School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science at the George Washington University

COMPREHENSIVE ANALYSIS AND IMPLEMENTATION OF SMALL-SCALE LLMS FOR HUMAN-LIKE MACHINE CAUSAL REASONING ACROSS MULTIPLE LANGUAGES

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Toghrul Tahirov

Supervisor Dr. Samir Rustamov

April 2024

ABSTRACT

This study was conducted with the primary focus of exploring the potential of small-scale Large Language Models (LLMs) to handle causal reasoning tasks through two primary interventions, namely, dataset augmentation and instruction fine-tuning. The employed **CRASS** and **Tübingen** datasets were augmented with GPT-4 to include additional examples and prompts that defined a wider range of cause-effect relationships and counterfactual reasoning examples. Moreover, the created augmented dataset was further extended to include an Azerbaijani-language counterpart through Google's MultiLingual BERT (MBERT) and manual modifications to address the lack of resources in language modelling for the Azerbaijani language. Despite the small volume of the currently developed dataset, 20000 samples at the moment of publication of this article, this contribution is of great value as it enables the researchers to test their yet to be developed LLMs that can efficiently handle the Azerbaijani Language on reasoning tasks.

During the investigation into the performance of small-scale LLMs on reasoning tasks, we identified a few unexpected failure modes where the LLM could not follow the instructions present in the prompt. Another interesting aspect of the outcomes was that even GPT4 displayed a tendency to prefer options that are presented first when multiple-choice types of questions are involved. This phenomena was less evident in smaller models like gemma-7b-it and Mistral-7B-int8, and could potentially be associated with improper placement of the transformer attention based on the context that resides within the text.

We then conducted instruction fine-tuning using the augmented Tubingen and CRASS datasets on Google's gemma-7b-it model. Low-Ranked Adaptation (LoRA) technique was utilized to reduce the computational requirements of the model during training time and Cosine Annealing Learning Scheduler and Cross Entropy Loss metrics were employed for quality control purposes. The fine-tuning process was carried on on both high-end Nvidia V100 GPUs and on the consumer level Nvidia RTX 3090 using the transformers library for model training, and the WandB service for monitoring and logging. The fine-tuning step used full-precision computations without quantization. The evaluation presented in this study demonstrated that the performance of small-scale LLMs can be significantly improved through a combination of dataset augmentation and instruction fine-tuning for causal reasoning tasks.

The results demonstrated that fine-tuned small-scale LLMs had comparable accuracies with only around 8 billion parameters to a much larger, 1.7 trillion parameter, GPT4 model when provided with properly augmented datasets and instructions for counterfactual scenarios. The accuracy of the fine-tuned **gemma-7b-it model** improved from **53%** to **85%** on the Tubingen dataset, and from **72%** to about **84%** on the CRASS benchmark. Therefore, it was concluded that smaller models could potentially match in performance with their much larger counterparts through elevated dataset quality and targeted instruction fine-tuning for deeper understanding in causal reasoning. We also discussed how development of task specific expert small-scale LLMs adept at causal reasoning can pave the way for widespread adoption of the said models across many industries where analytical thinking, decision making and problem solving are at the core of operations. In this context, our results demonstrate that by further fine-tuning small-scale LLMs to become experts at a specific reasoning task, development of a net of small-scale LLMs that offer different perspectives into a problem through controlled bias to aid humans in fast, accurate and thorough decision-making is just a few papers away.

For reproducibility purposes, the code base and the results of this study is presented in this Github repository [**?**], while the Azerbaijani dataset and an interactive comparison dahsboard of the LLMs fine-tuned or studied in this work is hosted in this HuggingFace repository [**?**].

1 Introduction

Causal reasoning is the process of establishing cause-and-effect relationships, which is a concept critical to human decision-making, situational analysis, and problem-solving. In the context of Artificial Intelligence, this notion refers to a model's ability to understand and predict potential outcomes from the underlying causal reasoning in the observed data. This ability is essential in various domains, such as healthcare, where it is used to diagnose diseases and prescribe treatment, finance for risk assessment and investing, and in autonomous systems for navigation, and analysis of an interdependent web of factors. AI models that can identify, follow, and predict forward from a set of circumstances and the correlations they reveal can also be used to develop autonomous systems that analyze a situation's factors from multiple perspectives. Such systems can help people make decisions by offering different perspectives and minimizing the bias of the final decision.

This specific domain of application is a particular focus of current AI research, as correlational models are notoriously unreliable in the less-ideal situations outside of data and the conditions present in the training environment. However, the achievement of human-like causal reasoning by AI models even in these circumstances where it differs little from advanced correlational analytics is a significant challenge. Typical machine learning models are proficient in identifying correlations between the variables but are unable to establish causation, leading to highly misleading, unpredictable, and erroneous conclusions in the presence of noise. In addition, the actual data is often complex and high-dimensional, and extracting the causal relationships governing the situation is often impossible. This difficulty is further exacerbated by the dynamic nature of the relationships, which change and evolve outside the experiment designs associated with them.

1.1 Role of Large LLMs in Causal Reasoning

Large Language Models (LLMs), having been trained on vast amounts of text data, have demonstrated an ability to understand and generate human-like language. This aspect of LLMs provides arises the question whether more sophisticated causal reasoning is possible. Given that the large body of readily available, human-written data, both in scientific and non-scientific literature, already contain cause-effect based correlations and sequences of events conditioned to lead to a specific outcome, it is expected for LLMs to be capable of understanding the general patterns with enough pretraining.

Thus, LLMs can possibly leverage their linguistic comprehension capabilities to infer causal relationships from text which could potentially lead to development of applications such as automated reasoning, question answering, and hypothesis generation. However, the architectural complexity of the large transformer models and resource requirements of training such models is an issue when it comes to widespread adoption, especially in real-time or resource-constrained environments. And even though proof of concept implementations of the encoder-decoder architectures, such as GPT and BERT were pivotal during the historical development of LLMs and advancements in Natural Language Processing (NLP), and the ground-breaking opportunities they present, these limitation are still rather preventative for most industries.

One such issue is that the LLMs training data is quality checked mostly for correlational integrality, that is, whether the text is grammatically, structurally, and somewhat logically coherent or not. Absence of strict policies regarding the underlying meaning, story-telling in the training data, and the fact that most of the high quality human

written text exists in copyright content, can create biases for LLMs that inhibits causal inferences. This means that the model pays more attention to grammatically correct text that somewhat makes sense than the ideas and the principles that exist in the text.

Moreover, these models typically usually need to be extensively fine-tuned to perform causal reasoning tasks effectively, which is costly for LLMs with very large parameter sizes. Additionally, the inherent black-box nature of these models complicates the interpretation of their reasoning processes, which is counter-productive for applications where understanding the model's decision-making is of importance. This lack of interpretability is a significant hurdle in sensitive fields such as healthcare and legal advisement, where explaining decisions made by AI systems is often required by regulation. Some workaround for this shortcoming, have recently been introduced in the form of LLMs that have been trained to provide extensive explanations when generating text or prompt engineering techniques that help achieve this.

Another limitation is the contextual generalization of LLMs. While they perform well within the distribution of their training data, their ability to generalize to new contexts or datasets—especially those with different underlying causal structures—is often limited. Given the fact that the firing of intermittent connections between the layers of the transformer architectures is not directly controllable, it is uncertain exactly which part of the distribution of the training data that has been mapped to a higher dimensional space the results will be generated from.

This leads to a discussion on another concern which is the quality, quantity of the data, and the breadth of the domains that were used as sources. Having widely varying sources of training data inevitably inhibits the model's ability to learn a specific domain in depth, leading to hallucinations. An LLM's inability to comprehend the sequences of human thought processes that weave together to explain one phenomenon becomes a hurdle for development of causal models. Presence of poorly written content, logical fallacies, implicit and vague reasoning in story telling only serve to further complicate the issue.

1.2 The need for smaller, more efficient LLMs for causal reasoning

As aforementioned, the inference and training time computational requirements of the large, multifaceted and generally very capable LLMs that perform well on varying benchmarks is a prohibiting factor for their wide adoption. At the current state of the industry, LLMs that are capable enough to be introduced to companies are offered mostly by big tech companies that have developed these models on proprietary data based on in-house developed architecture and technologies. This is a welcomed development as the end user does not have to shoulder the infrastructure costs. However, this introduces well-rooted security and privacy concerns. For this reason, companies or industries that have strict security policies are looking for alternative approaches that will not require purchase and maintenance of extremely expensive hardware.

Furthermore, it is important to consider that a lot of systems that enable the currently available technologies used on a daily basis are built upon edge, embedded or mobile platforms where computational resources are limited. This further complicates the introduction of causal reasoning capable systems into crucial infrastructure.

To address the said drawbacks, it is imperative to develop smaller, more efficient models that have the advanced capabilities of larger models but at a fraction of the computa-

tional cost. These models would be particularly advantageous for real-time applications, such as mobile apps, wearable technology, and edge devices, where quick and efficient processing is needed. Another noteworthy domain of application is having access to a swarm of smaller, faster causal models that can represent diverse viewpoints and analyze the situation with selective bias.

Reduction in size, improvements in efficiency of LLMs and development of tailored data for decision-making involves various strategies, including model pruning, knowledge distillation, and quantization. Model pruning reduces the model size by eliminating non-critical parameters, potentially without significant loss in performance. Knowledge distillation involves training a smaller "student" model to replicate the behavior of a larger "teacher" model, effectively transferring the capabilities without the original model's size. Quantization reduces the precision of the numerical values used in computations, and decreasing the model's memory requirements and speeding up inference.

However, employment of these strategies effectively while ensuring that the pre-training based causal reasoning capabilities of the models are not compromised remains a significant research gap. This specific gap represents a critical area of development in the field of AI.

1.3 Objectives of this Study

1.3.1 Exploring the Potential of Small-Scale LLMs in Human-like Causal Reasoning

The main objective of this study was exploration and analysis of the capabilities of small-scale Large Language Models in performing human-like causal reasoning. Through this, we investigated whether reduced-scale models can achieve performance levels comparable to their larger counterparts in detecting, understanding and generating causal relationships. Due to the aforementioned current constraint in the applicability of large LLMs, our study is crucial for broadening the adoption of LLMs. Special point of concern is the environments where computational resources are limited, but the need for causal reasoning in decision-making remains high.

