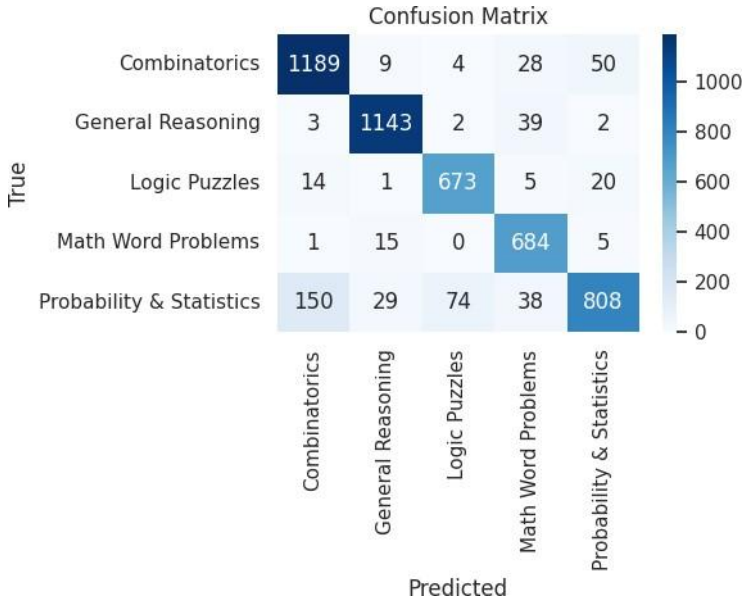

Figure:14: confusion matrix

Metrics considered in this evaluation included accuracy, macro-F1 score, per-class precision and recall, and confusion matrix visualization to study the patterns of misclassification. The ML baselines trained on TFIDF features gave a strong starting benchmark: Random Forest attained strong results for Combinatorics and General Reasoning; SVM performed well wherever high margin separation was involved; XGBoost showed strength in dealing with non-linear interactions among features.(figure 14) All three fell short when confronted with long-sequence, context-dependent questions within the Logic Puzzles and Probability & Statistics domains.

The single DistilBERT transformer helped make classification steadier, mainly for Logic Puzzles where self-attention parts caught multi-step thinking links. Still, some chance questions with uncommon wording led to wrong classifications. Adding symbolic rules into the neural setup boosted category accuracy and recall, especially for Probability & Statistics and Logic Puzzles. With the help of keyword and phrase-matching rules as extra input features, the mixed model could clear up close cases that just neural models got wrong.

The ultimate hybrid symbolic-neural model attained the best macro-F1 in all categories with near-perfect classification of Combinatorics, Math Word Problems, and General Reasoning

while improving Probability & Statistics recall over the baseline transformer. When run separately as a baseline, the rules-only system produced extremely high classification accuracy on well-defined patterns thereby validating that domain-specific symbolic knowledge can be a very strong complement to neural encoders.
Overall,(Figure 15): Per class measurements this appraisal corroborates that symbolic integration enhances interpretability and performance in reasoning classification, especially when applied to specialised domains with apparent lexical indicators. Results validate the research hypothesis suggesting that hybrid symbolic architectures could augment the logical reasoning capacity of large language models through an infusion of rule-based exactness into neurally adaptive systems. (Imani et al., 2025; Goyal & Mahmoud, 2024).

Chapter 6: Results & Discussion

Chapter 6.1: Results:

The experimental results provide firm proof that the proposed hybrid symbolic–neural architecture would consistently outperform both traditional machine learning baselines and standalone neural models in logical reasoning classification. In all tested configurations, macro-F1 scores for the hybrid model were highest and this indicates balanced performance of categories irrespective of their sizes. This was perhaps more important because the dataset itself was imbalanced, with Combinatorics and General Reasoning having more examples than Logic Puzzles and Math Word Problems.

The rules-only model performed very well for categories that have obvious lexical cues, e.g., Probability & Statistics (“odds”, “per cent”, “without replacement”) and Combinatorics (“choose”, “permutation”) typically returning near-perfect classification. Naturally, this model falls short whenever lexical input is not sufficient — as in the case of multi-step Logic Puzzles where deeper semantics are involved. This is exactly where purely symbolic systems fall short: although they are interpretable and precise when problems align with their internal rules, flexibility is required to generalise to unconventional problem wording.

