**Do LLM’s (Large Language Model) understand human logic? If so, at what level? Can LLM’s understanding of human logic be improved?**

requirements document

**Data**

We will create a data set with few hundreds of questions templates (each time user can change certain aspect of the question so the model won’t recognize that question). we will also create a prompt database (those that was most successful) so the user could check what prompt works better with which type of questions.

**Data Input (Prompts)**

The system will be able to receive textual inputs from the user in the form of multiple-choice questions (which include multiple possible answers), binary classification questions (questions where the answer is either "True" or "False"), or open-ended questions (where the user provides a free-text answer). The inputs can include:

* Multiple-choice questions with several answer options.
* Binary classification questions with one answer from two options (True/False).
* Open-ended questions where the user provides a free-text response.

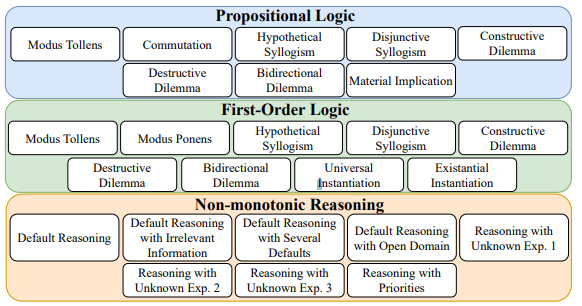
**LLM**

LLM’s (Large language Models) and how should we use them

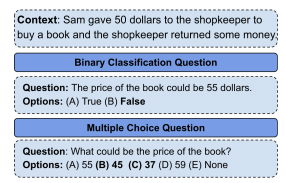
* We would choose models from the most popular LLM’s that currently working
* We will asses those model and choose the best according to our initial score assessment and try to improve it.

**logic**

* The research will test a few types of human logic - Propositional Logic, Propositional Logic and Non-monotonic Reasoning.



* In the research, we will test the logic in a few ways – binary classification questions, multiple-choice questions, and open-ended questions.



**User experience**

The user will be able to choose a LLM , a question type and a prompt and the system will give the LLM a score based on his performance on the question, so the user will experience our study result through the app.

**Output**

* By the end we expect to have the best model in each logical question category we examine.
* Also we expect to get the improvement (if we succeed) value for each model we test .

**Score assessment and bias inspection**

* We will build a Database with questions templates and in every iteration we will change some aspect of the question and evaluate the model based on it
* The model will be evaluated based on its ability to identify correct or incorrect logical arguments and provide answers that correctly apply logical principles.
* Clear criteria must be defined for evaluating the accuracy of the model's answers in relation to correct or incorrect logical arguments.
* To prevent Bias we will Have multiple question from each Logical category and compare to other studies that have already done this kind of assessment.

**Success Metrics for the Test:**

**Success Metrics for Multiple-Choice Questions**For multiple-choice questions, the success metrics include the following:

* Percentage of correct answers: The percentage of correct choices out of all the options in the question.
* Logical accuracy: How well the chosen answer aligns with the logical principles and reasoning behind the question. Does the system consistently rule out the other options based on logical judgment?
* Relevance: How directly the answer addresses the question asked, providing a solution to the exact problem presented.
* Alignment with human logic: How well the chosen answer aligns with human understanding of the question and the reasoning behind it.

**Success Metrics for Binary Classification Questions**For binary classification questions, the success metrics include the following:

* Percentage of correct answers: The percentage of correct answers (True/False) in relation to the questions.
* Logical accuracy: How well the answer aligns with the logical reasoning of the question (for example, if the question deals with a factual statement, is the answer logically correct?).
* Clarity of the answer: Whether the answer provides a brief and clear explanation about the question and the decision made.

**Success Metrics for Open-Ended Questions**For open-ended questions, the success metrics include the following:

* Accuracy of the answer: How correct the answer is in terms of logic and the principles behind the question.
* Clarity and explanation: How clear and understandable the answer is. Does it provide a full and understandable explanation that justifies the given answer?
* Relevance of the answer: How directly the answer addresses the question and provides an accurate solution or explanation.
* Academic and professional level: For questions requiring academic or professional processing, whether the answer meets the correct and professional standards of the field.

**Overall Success Metric**The overall score of the test will be calculated as the average score across all metrics, with each question assessed separately according to its type (multiple-choice, binary classification, or open-ended). Each metric will be evaluated based on the following criteria:

* Percentage of correct answers: The system's ability to choose the correct answer from all the options.
* Logical accuracy: The system's ability to identify and respond logically, based on principles.
* Clarity of answers: Whether the answers are presented clearly and understandably to the user.
* Relevance of answers: How well the answers fit the questions asked.