1.3.2 Enhancing the Performance of Small-Scale LLMs Models Through Fine-Tuning on Causal Reasoning Data and Quantization

The second objective is optimizing the observed performance of small-scale LLMs specifically on causal reasoning tasks. We employed two main strategies, namely fine-tuning and quantization. Fine-tuning small-scale LLMs on specifically curated datasets that focus on causal relationships will help in adapting the pre-trained transformer models to better align with the demands of this domain of application. This process not only enhances the model's accuracy in causal inference but also its generalizability across different types of causal data.

Quantization, on the other hand, was used for the purposes of reducing the numeric precision of the values that represent the model parameters, thus decreasing the model's memory footprint and computational requirements. An example is a scaled shift from using full precision, 32-bit floating point, to 8-bit integer for in-RAM storage of the model parameters. This technique is particularly useful for deploying AI in resource-constrained environments. It allows the streamlined models to operate more efficiently without a substantial compromise in performance. The research explored various levels

of quantization to find an optimal balance where the reduction in resource consumption does not detrimentally affect the model's ability to perform complex causal reasoning.

1.3.3 Development of Causal Reasoning Benchmarks in Azerbaijani language

Development of LLMs capable of efficiently handling Azerbaijani language is an area of study that is still in its early stages. It is for this reason that significant effort was exerted into fact checking and development of causal reasoning datasets similar to those employed for the testing in English. This is a significant step as availability of such datasets for the future work into LLMs in this language will allow for fast-paced development of capable decision-making models.

In retrospect, the above stated objectives aim to push the boundaries of what is currently possible with smaller LLMs in causal reasoning. By fine-tuning small-scale LLMs on specialized causality datasets and applying quantization techniques, this research creates a new paradigm where smaller, more efficient models do not merely mimic but are also well performant. Considering that causal reasoning tasks were thought to be a domain of larger LLMs. This shift has highly crucial implications for the deployment and integration of small-scale LLMs in everyday applications, making sophisticated AI tools more accessible and practical for a broader set of domains and industries.

Overall, successful achievement of the objectives mentioned above significantly contributes to the field of AI by demonstrating that size and scale can be decoupled from not only from performance, but also functionality in terms of causal reasoning. Moreover, with these points proven, wide adoption of a fleet of small-scale LLMs that offer different perspectives into a case based on the governing causal dependencies is just a few papers away.

It should also be considered that wide deployment of small-scale LLMs would not only fulfill a technical need but also pave the way for more sustainable and ethical AI deployment. Minimizing the environmental and economic costs associated with large-scale computational processes is an added advantage of the outcomes of this study.

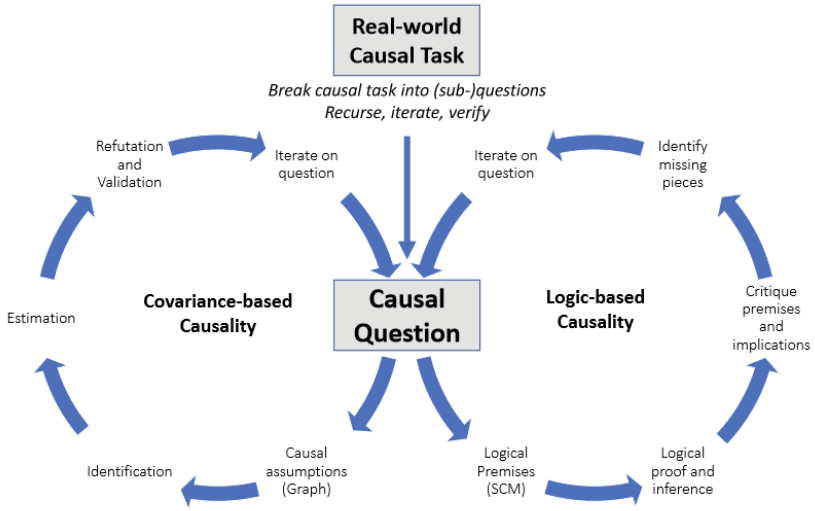

Figure 1: An overview of human causal reasoning process.

2 Literature Review

In Artificial Intelligence, as far as LLMs are considered, contextual reasoning refers to the ability of the model to comprehend, interpret and answer follow-up queries based on the underlying logical context of the data it processes. This is different as compared to basic reasoning where one might only consider direct inputs, contextual reasoning involves integrating several components including, but not limited to, situational awareness, relational understanding and world knowledge to generate more accurate and contextually relevant predictions. Causality, in its essence, is the concept of a sequence of pivotal events leading to a certain outcome and not the others given the variables conditions under which the said events took place [**?**]. To this extent, an AI model capable of causal reasoning is a model that can identify this sequence, understand the logical sequence and tie this understanding to the emergence of the outcome event.

As has been depicted in Figure 1, human causal reasoning, as far as neuroscience and social sciences are concerned, involves many steps that require strategically alternating between and iterating over different modes of reasoning such as, logical and covering-based reasoning developed as a result of a lifetime of experiences. Verifying important assumptions, designing and testing sub-questions and recalibrating the currently pursued approach is an integral part of this process [**?**]

2.1 Overview of LLMs and Their Evolution

LLMs like the GPT, BERT, Llama and their successors have caused a fundamental shift in the landscape of NLP, demonstrating remarkable capabilities in generating coherent text and understanding complex linguistic structures. Having been trained on extensive corpuses of text data from various sources in countless disciplines and domains, LLMs developed a broad understanding of the human language and context. As they evolve, LLMs increasingly handle more sophisticated tasks beyond mere text generation, venturing into areas that require nuanced understanding and reasoning.

As it is lucid from above, language modelling has generally seen a significant improvement in the last decade. The first attempts at Neural Language Models (NLM)

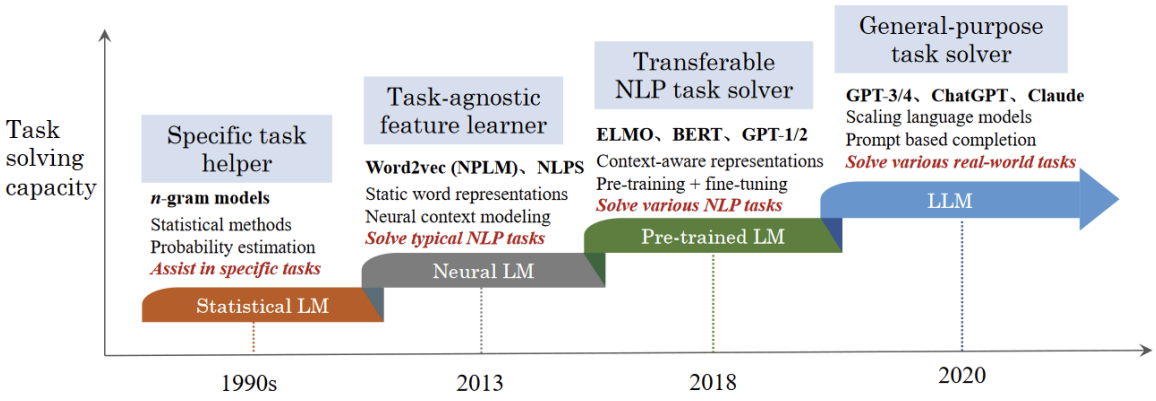

Figure 2: The evolutionary process of Language Modelling and NLMs over the last 3 decades categorized based on the task solving capabilities.

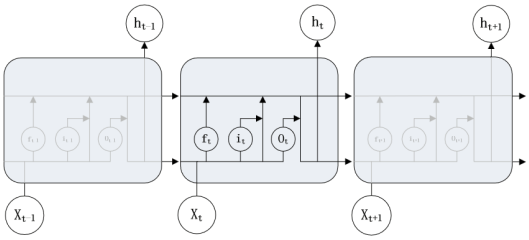

Figure 3: LSTM Cell architecture

were mostly composed of multi-layer perceptron (MLP) and recurrent neural net- works (RNNs). Later introduction of Long Short-Term Memory concept into RNNs revolutionized the NLM domain by making a wide range of applications possible. This relied on the fact that a generally well-trained base model made of an LSTM architecture can act as the fundamental knowledge block for specific down-stream NLM tasks [?] [?] [?]. Figure 2 showcases the development in the domain of Language Modelling from the perspective of the domain of application they were intended for

Despite the progress made with bidirectional LSTMs (Figure 3), there were still issues regarding the generalizability of the models. Despite being trained on large corpuses of data, longer sequence generation stood as a challenge.

2.1.1 The Transformer

A new frontier was brought into the NLM space when Google published the first ever transformer architecture based language model in its paper titled "Attention Is All You Need" [?]. In this paper, two revolutionary concepts, namely the scaled dot-product attention and multi-head attention.

Scaled Dot-Product Attention, is performed by calculating the dot products of the original queries, obtained by performing a token mapping of the input, with keys, and adjusting the scale of the result by the square root of the dimension. This is followed by applying a softmax function to obtain the final weights for the values. This vectorized process is efficient as it employs packs the queries, keys, and values into matrices Q, K, and V respectively. This, in turn, allows for a swift calculation of the output matrix.

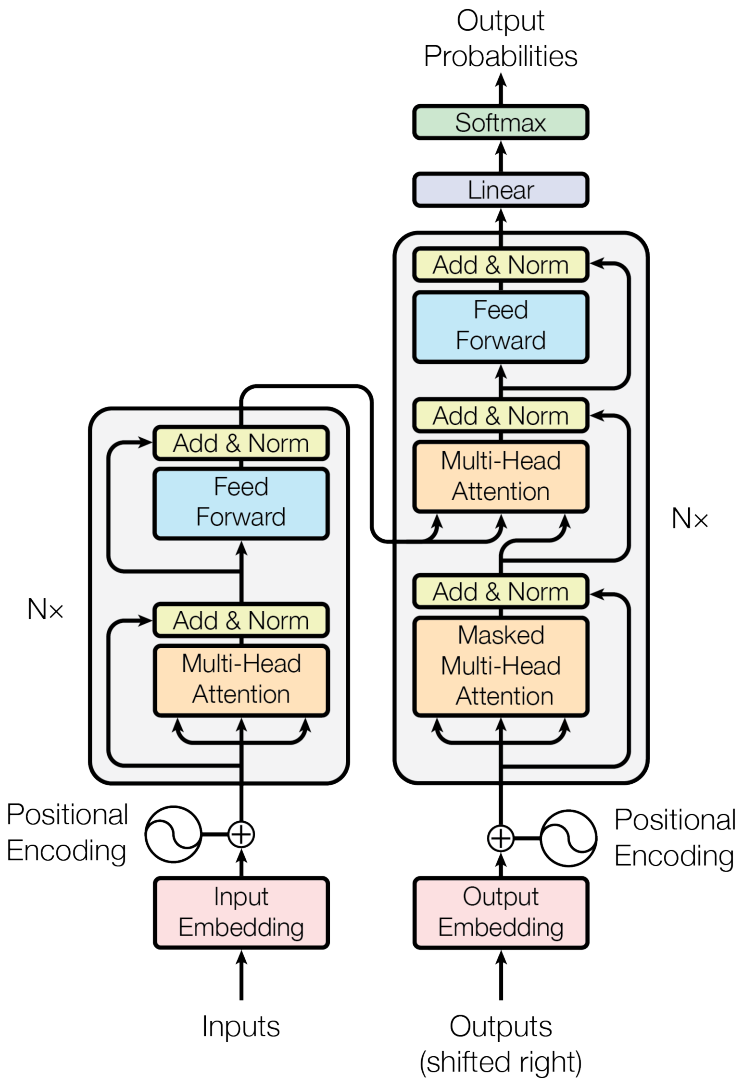

Figure 4: The Transformer - model architecture as depicted in [**?**]

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

This method, coupled with multi-head attention, resulted in state-of-the-art performance on many metrics on Language Modelling tasks such as English Language understanding and English to French translation [**?**]. It should also be mentioned that the training costs associated with this methodology was at a fraction of the other SOT techniques.