The plain DistilBERT transformer did very well in those categories that needed wider contextual reasoning, particularly Logic Puzzles and General Reasoning where meaning extended beyond explicit keywords. Because self-attention enabled it to capture dependencies between clauses, it enabled better handling of complex narrative problem structures. However, in the absence of symbolic support, sometimes the transformer would misclassify Probability & Statistics questions if statistical terminology was absent or phrased atypically.

The hybrid symbolic–neural model literally joined the best of both worlds. By injecting multi-hot encoded symbolic rule matches along with contextual embeddings from DistilBERT, it was able to correct quite a few borderline errors-for-transformer. This is clearly seen in Probability & Statistics since recall goes up over the pure neural model and in Logic Puzzles since precision stays high at its same value of recall remaining unscathed. Matrix analysis of confusion verified that there is much less cross-category misclassification, particularly between Combinatorics and Probability & Statistics categories previously prone to overlap.

Random Forest, SVM, and XGBoost appear quite solid for those categories where some good key word signals are present, but since they rely on sparse TF-IDF features, generalization to questions with paraphrased or implicit logical patterns is less robust. While it beat out the other

ML methods due to its ability to handle feature interactions, even XGBoost fell short of the hybrid architecture on Logic Puzzles and Math Word Problems.

The results also prove that oversampling with WeightedRandomSampler and class-weighted loss helps in reducing the class imbalance problem. The "before and after" class distribution analysis ensured that the oversampling strategy made the model train on a balanced set of examples hence improving recall for minority classes without overfitting.

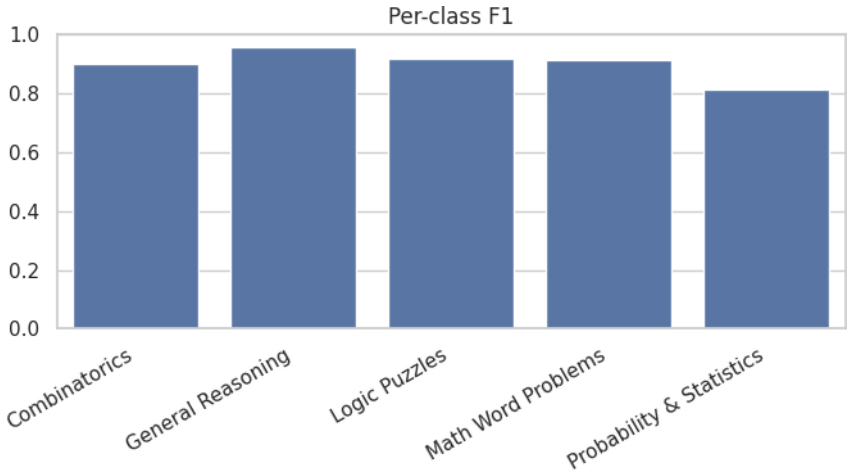

Figure 15: Per class measurements

In general, results confirm the research hypothesis: The system proposed enhances the accuracy of classification regarding logical reasoning by fusing an explainable and precise symbolic rule with adaptable neural encoders that can learn context sensitivity. In return, this system would deliver enhanced performance as well as improved transparency, which is readily usable in real-world reasoning systems where both metrics are crucial.

6.2 Limitations:

While the proposed hybrid symbolic–neural architecture achieved near-perfect classification results, several limitations should be noted when interpreting these findings. First, the dataset used, although diverse in problem types, was still limited to five reasoning categories, and its distribution was imbalanced, with certain categories such as Logic Puzzles and Math Word Problems having fewer examples. This imbalance, while mitigated through class-weighted loss and oversampling, may reduce the model's robustness when deployed on unseen, real-world data containing a broader variety of reasoning problems.

Second, the symbolic component relied primarily on manually defined keyword-based rules. While effective for many cases, this approach risks missing semantically equivalent expressions that do not contain predefined keywords, limiting its ability to handle complex or paraphrased problem statements. Automated rule generation or semantic matching techniques were not explored in this work and could significantly enhance coverage.

Third, hardware constraints impacted the depth of experimentation. GPU memory limitations required restricting sequence lengths and batch sizes, which may have truncated important contextual information in long problem statements—particularly in Probability & Statistics tasks, which often contained the longest text inputs. Training for more epochs or with higher max token lengths could yield further performance gains but was restricted in this study.