2.2 Role of LLMs in Complex Reasoning

LLMs were demonstrated to be pivotal in complex reasoning tasks that require understanding and manipulating language-based information. In diverse applications such as summarization, translation, and even answering sophisticated causal and factual questions LLMs are employed due to their versatility. However, in order for LLMs to accurately perform these tasks there are important considerations to be made such as the depth of their training and the specificity of the prompt engineering techniques utilized in the

process. These factors are crucial when attempting to handle certain types of reasoning, such as deductive, inductive, and causal reasoning [**?**].

Despite the ongoing efforts into the matter, it is necessary to keep in mind that there are various types of reasoning tasks that we, as humans, perform in our day-to-day tasks. Thus, a monotonous analysis of LLMs can easily be misleading, and this analysis should be conducted in a more rigorous and detailed manner. As explained in [**?**], various modes of reasoning require specialized approaches.

2.3 Types of Reasoning in AI

2.3.1 Deductive Reasoning

Deductive reasoning is fundamental to logical inference in AI and works off the assumption that if something is true of an entire class of objects, it must also be true for all members of that class [**?**]. While this application of AI is widely used in situations where the user knows a priori that the premises are true, including formal verification and legal code reasoning to guarantee a strict logical standard, this approach is vulnerable to brittle from uncertain or incomplete data. Indeed, logical rules can lead to false systems of reasoning if they are not strictly robust.

2.3.2 Inductive Reasoning

This aspect of reasoning is all about drawing the widest general rules from the narrowest particular observations. It underpins most machine learning algorithms perfect for building large models to make predictions from vast data sets. It has also applications in everything from financial forecasting to health diagnostics. However, inductive reasoning is only as good as the data and experience it is built upon [**?**]. Ultimately, this risks peril is that the training scenarios will deviate significantly from real-world applications, severely limiting the model's ability to transfer. There is also an ever-present risk of overfitting, mitigated through increasingly sophisticated regularization techniques.

2.3.3 Abductive Reasoning

Also known as inferential reasoning, it attempts to generate the most likely explanation for particular observed data. Abductive reasoning is of importance in sectors such as medical diagnosis or system troubleshooting, which uses it to deduce likely causes from observed effects. However, this kind of reasoning faces the critical challenge of data reliability and comprehensiveness. Poor-quality data may lead the AI systems to make biased or misleading hypotheses; thus, comprehensive data handling and preprocessing approaches are essential.

2.3.4 Counterfactual Reasoning

Involves reasoning about the outcomes of counterfactuals of possible alternatives to the actual history of the world. Counterfactual reasoning is crucial for causal inference, since it allows a decision-maker to judge the impact of an action by simulating under counterfactual scenarios. Implementing this reasoning in AI systems will require a more sophisticated understanding of causal relationships, as well as statistical techniques capable of estimating the causal effects of possible interventions. Counterfactual reasoning is also

critical for AI systems that plan and make decisions under hypothetical scenarios. Building AI systems around counterfactuals will thus enable applications to be used in strategic planning and policy analysis.

2.3.5 Causal Reasoning

Another important type of reasoning in AI is causal reasoning. In particular, causal reasoning refers to the ability of AI systems to distinguish and comprehend the relationships between cause and effect . Many AI applications in such spheres as economics, healthcare, and social science research depend on causal reasoning as well. What distinguishes this type of reasoning from deductive and probabilistic reasoning is that causal reasoning draws on the notion of causality between one or several causal variables and other effect variables . While statistical methods are sufficient for identifying such relationships, they only help to estimate the true causal effect without actually understanding why things happen. It is also often the case that identifying one's cause of interest comes across a variety of confounds or possibly confusing variables that appear when stating the relationship between the variables.

2.3.6 Limitations in Handling Various Types of Reasoning

Current Large Language Models have proven to be quite proficient in handling structured reasoning tasks falling under their training data. However, dynamic forms of reasoning requiring either the ability to conduct counterfactual reasoning or scenarios that need profound causal understanding, such as high-stakes decision-making are still a considerable challenge for these models. As to LLMs, including GPT-4, these models have been shown to present a vastly increased capability to generate plausible content while still struggling with reasoning tasks requires a great deal of nuance in causal understanding. These tasks also need to possess a level of demand for precision, as in the case of high-stakes decision-making where the correct answer has an essential relevance and for reasoning over the underlying mechanistic level of information [**?**].

While progress regarding the state of the art demonstrates an increase of accuracy with GPT-4, there remains a significant gap in causal reasoning, the ability of the model to generate correct inferences being strictly dependent if the causal factor has been part of the training data. If the models fall short of data from which to rely on the causality, the LLMs will often provide the answer based on the correlational evidence, instead of generating the answer based on true causality [**?**]. The current training regime premised on the estimations of probing formalisms and correlational data does not develop a model possessing a true facility for causal inference. It should also be noted that suitability of the current paradigms for training eliminating confounders needs to be discussed.

The performance of LLMs is largely dependent on the quality and quality of the training data. This becomes problematic when the data does not capture real-world variation or is not deep enough for reasonable learning. As a result, the empirical models can misconstrue or miss out on critical details that can lead to inaccuracies, particularly in high-stakes applications. It is also evident that existing evaluation protocols for reasoning tasks are not extensive enough to capture the complexities of real-world decision making [**?**]. As such, there is an overestimation of the capacity of these models to positively drive reasonable decision-making. Existing benchmarks and evaluation protocols do not capture the depth and flexibility of reasoning required in field applications, including the ability to handle multifaceted causal queries and provide correct, but more importantly,

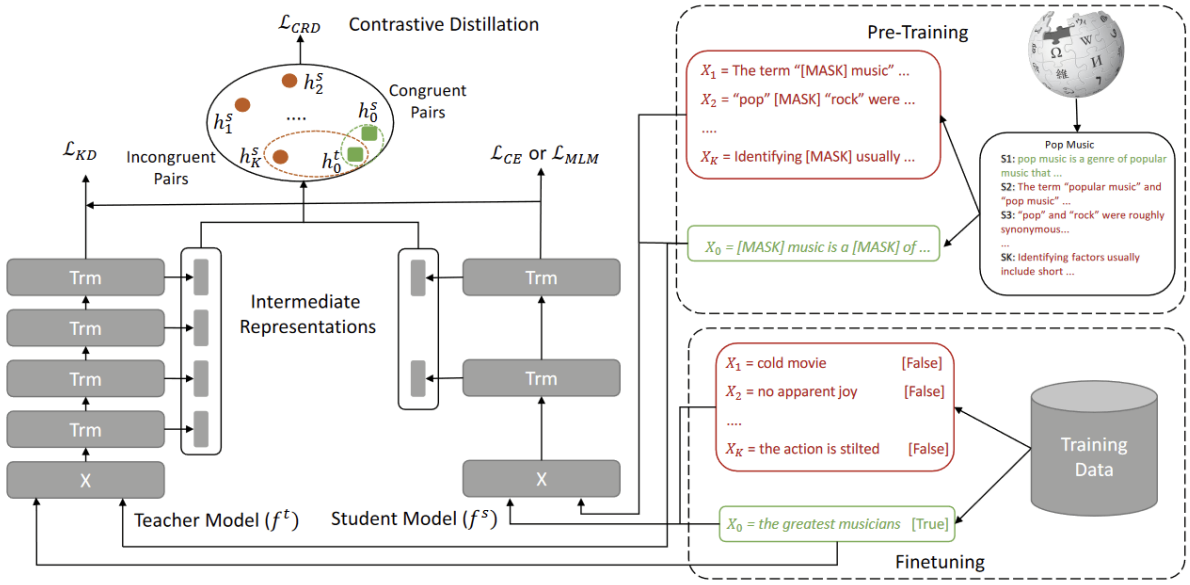

Figure 5: Illustration of CoDIR framework. Depicts the compression workflows for language models in pretraining and finetuning stages [**?**].

relevant responses. It is on this backdrop that we evaluate the capacity of some of the new language models to perform zero-shot reasoning when using a chain of thought prompting.

One of the studies, [**?**], established that LLMs, such as InstructGPT and PaLMs, can effectively perform zero-shot reasoning. The researchers used a single prompt to participants to encourage them to think through problems before answering while requiring no task-specific exemplars. They found that CoT prompting led to remarkable improvements across several classical reasoning tasks, such as arithmetic and symbolic reasoning. For InstructGPT, for instance, accuracy on MultiArith increased from 17.7% to 78.7%, and for GSM8K, it moved from 10.4% to 40.7%. Overall, the research established that LLMs have an untapped zero-shot reasoning capacity which can be exploited through the chain of thought chatting. As a result, users can now train these models without the need to generate fine-tuned datasets or feed them all possibilities during the few-shot exercise.

2.4 Challenges in Model Training and Adaptation

The nature of the reasoning task, unfortunately, is not the only prohibiting factor, scale and computational resource heaviness is still a bottleneck. The growing demand for the use of artificial intelligence in resource-constrained mobile devices and embedded systems has lead to the need of development of small-scale efficient models. These models provide a significant outcome for the amount of computation spent while reducing energy consumption and enabling their wider use especially for real-time processing [**?**].

Furthermore, a number of model optimization techniques that overcome the disadvantage of high memory utilization by the most performant LLMs have been developed and researched.