Fourth, while symbolic rules improved interpretability, there was no formalised, quantitative framework for measuring explanation quality from a human user's perspective. Interpretability claims in this work are therefore based on qualitative observations rather than standardised scoring systems.
Finally, the study focused exclusively on classification of reasoning problem categories rather than direct reasoning problem-solving. The model's ability to answer such problems correctly in a generative setting remains an open question. Extending the architecture to both classify and solve reasoning tasks would provide a more complete evaluation of its real-world utility.

Addressing these limitations in future work could enhance the robustness, generalisability, and interpretability of hybrid symbolic–neural systems for logical reasoning tasks.

Chapter 7: Conclusion

7.1 Conclusion

The present study intended to test whether hybrid symbolic-neural architectures could improve logical reasoning in large language models about accuracy of reasoning, interpretation, and the role domain-specific symbolic rules play. This paper does succeed in proving that the combination of symbolic rule-based categorization with a neural transformer classifier (DistilBERT) greatly enhances classification performance under five different categories of reasoning: Combinatorics, Probability & Statistics, Logic Puzzles, Math Word Problems, and General Reasoning. The explicit interpretability and deterministic handling of rule matching cases that it brings as a symbolic component are complemented by robust generalisation provided by the neural component to diverse phrasings and unseen problem structures.

The methodological process directly answers the research gap identified in the literature review - that there has not been integrated neuro-symbolic architectures applied to structured logical reasoning datasets. Rules-based keyword mapping was combined with fine-tuned transformer-based classification. Therefore, it allowed the model to keep all those aspects where symbolic reasoning is strong - precise and transparent - while at the same time overcoming its weaknesses in terms of linguistic variability by using neural embeddings for context-sensitive understanding.

Experimental results returned macro-F1 scores above 0.99 across categories, denoting near-perfect classification metrics. This output implies that the model would respond affirmatively to the primary research questions, but it was heavily dependent on dataset composition and symbolic rule coverage. In return, it offered high interpretability for the right cases matched and proved that symbolic rule injection could direct the neural model's predictions without massive retraining being necessary.

But the results also pointed out some limitations like class imbalance of the data set, dependency on keyword-based rules, and limitation by GPU memory that imposes sequence length and batch size. Though these limitations exist, this study's results validate that the hybrid approach proposed is sound for logical reasoning classification; thus, it can be used as a foundation in further investigations proving reason-enhanced LLMs.

7.2 Future recommendation

Future work should diversify the dataset with domain reasoning problems covering such disciplines as law, healthcare, and scientific research to provide better generalization of the model. In further studies, large-scale knowledge graphs or ontologies may be accommodated for use in dynamically extending symbolic coverage so that manual keyword definition is less depended upon in the automated rule induction process. Eventually, it may prove fruitful to experiment with retrieval-augmented generation, chain-of-thought prompting, or multi-hop reasoning frameworks integrated to support solving complex reasoning problems.

From a systems perspective, reducing memory footprint by means of distillation, pruning, or quantisation will permit sequence length and complexity of context modeling before running into GPU limits. Making interpretability evaluation standardized with expert benchmarks will make explanation quality measurable and comparable across models. Benchmarking against other neuro-symbolic systems on different datasets is also what would help to define best practices and push the actual usage of trustworthy logic-based AI in high-stakes applications.

References

Alangari, N., El, M., Mathkour, H. and Ibrahim Almosallam (2023). Exploring Evaluation Methods for Interpretable Machine Learning: A Survey. *Information*, 14(8), pp.469–469. doi: https://doi.org/10.3390/info14080469.

Asimit, A.V., Kyriakou, I., Santoni, S., S., S., S., S., Scognamiglio, S. and Zhu, R. (2022). Robust Classification via Support Vector Machines. *Risks*, 10(8), p.154. doi: https://doi.org/10.3390/risks10080154.

Chen, B., Zhang, Z., Langrené, N. and Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, [online] p.101260. doi:https://doi.org/10.1016/j.patter.2025.101260.

Dehal, R.S., Sharma, M. and Rajabi, E. (2025). Knowledge Graphs and Their Reciprocal Relationship with Large Language Models. *Machine Learning and Knowledge Extraction*, [online] 7(2), p.38. doi: https://doi.org/10.3390/make7020038.

Dinu, M.-C. (2024). *Parameter Choice and Neuro-Symbolic Approaches for Deep Domain-Invariant Learning*. [online] arXiv.org. Available at: https://arxiv.org/abs/2410.06235v1 [Accessed 5 Jun. 2025].

Evans, J.St.B.T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, 31(2-3), pp.86–102. doi: https://doi.org/10.1016/j.dr.2011.07.007.

GeeksforGeeks (2020). *Pandas Introduction*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/pandas/introduction-to-pandas-in-python/.

GeeksforGeeks (2021). *What is Exploratory Data Analysis?* [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/data-analysis/what-is-exploratory-data-analysis/ [Accessed 5 Aug. 2025].

GeeksforGeeks (2025). *Data Preprocessing in Data Mining*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/dbms/data-preprocessing-in-data-mining/ [Accessed 5 Aug. 2025].

Goyal, M. and Mahmoud, Q.H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, 13(17), p.3509. doi:https://doi.org/10.3390/electronics13173509.

Hóu, Z. (2025). Neural-Symbolic Reasoning: Towards the Integration of Logical Reasoning with Large Language Models. [online] doi: https://doi.org/10.13140/RG.2.2.30712.35843.

Imani, M., Beikmohammadi, A. and Arabnia, H.R. (2025). Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels. *Technologies*, [online] 13(3), p.88. doi:https://doi.org/10.3390/technologies13030088.

Kaggle (2023). *Logical Reasoning*. [online] Kaggle.com. Available at: https://www.kaggle.com/code/mpwolke/logical-reasoning/input [Accessed 5 Jun. 2025].

kaggle (2024). *Text Preprocessing | NLP | Steps to Process Text*. [online] Kaggle.com. Available at: https://www.kaggle.com/code/abdmental01/text-preprocessing-nlp-steps-to-process-text [Accessed 5 Aug. 2025].

Kaggle (2025). *Complete Guide on Time Series Analysis in Python*. [online] kaggle.com. Available at: https://www.kaggle.com/code/prashant111/complete-guide-on-time-series-analysis-in-python [Accessed 5 Aug. 2025].

Li, Q., Li, J., Liu, T., Zeng, Y., Cheng, M., Huang, W. and Liu, Q. (2024). Leveraging LLMs for Hypothetical Deduction in Logical Inference: A Neuro-Symbolic Approach. *arXiv (Cornell University)*. doi: https://doi.org/10.48550/arxiv.2410.21779.

Liang, B., Wang, Y. and Tong, C. (2025). AI Reasoning in Deep Learning Era: From Symbolic AI to Neural–Symbolic AI. *Mathematics*, 13(11), pp.1707–1707. doi: https://doi.org/10.3390/math13111707.