2.4.1 Pruning

Pruning is a useful method for improving the efficient of the model by removing weights in case of neural networks or connections contributing the least to output's accuracy. There are various methods among which are magnitude-based pruning that removes a certain number of weights with the smallest absolute values and structured pruning that deletes entire channels or filter according to a certain criterion. These both methods have already been used in practice and demonstrated certain advantages as they decrease the computational time during testing but do not distort the neural networks model's output or accuracy significantly. Furthermore, since pruning reduces the models' size it results in a better generalization as a method to avoid overfitting. Pruned networks with subsequent fine-tuning also show competitive results or even outperform non-pruned networks. [**?**]

2.4.2 Quantization

This method is predominantly used to lower the use of high precision numerical values, such as from 32 bit floating-point representations to lower bit integer representations, such as 8-bits, or lower through scaling. Quantization reduces the model's memory footprint and helps in speeding up the computation processes by enabling the use of integer-only arithmetic during the model's inference [**?**]. Two of the common approaches to quantization are post-training quantization and quantization-aware training. The former applies the quantization process after a model has been trained. The application of quantization after the model is trained can reduce the implementation complexity. The latter, quantization-aware training, applies the quantization process to the models during training to minimize the impact of quantization on the model's accuracy. Research shows that quantized models can have high-accuracy rates if implemented within the correct boundaries. This makes the quantized models suitable for deployment to edge devices with low computation capabilities.

2.4.3 Knowledge Distillation

Knowledge distillation is a technique where you train a smaller "student" model to reproduce the behavior of a larger "teacher" model. The "student" model learns from both the hard targets that represent the true labels and the soft targets that are indicative of the probability outputs of the "teacher" model. These data signals from the "teacher" model are incorporated to convey a supplementary message or information about the data distribution that is not seen in the hard labels alone. This helps the already trained student model closely mimic the performance of the larger "teacher" model while also being smaller and much faster to perform. For example, Distillation, especially Contrastive Distillation on Intermediate Representations, has been one of the most effective methods of doubling smaller model performance in real-time applications while maintaining much of the complexity of the decision boundaries learned by the teacher. The mechanisms that transfer the knowledge differ from each implementation of the methodology. For instance, the implementations facilitating the transfer do not just concentrate on generated content and their probabilities, but on the intermediate representations, i.e. inner layer hyperplane mapping of the input and generated tokens. Figure 5 illustrates the Contrastive Distillation process. The reported results achieved in the study show that the model developed through application of the CoDIR methodology achieved far better performance on benchmarks like CoLA, SST-2, MRPC, etc than a 125 million parameter

RoBERTa model [?]. [?].

As evident, knowledge distillation process can achieve significant reduction in the architectural and memory footprints of LLMs with minimal loss. However, it should be noted that this process requires training and hyperparameter tuning on different architectures and is rather computationally costly. Our goal, in this paper, is to explore a much more easily accessible method that requires minimal to no training. Thus, our main focus was on the quantization techniques and their impact.

2.5 Datasets and Methods Used in Enhancing LLMs for Reasoning

The importance of expanding datasets for training Large Language Models cannot be overestimated since it is crucial for enhancing reasoning accuracy and model performance. To expand the datasets, there are specialized tools provided in a ToolQA [?] dataset as the standard dataset. In addition, the investigation and CMExam set consisting of medical data and other relational concepts help the LLMs to develop their abilities in the utilization of external information and reasoning on specific topics [?] data. However, these standard datasets generate growth in the quality parameters, and it is needed to expand the datasets due to the intervals of data accuracy and further, expanding the datasets.

It is evident that human-AI partnerships can diversify and expand the datasets that are necessary for accurately and objectively training LLMs. The application of logit suppression and label replacement are considered successful for boosting the accuracy and diversity of the data. As a result, the correct data distribution has increased the abilities to enhance the model, and it has been evident in class calibration and standard case study with SAT test data. The implemented logit suppression has shown a 48% improvement in model development and standard operating procedure .

Standard datasets, such as SQuAD, GLUE, and bAbI have been successful and standard-notching to test and investigate the reasoning of their model and performance. Standard aspects and cognitive tasks for the benchmark of reasoning ability of such models, include reading comprehension, natural language inference, and question answering standardized in a SQuAD dataset. It is obvious that the testing and evaluation of compliance through legitimate and related data have been successful in creating a suitable environment. This data was later used to evaluate the proficiency of the model in language understanding and logical reasoning.

Impacts of Expanding Dataset Diversity on Model Performance and Reasoning have been proven to benefit the down-stream project adoption. One of the most significant challenges in a training large language model is to expand the dataset diversity, as it has a direct and dramatic impact on the performance and reasoning accuracy of the LLMs. The process of exposing these models to a broader range of language use and reasoning complexity allows them to build more sophisticated generalization capabilities and avoid some of the most common pitfalls, such as the overfitting to specific linguistic patterns or biases encountered in a more homogenous and closed dataset. For instance, if the model has only been trained on a specific type of data and from the limited number of sources and authors, the understanding of the nuances of general or regional uses of language in writing would not apply for tasks like language translation or text personalization. At the same time, the exposure to specialized datasets and reasons can significantly improve the reasoning accuracy of large language models while making them capable of

processing more complex selection reasoning scenarios or understanding the current state of the dialogue to select the most relevant response. As a prominent example, one paper is providing the results of an experiment to expand the multilingual commonsense reasoning datasets by using LLMS. One study [**?**] augmented three very common datasets, XCOPA, XWinograd, and XStoryCloze.

Another paper builds on the idea of improving the reasoning performance of LLMs by introducing the input prompt diversity. The authors suggested the DIV-SE and IDIV-SE methods, which generate several variations of the prompt and ensemble the multiple generated responses. As the result, both of the proposed methods have significantly surpassed the performance of all existing baselines in several reasoning benchmarks while using the variants of the GPT-3.5 heuristic and two advanced LLMS, GPT-3.5 and GPT-4. This experiment has demonstrated the importance of prompt diversity in achieving more effective reasoning performance and building LLMs with the best generalization capabilities at the lowest possible cost.

2.5.1 Retrieval-Augmented Generation

Retrieval-augmented generation is a significant advancement in the use of external knowledge sources to extend the capabilities of large language models. This technique is based on the integration of the retrieval process into the generation of an LLM, thus, allowing the model to access external information dynamically during the generation process. This approach is vital for tasks or questions that require the knowledge of an expert. In the field of academic research or technical support, external information can be used in a database or encyclopedia. Furthermore, when using a dataset or even a sophisticated corpus during the generation process, the provided response is not only contextual but also current and correct. Google's RAG model is an example of such a system that surpasses other LLMs in tasks that require a strong knowledge base, adding appropriate information dynamically during runtime. Moreover, the ability to augment the reasoning process with external data makes the model more capable of performing even better and in more sectors. An example of this advancement is the work titled "Scaling Relationship on Learning Mathematical Reasoning" [**?**]. The research is based on datasets and examines the effect of pre-training loss, the number of supervised data, and the amount of augmented data on LLMs.

The research presents that pre-training loss is a more consistent predictor of when models perform better and that the number of parameters in a model is not as relevant. Additional log-linear relationships were observed when BERT-based models were refined by different quantities of supervised fine-tuning (SFT). It was found that models with a higher performance tend to take advantage of more supervised datasets, but their performance improvements are lower. To facilitate the enhancement of the performance of a model without the intervention of humans, the research suggests the Rejection sampling Fine-Tuning (RFT). In this approach, a supervised model is employed to generate correct reasoning paths, which are then collected as fine-tuning datasets. There is evidence of improved rejection sampling when more than one models are used to fine-tune or shape the rejection samples. The research also demonstrates that with a greater variety of distinct reasoning paths, RFT significantly improves mathematical reasoning in LLMs, particularly for less performant models.

On the other hand, the paper named "Evaluating the Effectiveness of Retrieval-Augmented Large Language Models in Scientific Document Reasoning" [**?**] focuses on the impact of the retrieval-augmented generation when it comes to the issue of LLMs

hallucinating during the generation of plausible but incorrect predictions when they lack a fact base to make estimations. These LLM designs include a non-parametric design to retrieve information from the dataset during training. This allows the model to come up with predictions that can be traced to some evidence data retrieved from the external knowledge base. The research assesses the effectiveness of such models in tasks of scientific document reasoning by tuning many variants of the models with similar science instructions. Results demonstrate that despite the use of scientific corpora as the pretraining data, the problem of justifications with fabricated or nonexistent evidence highlights the difficulty of eliminating the risk of evidence fabrication.

2.5.2 Development of Language Modelling Datasets for the Azerbaijani Language

Development and maintenance of datasets for the low resource languages has been a popular topic of research in the recent years. As the Azerbaijani language belongs to the same category as well, there have been many works that try to expand the body of available dataset for development of language models. As mentioned before, data quality is an important concern when it comes to development of LLMs and thus, it cannot be taken for granted that a model developed based on the automatically scraped online data, which is the case for a large portion of the available conversational datasets in Azerbaijani, will have reasoning capabilities.

There have been many works that have contributed to the development of the language modelling corpora for this language such as [**?**], [**?**] and [**?**]. However, Causal Reasoning remains a wide gap in this research domain. Thus, in this study, we have developed a 20000 sample dataset for LLM evaluation on reasoning tasks in this language

2.6 Analysis of Model Failures in Reasoning

The study of the points of failure of AI reasoning models is a critical aspect of improving their reliability and robustness. An enumeration of a significant portion of the literature reveals that the most common types of failures are due to training dataset bias, inability to generalize to new examples, and uncertainty in reasoning about edge cases or entirely novel situations. Understanding these failure points is not possible using existing methods, such as metrics of success, and requires a comprehensive study with such methodologies as stress tests and adversarial evaluations designed to test models in situations where they fail the most often.

Multiple common failure points of AI reasoning models have been identified. The first type of failure is data bias, meaning that the model inherits and often amplifies the biases in the dataset on which it was trained. For example, such a model may prioritize certain groups over others when making recommendations for a judicial sentence, or disproportionately fail to approve loans to some groups when deciding to approve lists of potential borrowers. In addition, models often struggle to make inferences if the data or the question to be answered is drawn from a significantly different distribution than the one that the model had seen before – it means that the model may fail to get the correct answer when faced with an example that actively requires abstract reasoning or transfer learning to solve. Methodologies for the systematic study of failures include adversarial testing, where the model is fed with results confusing for it, and evaluation of the assumptions of the model using interpretability tools to follow an AI system in making a decision to find why decisions are made and not other alternatives.

In a paper titled "Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks" [**?**] authors consider that intuitive psychology being a fundamental part of common-sense reasoning. The authors of the source paper criticize the work of LLMs on Theory-of-Mind tasks that are anticipated to measure a specific ability, including belief attribution. It is explained in the paper that although there are great successes in the area, many LLMs fail on ToM tasks and show "systematic fragility to very minor variations in tasks that all correctly adhere to the underlying design of ToM" [**?**]. For this reason, the authors state that the evaluation of such systems should be seen critically, and their consistent failure is more important than an average rate of success. It is an urgent problem, and to address it, the best practices require that for the training of a model such criteria as fairness and robustness should also be considered in addition to accuracy. The ideas that may result in the enhancement of the level of generalization and lower the problems of overfitting are domain adaptation, regularization, ensembles, and creating datasets of different types. Before deploying a model, it is also crucial to test it appropriately, and in the case of common-sense reasoning, this procedure may become continuous because, with time, it will be possible to learn more about a model's failures in new circumstances. Feedback loops are also desirable, and they may help in situations when a model learns on its mistakes, changing its parameters to become better.

In [**?**] the authors discuss the problem of self-consistency of LLMs when these systems perform multistep reasoning tasks. They assert that such Self-Consistency can be of two types: hypothetical and compositional. Hypothetical consistency is a condition when a model can predict accurately what its output would be in a different hypothetical context to guarantee it would be the same. Another definition is given to describe the "compositional consistency," and it is the term used to explain that a model's set of final outputs should be the same as performing those sub-steps removed and replaced with the model's outputs". The research shows that the rates for both types of consistency are poor for GPT-3 and GPT-4 models, implying that current LLM systems cannot effectively reason over a sequence of steps. Finally, the paper implies that methodologies to analyze these systems should include the assessment of self-consistency, and the latter should be measured to check whether a model copes with complex reasoning tasks.

The issue of "Assessment Misalignment" has been addressed in another study [**?**] on fine-tuned LLMs with the typical growth of scores of the subpar reasoning processes. For this purpose, an Alignment Fine-Tuning process is proposed, which focuses on fine-tuning with chain of thought data, generating two responses, and calibrating scores. The effectiveness of the AFT process has been evaluated by its performance on a number of benchmarks ones (Wang, et al., 2023).

Another study explored integration of free-text explanations in the in-context provided learning for LLMs to enhance the performance of the training of smaller models to promote their cost-effectiveness for real-production deployment. The results of the study are impressive for much smaller models to fine-tune with their explanations generated by the LLMs outperform large GPT3-like models [**?**].

In another, the results were fine-tuned with a novel DaSLaM process [**?**], which generated decompositions of the initial, relatively large, problem into simpler subproblems. To this end, a small LM was used to be the decomposition generator of a complicated task to be solved by a bigger LM solvers. The feasibility of the process has been illustrated by the competitive results across the reasoning disclosed datasets regarding the unlimited solver scale.

The Fine-tune-CoT process [**?**] has been used in still another to make large LM Rea-

soning teachers capable of the training of smaller ones to master the complex reasoning. Fine-tune-CoT is based on the generation of the reasoning samples for the big teacher models to learn them to the smaller ones to boost the substantial improvements in the capabilities of the reasoning ones.

Yet another study employes the two-stage fine-tuning strategy, which has been developed based on the Maniskill2 benchmark to enhance the capability of the generalization of the model. The fine-tuning performance is impressive for its effective enhancement of the generalization ability for the practical uses.

3 Research Methodology

This research project that aims to assess the performance of LLMs in counterfactual reasoning, which is a fundamental aspect of causal understanding. The following sections briefly articulate the specific goals.

3.1 Evaluate the Performance of State-of-the-art small-scale LLMs on Established Eausal Reasoning Benchmarks

The first benchmark is the CRASS dataset, which is specially designed for counterfactual reasoning. The second benchmark is the Tubingen dataset, which is designed to test cause-and-effect predictions. In this task we will use a range of LLM architectures including gpt-4-1106-preview and gpt-4-0125-preview, to test the effect of model architecture and training data on the ability of the model to perform causal reasoning.

3.2 Investigate the impact of prompt engineering and instruction fine-tuning on LLM causal reasoning

In this analysis, our interest will revolve around determining how to create prompts that facilitate better causal reasoning for LLMs. In particular, we aim to determine whether fine-tuning techniques like LORA and QLORA can increase the causative reasoning capabilities of LLMs. To accomplish this feat, we will determine prompts that explicitly convey the nature of the task at hand and encourage LLMs to infer alternative scenarios from the established state. To complete this objective, we will analyze how the introduction of specific phrases in the statements provided to LLMs, such as "What would have happened if...?" or "How would...be different if...?" may affect their capacity for causative reasoning. In other words, we will seek to establish which prompt configurations will lead to stronger causative reasoning abilities in LLMs. This process may be described as the adjustment of internal LLM parameters for the destination problem. The performance of the resulting fine-tuned LLMs will be compared with their non-fine-tuned counterparts in the case of the benchmark assignments associated with the causative reasoning tasks. We have also conducted tests concerning quantized sample LLMs Mistral-7B-int8 and Gemma-7b-it-int8, as well as the effects of instructions and their size and performance. We will analyze whether a similar relationship applies to the achievable level of performance for the cause-and-effect tasks.

3.3 Dataset Augmentation and Development of Causal Reasoning Benchmarks in Azerbaijani language

We have extended the evaluation dataset and provided the opportunity for real-world application of our instruction fine-tuning. The new dataset focuses on diagnosing neuropathic pain and provides examples where knowing the causing ailment was the only way to solve the diagnostic issue. By testing LLMs performance on this dataset, we can conclude how much our findings and patterns are generalizable to a different domain and the random generating of causal relationships. Moreover, if the fine-tuning is indeed effective in training the models to reason about causes and their corresponding effects accurately, we would publish the causal reasoning dataset in Azerbaijani for the interested researchers to delve into. The creation of this dataset would involve translating up-to-date causal reasoning benchmarks to Azerbaijani or, if unavailable, creating the new ones and making sure they are linguistically and culturally appropriate. It would serve as a stepping stone for developing Azerbaijani causal reasoning LLMs and empower researchers and developers to explore the applications in this under-represented language area. For instance, developers may create educational applications, healthcare applications, or even use LLMs to make scientific discoveries among the Azerbaijani-speaking.

Having the ability to reason causally is one of the forms of basic reasoning for intelligent systems. For example, knowing the cause and effect can result in forming predictions, recognizing problems and offering effective determinants. Counterfactual reasoning is one of the most difficult forms of such reasoning and refers to the thinking about alternatives to the events of the past. As LLMs become more and more flexible and continue obtaining access to the knowledge of the world, it is crucial to evaluate their level of counterfactual reasoning. being a perfect basis for this evaluation, the CRASS dataset is a valuable benchmark for estimating LLMs' performance. Through the identification of the weak aspects of these mechanisms, it is possible to conclude on their potential for success in human-like reasoning and justify the development of an effective program to enhance these algorithms and use them in the creation of powerful AI.

3.4 Description of the CRASS (Counterfactual Reasoning Assessment) Dataset

CRASS is a deliberately structured dataset that allows for testing large language models at counterfactual reasoning. Unlike many other causality datasets, it has a comparatively narrow focus on this single question while stretching the limits of questions too unrelated to simple fact recall. Each scenario includes the following components:

- **Premise:** the situation or event where the counterfactual reasoning will be built. In other words, a premise is a behind-the-counterfactual alternation made.
- **Counterfactual Question:** the central part of a scenario, always starting with "What would have happened". Then, the given alternation is brought to the premise; the counterfactual reasoning is expected.
- **Correct Answer:** the most plausible resulting premise LLM would reach with counterfactual reasoning. The "best" option is determined by crowdsourcing and is usually a rather logical outcome.

- **Options (A, B, C, D):** four possible resulting premises created by modifications. Large language models need to evaluate the plausibility of each.
- **Correct Answer:** the most plausible resulting premise LLM would reach with counterfactual reasoning. The "best" option is determined by crowdsourcing and is usually a rather logical outcome.

3.4.1 Qualitative analysis of the dataset

One of the reasons why the CRASS dataset was chosen is that while many other datasets happen to serve the purpose of causal reasoning, CRASS makes it an exclusive and narrow focus. Models are tested on their capability of suggesting "ifs" to statements and causes related outcomes. This process not only establishes the result of certain events in the past (strictly causal reasoning) but allows to hypothetically model future situations.

Another noteworthy factor is that the majority of examples are neither under described nor specifically complicated. Both the premise and the counterfactual question are simple to understand which lets LLM concentrate on reasoning instead of understanding. The provided options are unambiguous and implicitly related to the premise, making it more challenging for the LLM to actually pinpoint the right answer since non of the choices are out of topic. Moreover, the general veracity of correct answers to questions is realistic and is within the human reasoning capabilities.

CRASS allows for testing of LLMs in an identical manner, which results in a series of favorable conditions to emerge. One of such benefits is ease of utility of quantifying the counterfactual reasoning capabilities. The dataset allows for application of to grade reasoning capacities of large language models employed. Identification of reasoning skewing is another advantage in that closer analysis of responses on both correct and incorrect categorizations would enable us to grasp the LLMs' tendencies when reasoning. Furthermore, LLM analysis would be beneficial on a grand scale. Understanding of where and why LLMs fail on CRASS creation would be able to help in the creation of causal models. Such models could handle a variety of complex and uncertain situations.

3.5 Description of the Tübingen Benchmark

The Tübingen Benchmark is an invaluable resource for one vital test of causality in LLMs. Essentially, the dataset's structure is elegantly simple while being thoroughly challenging. This dataset, contrary to CRASS, requires a scientific background in the decision-making process to get to the correct answer. Each point of data is composed of a pair of physical phenomena:

- **Cause Pairs** This data point is a pair logical of occurrences one of which is the cause with the other being the effect.
- **Label:** each of the pairs of is accompanied by a label that indicates which item in the pair is the cause

It should be noted that the pairs are correctly depicting a possible causal link, according to common knowledge and scientific reasoning, and the LLM is expected to find out this link.