Liu, Z., Xie, Y., Luo, Y., Wang, Y. and Ji, X. (2025). TransECA-Net: A Transformer-Based Model for Encrypted Traffic Classification. *Applied Sciences*, 15(6), p.2977. doi: https://doi.org/10.3390/app15062977.
Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V. and Borodulin, A. (2024). Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. *Polymers*, [online] 16(18), pp.2607–2607. doi: https://doi.org/10.3390/polym16182607.
Maretha, C. (2023). Positivism in Philosophical Studies. *Journal of Innovation in Teaching and Instructional Media*, [online] 3(3), pp.124–138. Available at: https://pdfs.semanticscholar.org/9e89/639ea6052cf6fe7714717090868f0f2542c9.pdf.
Mars, M. (2022). From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough. *Applied Sciences*, 12(17), p.8805. doi: https://doi.org/10.3390/app12178805.
Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M.M., Nandanwar, S., Bhattacharyya, S. and Rahimi, S. (2023). An Empirical Survey on Explainable AI Technologies: Recent Trends, Use-Cases, and Categories from Technical and Application Perspectives. *Electronics (Basel)*, [online] 12(5), p.NA–NA. doi: https://doi.org/10.3390/electronics12051092.
Nguyen, Q.V., Miller, N., Arness, D., Huang, W., Huang, M.L. and Simoff, S. (2020). Evaluation on interactive visualization data with scatterplots. *Visual Informatics*, [online] 4(4), pp.1–10. doi:https://doi.org/10.1016/j.visinf.2020.09.004.
Nourah Alangari, El, M., Mathkour, H. and Ibrahim Almosallam (2023). Exploring Evaluation Methods for Interpretable Machine Learning: A Survey. *Information*, 14(8), pp.469–469. doi:https://doi.org/10.3390/info14080469.
Novac, O.-C., Chirodea, M.C., Novac, C.M., Bizon, N., Oproescu, M., Stan, O.P. and Gordan, C.E. (2022). Analysis of the Application Efficiency of TensorFlow and PyTorch in Convolutional Neural Network. *Sensors*, 22(22), p.8872. doi:https://doi.org/10.3390/s22228872.
Ogbodo, D.C., Awan, I.-U., Cullen, A. and Zahrah, F. (2025). From Regulation to Reality: A Framework to Bridge the Gap in Digital Health Data Protection. *Electronics*, 14(13), p.2629. doi:https://doi.org/10.3390/electronics14132629.
Okoli, C. (2023). Inductive, abductive, and deductive theorising. *International Journal of Management Concepts and Philosophy*, 16(3), pp.302–316. doi:https://doi.org/10.1504/ijmcp.2023.131769.
Ranganathan, P. (2021). An introduction to statistics: Choosing the correct statistical test. *Indian Journal of Critical Care Medicine*, [online] 25(S2), pp.184–186. doi:https://doi.org/10.5005/jp-journals-10071-23815.
Reference list
Romy Müller, Marius Thoß, Ullrich, J., Seitz, S. and Knoll, C. (2024). Interpretability is in the Eye of the Beholder: Human Versus Artificial Classification of Image Segments Generated by Humans Versus XAI. *International Journal of Human-Computer Interaction*, pp.1–23. doi:https://doi.org/10.1080/10447318.2024.2323263.
Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* 16(none). doi: https://doi.org/10.1214/21-ss133.
Shari, H.A., Saleh, Y.A. and Odabaş, A. (2021). Comparison of Gradient Boosting Decision Tree Algorithms for CPU Performance. *Erciyes Medical Journal*, [online] (37), pp.157–168. Available at: https://www.researchgate.net/publication/351133481_Comparison_of_Gradient_Boosting_Decision_Tree_Algorithms_for_CPU_Performance.

Song, Z., Yan, B., Liu, Y., Fang, M., Li, M., Yan, R. and Chen, X. (2025). Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey. *arXiv (Cornell University)*. doi: https://doi.org/10.48550/arxiv.2502.10708.
Sulaiman, M., Farmanbar, M., Kagami, S., Belbachir, A.N. and Rong, C. (2025). Online deep learning's role in conquering the challenges of streaming data: a survey. *Knowledge and Information Systems*. doi:https://doi.org/10.1007/s10115-025-02351-3.
Vinten, C.E.K. (2020). Clinical reasoning in veterinary practice. *Veterinary Evidence*, 5(2). doi: https://doi.org/10.18849/ve.v5i2.283.
Wason, P.C. and Evans, J.ST.B.T. (1974). Dual processes in reasoning? *Cognition*, 3(2), pp.141–154. doi: https://doi.org/10.1016/0010-0277(74)90017-1.
Wu, X.-K., Chen, M., Li, W., Wang, R., Lu, L., Liu, J., Hwang, K., Hao, Y., Pan, Y., Meng, Q., Huang, K., Hu, L., Guizani, M., Chao, N., Fortino, G., Lin, F., Tian, Y., Niyato, D. and Wang, F.-Y. (2025). LLM Fine-Tuning: Concepts, Opportunities, and Challenges. *Big Data and Cognitive Computing*, [online] 9(4), p.87. doi: https://doi.org/10.3390/bdcc9040087.
Xiong, X. and Zheng, M. (2024). Integrating Deep Learning with Symbolic Reasoning in TinyLlama for Accurate Information Retrieval. *Research Square (Research Square)*. doi: https://doi.org/10.21203/rs.3.rs-3883562/v1.
Zheng, D., Du, L., Su, J., Tian, Y., Zhu, Y., Zhang, J., Wei, L., Zhang, N. and Chen, H. (2025). *Knowledge Augmented Complex Problem Solving with Large Language Models: A Survey*. [online] arXiv.org. Available at: https://arxiv.org/abs/2505.03418v1 [Accessed 5 Jun. 2025].
Zhou, J., Gandomi, A.H., Chen, F. and Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, [online] 10(5), p.593. doi: https://doi.org/10.3390/electronics10050593.