Fundamentally, the purpose of the Tübingen Benchmark is to establish whether an LLM can indeed understand positive and negative causality from natural language. At

Premise	Counterfactual Question	Options
A woman opens a treasure chest.	What would have happened if the woman had not opened the treasure chest?	The treasure chest would have remained closed.; The treasure chest would have been open.; That is not possible.
A police officer calms down a hostage-taker.	What would have happened if the police officer had not calmed the hostage-taker?	The hostages would have remained in danger.; That is not possible.; The hostage-taker would have released the hostages anyway.
A man talks about a lion.	What would have happened if the man had talked to the lion?	Without a barrier, the man would have been eaten.; Without a barrier, the lion would have been eaten.; Nothing special would have happened.
A girl kisses a boy.	What would have happened if the girl had slapped the boy?	The girl would have been angry.; The girl would have been happy.; That is not possible.
A girl kisses a boy.	What would have happened if the girl had killed the boy?	She would have been liable to prosecution.; That is not possible.; The boy would have been arrested for assault.; The boy would have kissed the girl.
A woman sees a fire.	What would have happened if the woman had touched the fire?	She would have been burned.; She would not have been burned.; That is not possible.; She would have seen fire.
A woman sees a fire.	What would have happened if the woman had fed the fire?	The fire would have become larger.; The fire would have been not hungry anymore.; That is not possible.; The fire would have been on the woman's shoulder.
A branch floats in a pond.	What would have happened if a corpse had floated in the pond?	The lake would have been a crime scene.; That is not possible.; The corpse would have swum back to the barge.; The corpse would have been shot.
A branch floats in a pond.	What would have happened if a branch had sung in the pond?	That is not possible.; It would have been the usual branch singing.; It would have been a great concert.; It would have got stuck.
A bird lands in a forest.	What would have happened if a plane had landed in the forest?	The plane would have crashed.; Everything would have been fine.; The plane would have landed safe and sound.; In a forest you will find lots of planes.

Table 1: Samples from the CRASS counterfactual reasoning dataset. In this scenario, the LLM is expected to select the contextually correct answer based on the causal relationship in the premise.

this point, it is important to clarify what is meant by understand. Indeed, it is possible for an LLM to learn that in a certain percentage of cases, the appearance of a certain phrase might mark an effect or an appearance of another phrase, linked to an action, identify it as a "cause". However, this kind of simple word-association pattern will not suffice. Specifically, the LLM has to perform several functions simultaneously. One of those functions is comprehension of the semantics. The LLM must properly grasp the meaning of the inputted sentence, and comprehend what happens in them, what key ideas are present, and how those items related to one another. Another aspect is world knowledge and scientific reasoning, where the LLM must apply general knowledge and have an understanding of simple physical phenomena, that we, as humans, intuitively understand, to determine whether one entity is a possible cause for another. In Table **??** a sample raw data is provided from the Tubingen dataset.

Complementary to CRASS, Tubingen benchmark also involves counterfactual reasoning. In CRASS, a model must predict the likely outcome of a change in correlation between two observed facts for the presence or the absence of a causal link between them. In contrast, the Tübingen Benchmark isolates causal reasoning by asking whether a given statement is likely to be causal, providing a more focused inquiry about the ability of LLMs to parse the context and reason about a relationship implied in the text.

Moreover, the text from many existing datasets for assessing causality makes the answers easily discoverable by simple means of the correlation calculation of word pair occurrences by modern LLMs. In addition, many questions that can be asked about the likely relationship between X and Y can be solved using simple patterns, which is not the

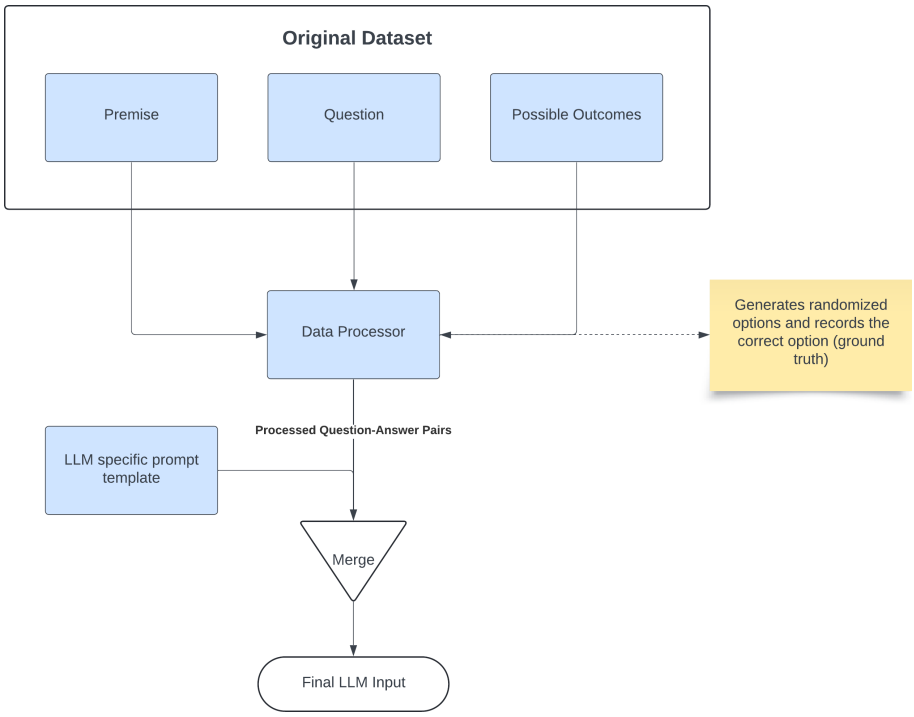

Figure 6: Data processing workflow for the CRASS and Tubingen datasets

goal of this study. The goal is for LLM to try to reason beyond the superficial cues and completely model the logical relationships of causality. If such a skill can be isolated, the extent to which specific models succeed in developing LLMs to build expansive views of causality and disentanglement systems can be measured. The test of such a skill is passed by passing the Tübingen Benchmark.

Success with the Tübingen Benchmark provides insight into a model's ability to distinguish whether a relationship is causal now, as written, and whether a change in one of the factors would likely have an impact on the change in the other. Both skills are essential for the development of versatile AIs that can operate in conditions where the relationship between various factors is essential for planning, prediction, or explanation tasks.

3.6 Data Processing and Prompt Engineering

The design of these prompts relied on a common workflow, an illustration of which is given in Figure 6, which allowed for consistency and provided space for specific nuances related to each respective dataset. The template centered around a few key elements:

- **System Message:** The role and task of the LLM is established through the system message; e.g., "You are a helpful researcher designed to work with data.". This was provided as a way of priming the LLM to answer in a specific manner and to ensure that the contextual understanding is the focus when generating the predictions. The decision to include this specific system prompt was a result of extensive efforts in prompt engineering, testing and analysis on GPT4.

- **Premise:** The situation or context of the case is established for the LLM, to be analyzed by the model in relation to the cause. The premise was taken from the original dataset.

```
SYSTEM: You are a helpful assistant designed to evaluate counterfactual
reasoning.
USER: A police officer calms down a hostage-taker.  What would have
happened if the police officer had not calmed the hostage-taker?
A:  The hostage-taker would have released the hostages anyway.
B:  That is not possible.
C:  The hostages would have remained in danger.
D:  The police officer would have been injured.
Let's work this out in a step-by-step way to be sure that we have the
right answer. Then provide your final answer within the tags
<Answer>A/B/C/D</Answer>
```

Figure 7: Example Prompt generated for the CRASS dataset

```
SYSTEM:  You are a helpful assistant designed to identify causal
relationships.
USER: Which cause-and-effect relationship is more likely?
A: Changing the price of a product causes a change in the demand for
that product.
B: Changing the demand for a product causes a change in the price of
that product.
Let's work this out in a step-by-step way to be sure that we have the
right answer.  Then provide your final answer within the tags
<Answer>A/B</Answer>
```

Figure 8: Example Prompt generated for the Tubingen dataset

- **Question:** As explained before, here the counterfactual element of the task is indicated, explaining to the LLM the difference in consequences based on the content of the premise. For example, "What would have happened if..."
- **Instruction:** A command provided to the LLM to format its output in a specific manner. This approach not only makes it easier to post-process the responses, but also to evaluate the instruction following capabilities of the LLMs

3.7 Experimentation and Evaluation on LLMs

Analysis of the results was conducted included both quantitative and qualitative methodologies. In terms of the utilized numeric metrics, the focus was on deeply investigating the model's performance to gain insights into the decision making of the LLM. On the other hand, qualitative exploration of the resulting responses was conducted to determine whether the input prompt was successful at priming the LLM, and if the response was within the guidelines provided to the LLM as an instruction.

If the model generated a response that was irrelevant to the question or had a severely hard time following the instructions, this response was marked as "Undetermined" which finds its reflection in the results part as well.

One of the metrics used to calculate the performance of the LLMs was the accuracy score which is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Where,

- TP **and** TN - True Positives/Negatives, instances that actually belong to the predicted class.
- FP **and** FN - False Positives/Negatives, instances that belong to the opposite class.

Another measurement involved in this study was the Precision score, which denotes the fraction of the instances that were predicted to fall into a certain class that actually belong to that class. This metric is computed as below:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall, on the other hand, is a measure of how many out of the members of the class being scrutinized were successfully predicted:

$$Precision = \frac{TP}{TP + FN} \tag{4}$$

F1 score, or the harmonic mean of Precision and Recall scores, is a very convenient metric that enables computation of the balance between the two. Ideally, we would like the said scores to be the same and closer to 1. From this perspective, the higher F1 score is indicative, although not directly attributable, of a better model.

$$Precision = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Another measurement used in this paper is the Confusion Matrix, which is essentially a convenient way of showcasing the TP, TN, FP and FN values, and determining if model confuses any of the classes with the others.

3.8 Dataset Augmentation and Instruction Fine-tuning

One of the motivations behind this research was to examine whether smaller LLMs could be trained to demonstrate stronger causal reasoning capabilities in the presence of large training data such as CRASS and the Tübingen Benchmark. Dataset augmentation and instruction fine-tuning were applied for a number of reasons that help achieve this goal.Firstly, a point of concern is data scarcity. Causal reasoning datasets appear to be underrepresented in comparison to datasets for language modeling in general used in training of LLMs. Thus, augmentation helped expand the original CRASS and the Tübingen Benchmark datasets, thus potentially exposing LLMs to a broader variety of cause-effect structures.

Improving the instruction following capabilities of small-scale LLMs was another vital point. Fine-tuning the instructions had a purpose of providing smaller LLMs with more explicit and descriptive instructions that could steer them towards causal reasoning. To this end, the input was manipulated so that the tasks were framed through the use of counterfactual states and scenarios.

Furthermore, the possible advantages resource made possible by utilizing small-scale LLMs is of high importance to this study. Large LLMs perform exceptionally well, but we were interested in exploring the potential of smaller models. Dataset augmentation as well as instruction fine-tuning were performed with smaller models that require less computation to help them grapple with real-world causal reasoning tasks.

The steps performed to augment already available English language datasets and develop an analogous dataset for the Azerbaijani language are described in the following paragraphs.

3.8.1 CRASS and Tübingen Augmentation

We used the GPT-4 model to generate additional examples and prompts for both datasets. All examples and prompts were reviewed so that the data kept its original counterfactual and cause-effect structures. A separate cause prompt structure was then chosen to create the instruction fine-tuning dataset. The augmented dataset is exemplified in 2.

3.8.2 Translation into the Azerbaijani language

The original CRASS and Tübingen datasets and the created augmented datasets were translated into Azerbaijani using Google's MBERT model. This work is not a simple translation – by developing an Azerbaijani dataset, we have created a resource that future LLMs designed for Azerbaijani or similar languages can be trained on and tested. Currently, only a very small number of very large LLMs with high levels of performance – like GPT-4 – are able to process the Azerbaijani language. Hefty effort was exerted into post-processing of this dataset to ensure logical, structural and grammatical quality.

A few examples from the instruction fine-tuning dataset are provided in Tables **??** and 3

Input	**Output**	**Prompt**
A man forgets to water his plants for a week. What would have happened if the man had remembered to water his plants? A: The plants would have died. B: The plants would have grown better. C: The plants would remain the same.	<Answer>B</Answer>	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Based on the context of the following question, identify the correct cause-effect relationship. Provide your final answer within the tags, <Answer>A/B/C/D</Answer>. ### Input: A man forgets to water his plants for a week. What would have happened if the man had remembered to water his plants? A: The plants would have died. B: The plants would have grown better. C: The plants would remain the same. ### Response:<Answer>B</Answer>

Table 2: A sample from augmented CRASS dataset transformed to be compatible with instruction fine-tuning techniques

3.9 LLM Fine-tuning

As a result of the previous steps, we had access to about 5000 samples of augmented instruction fine-tuning data for both datasets. We proceeded with fine-tuning Google's gemma-7b-it model on this dataset using the widely adopted transformers library. Training was continued for 3 epochs, which is the recommended number of epochs for fine-tuning

English Prompt	Azerbaijani Prompt
SYSTEM: The following is a cause-and-effect relationship. USER: Which cause-and-effect relationship is more likely? A. changing the altitude causes a change in temperature. B. changing the temperature causes a change in altitude. Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.	SİSTEM: Aşağıdakılar səbəb-nəticə məntiqi əlaqəsinin bir nümunəsidir. İSTİFADəÇİ: Hansı səbəb-nəticə məntiqi ardıcıllığı daha böyük ehtimalla mümkündür? A. hündürlüyün dəyişməsi temperaturun dəyişməsinəsəbəb olur. B. temperaturun dəyişməsi hündürlüyün dəyişməsinəsəbəb olur. Düzgün cavabı əldə etmək üçün suala mərhələli şəkildə cavabla. Son cavabını <Answer>A/B</Answer> formatına uyğun şəkildə təqdim et.
SYSTEM: The following is a cause-and-effect relationship. USER: Which cause-and-effect relationship is more likely? A. changing the altitude causes a change in precipitation. B. changing the precipitation causes a change in altitude. Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.	SİSTEM: Aşağıdakılar səbəb-nəticə məntiqi əlaqəsinin bir nümunəsidir. İSTİFADəÇİ: Hansı səbəb-nəticə məntiqi ardıcıllığı daha böyük ehtimalla mümkündür? A. hündürlüyün dəyişməsi yağıntının dəyişməsinəsəbəb olur. B. yağıntının dəyişməsi hündürlüyün dəyişməsinəsəbəb olur. Düzgün cavabı əldə etmək üçün suala mərhələli şəkildə cavabla. Son cavabını <Answer>A/B</Answer> formatına uyğun şəkildə təqdim et.
SYSTEM: The following is a cause-and-effect relationship. USER: Which cause-and-effect relationship is more likely? A. changing the longitude causes a change in temperature. B. changing the temperature causes a change in longitude. Let's work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>A/B</Answer>.	SİSTEM: Aşağıdakılar səbəb-nəticə məntiqi əlaqəsinin bir nümunəsidir. İSTİFADəÇİ: Hansı səbəb-nəticə məntiqi ardıcıllığı daha böyük ehtimalla mümkündür? A. uzunluq dairəsinin dəyişməsi temperaturun dəyişməsinəsəbəb olur. B. temperaturun dəyişməsi uzunluq dairəsinin dəyişməsinəsəbəb olur. Düzgün cavabı əldə etmək üçün suala mərhələli şəkildə cavabla. Son cavabını <Answer>A/B</Answer> formatına uyğun şəkildə təqdim et.

Table 3: Sample from translated Azerbaijani Tubingen dataset

small-scale LLMs, and we used Cosine Annealing Learning Scheduler, Eq.7, proposed in [**?**], and the next token based Cross Entropy Loss, Eq.6.

$$L(Y, P) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \tag{6}$$

where:

- $L(y, p)$ - loss function.
- N - number of classes.
- M - number of samples.
- y_j - true label of the j-th sample (i).
- p_j - predicted probability of the j-th sample being of class i.

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})(1 + \cos(\frac{T_{cur}}{T_{\max}}\pi)) \tag{7}$$

where:

- η_t - learning rate at epoch t.
- $\eta_{\min}$ - minimum learning rate.

- $\eta_{\max}$ - maximum learning rate.
- T_{cur} - current epoch number.
- $T_{\max}$ - maximum number of epochs.

article algorithm algpseudocode

The training cycles were conducted using Low-Rank Adaptation (LORA), which is method of applying matrix decomposition, namely Singular Value Decomposition on the weight matrices. In this study, we applied LORA, [**?**] on all linear layers in the architecture. This includes the linear layers at the embedding layers and the decoder layer. WandB service was utilized for tracking and logging of the training sessions.

Low-Rank Adaptation (LoRA) adapts pre-trained models by inserting trainable low-rank matrices. This allowed for the fine-tuning on full-precision, without any quantization, to take place on a GPU system equipped with 2 Nvidia V100 units each of which had 16GB of RAM with the batch size of 32. The same training run was successfully reproduced on a 24 GB consumer grade Nvidia RTX 3090 GPU with the batch size of 16 as.

$$W' = W + AB^T \tag{8}$$

$$y = \phi(xW') \tag{9}$$

where:

- W - original pre-trained weight matrix.
- W' - adapted weight matrix.
- A and B - trainable low-rank matrices of rank r.
- AB^T - low-rank update to the weight matrix W.
- x - input to the layer.
- y - output of the layer.
- ϕ - activation function.

Figures 9 and 10 showcase the training and validation losses respectively.

4 Results and Discussion

As mentioned in the methodology section of this paper, models were tested on different datasets with varying levels of quantization and diverse architectures. The following section addresses the results obtained as a result of testing on baseline models without any fine-tuning performed. It should be noted that some of the LLMs used in this study are not open-source and were accessed using the public APIs developed by for-profit companies that own the models. Hence, models like GPT-4 do not provide a quantization end-point.

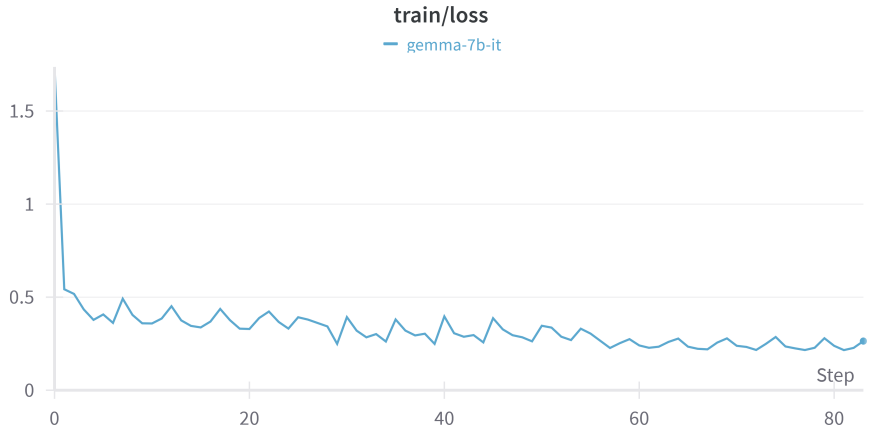

Figure 9: Training loss during instruction fine-tuning of gemma-7b-it. Only the best run is displayed

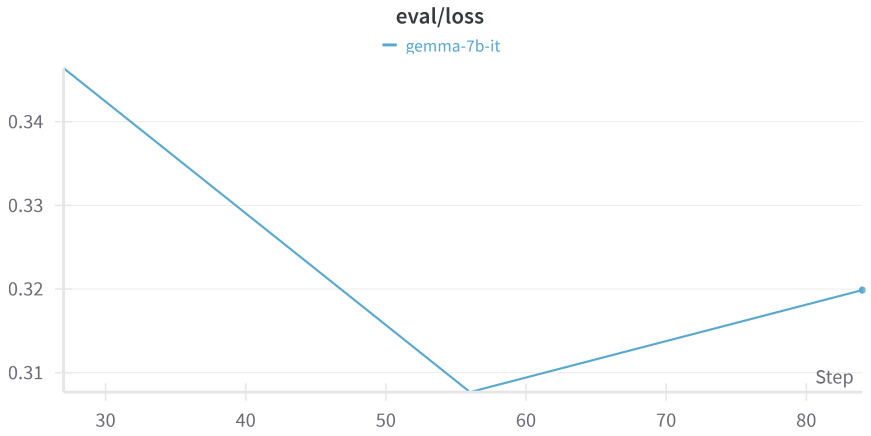

Figure 10: Validation loss during instruction fine-tuning of gemma-7b-it. Only the best run is displayed

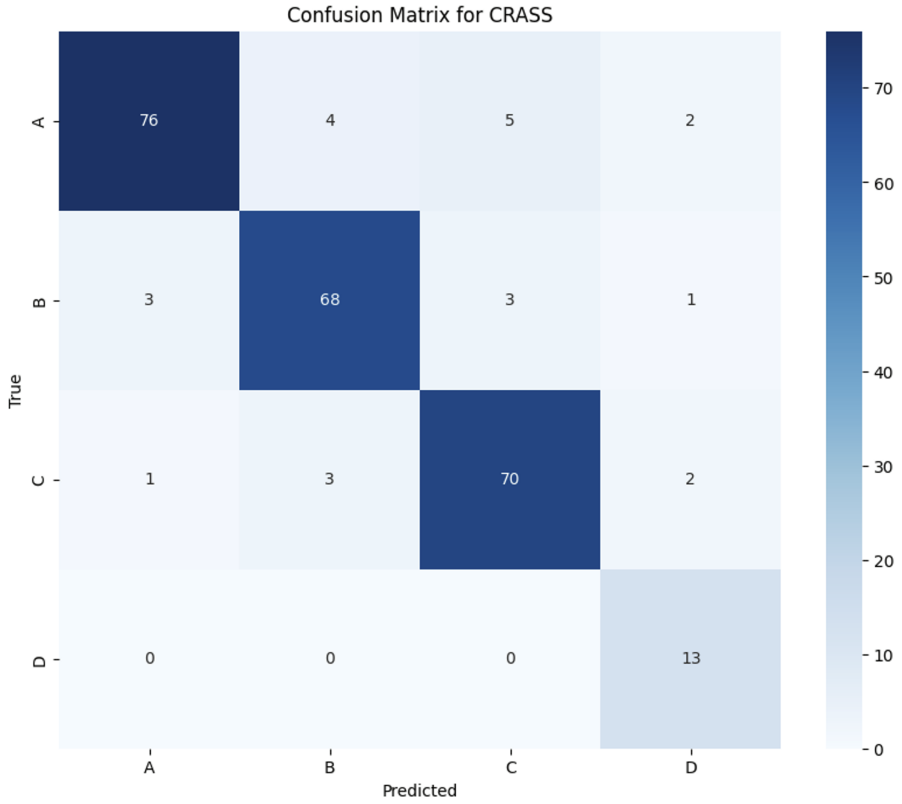

Figure 11: Confusion Matrix for gpt-4-0125-preview on CRASS dataset

4.1 LLM Performances on Causal Reeasoning

Codenamed gpt-4-0125-preview, the model yielded promising results with an overall accuracy of 90% on the CRASS dataset 4.1, with a similar performance of 89% on the Tubingen dataset. It should be noted that, as mentioned previously, the original datasets were not balanced in terms of the distribution of the correct answers on the options due to the randomness introduced to the process. The results are better depicted below in the following summary.

Considering that no previous context was provided for any LLMs when prompting for an answer over the test samples, it was surprising to observe that gpt-4-0125-preview was less likely to select the last option provided in the prompt as its F1 score on the CRASS dataset drops significantly when the correct answer is in option D compared to all other options. The steep decline in performance, from 91% for options A, B and C to 84% for option D is perplexing and could potentially point to model's inability to balance out the attention on a contextual basis for the later parts of the text.

The same tendency was observed over the Tubingen dataset with a sharper decline in the observed F-1 score from 92% for option A to about 82% for option B. This, coupled with the fact that due to the unavailability of previous user and model interactions in the context it is not feasible for the LLM to prefer an earlier posed option due to its previous response, reinforces the statement made above. This is a rather interesting aspect to research as to why the model fails to pay proper attention when there are many options available.

When it comes to smaller models, like Gemma-2b-it, Gemma-7b-it, and Mistral-7b-it, we are faced with another dilemma.

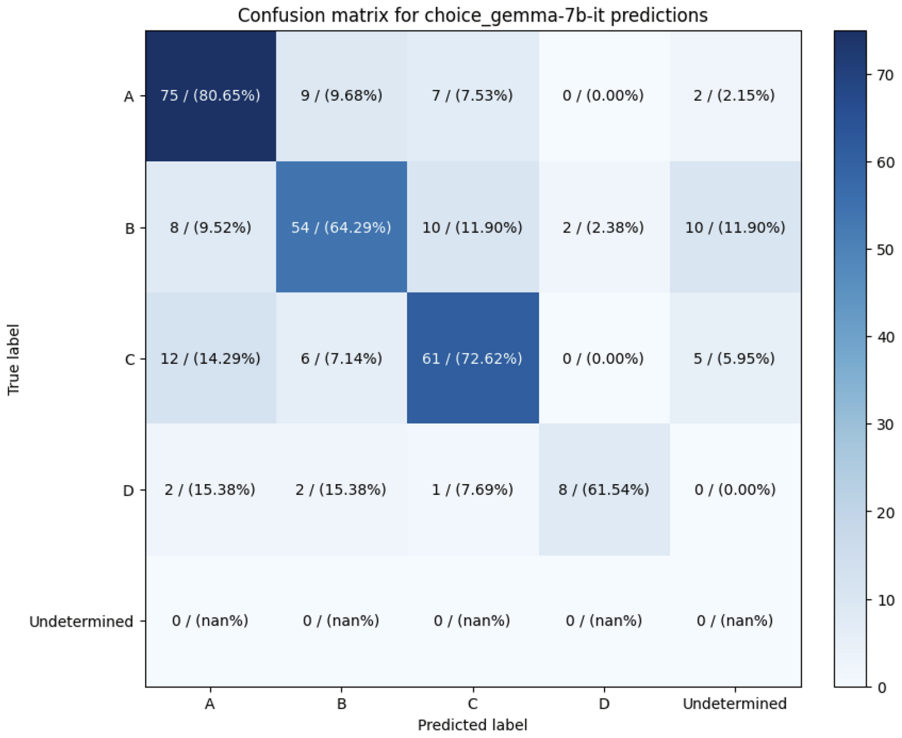

Figure 12: Confusion Matrix for gemma-7b-it on CRASS dataset

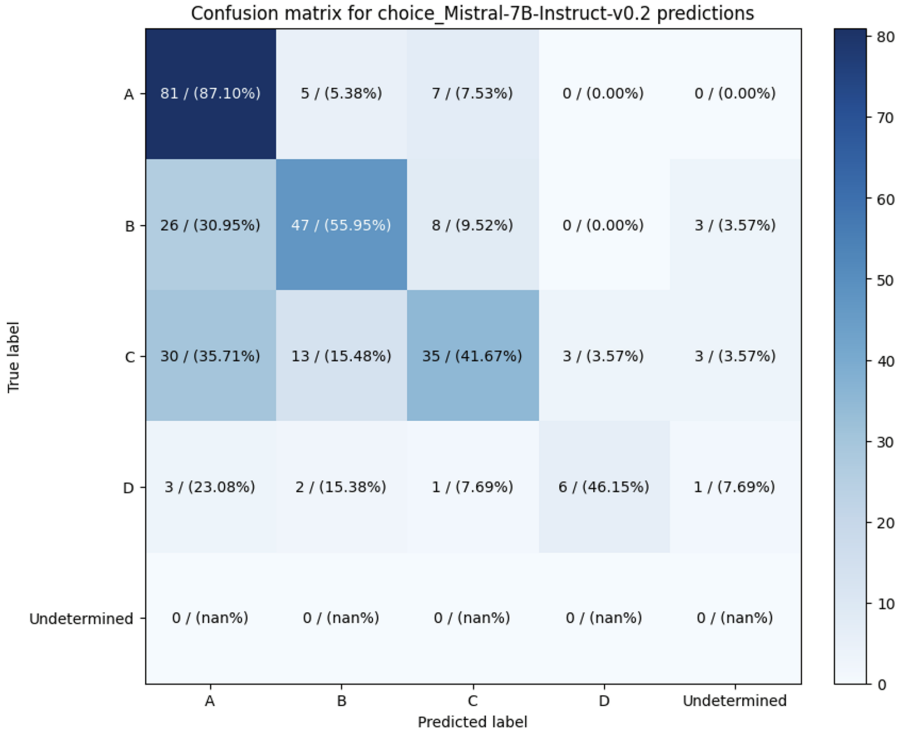

Figure 13: Confusion Matrix for Mistral-7B-Instruct-v0.2 on CRASS dataset

Here, as can be seen from Tables 4.1 and 4.1, we have another category of results, labelled as "undetermined". This label denotes a general subset of model responses where the LLM either failed to follow the instructions and provide a favorable answer or had severe hallucinations that did not match the posed question. For a better understanding, Figure 13 presents the confusion matrix for the labeled responses of gemma-7b-it over the CRASS dataset. As we can see, especially for the cases where the correct answer lies in option B, the model hallucinated quite often. The same phenomenon was observed over the results obtained from Mistral-7B-Instruct-v0.2 as well.

An identical phenomenon was observed when analyzing the results of the predictions on the Tubingen dataset as well. As depicted in Figures 15 and 16, there were errors made by the LLMs when following the instructions and providing the necessary outcome.

One unexpected aspect of the previously discussed results is that in the CRASS dataset Google's gemma outperformed Mistral-7b which is a Llama2 based model, by an astounding 10% difference, accuracy scores of 72% compared to 62% in the specified order. However, the reverse was true for the Tubingen benchmark. On this specific dataset, Mistral-7B-Instruct-v0.2 was significantly better than gemma-7b-it with an accuracy of 79% and 53% respectively. In fact, gemma's results are barely better than a random guess. This further illustrates the difference between the LLMs in terms of their capabilities and the training data used. CRASS and Tubingen benchmarks are fundamentally different in that the former relies on mostly commonsense reasoning abilities for counterfactual causal reasoning while the latter relies on scientific knowledge.

Metric	**F-1 scores (%)**				**Overall Accuracy (%)**	**Number of Parameters (10^9)**
Model	**true label A**	**true label B**	**true label C**	**true label D**		
gpt-4-0125-preview	**91**	**91**	**91**	**84**	**90**	1760
gemma-7b-it	79	70	75	70	72	8.54
Mistral-7B-Instruct-v0.2	70	62	52	55	62	**7.24**
gemma-7b-it (fine-tuned)	**87**	**84**	**85**	**83**	**84**	8.54

Table 4: Comparison of the results of LLMs tested in this study on the Crass dataset based on the per category F1 score, and the overall Accuracy

The above findings let us conclude the presence of a significant influence of the training data over the causal reasoning capabilities of LLMs for smaller LLMs. This fact is indicative of the fact that small scale LLMs require more specialized data to understand the underlying patterns in data and perform well enough under varying circumstances.

Regarding whether a similar issue in terms of the attention mechanisms of these LLMs, favoring the initial options over their contextual relevance to the question, is present in the smaller models, this still seems to be a common pitfall, albeit less evident. We can see that for Mistral-7B-Instruct and gemma-7b-it models there was a steep drop of 20% and 13% respectively for their accuracies on the Tubingen dataset. On the CRASS dataset,

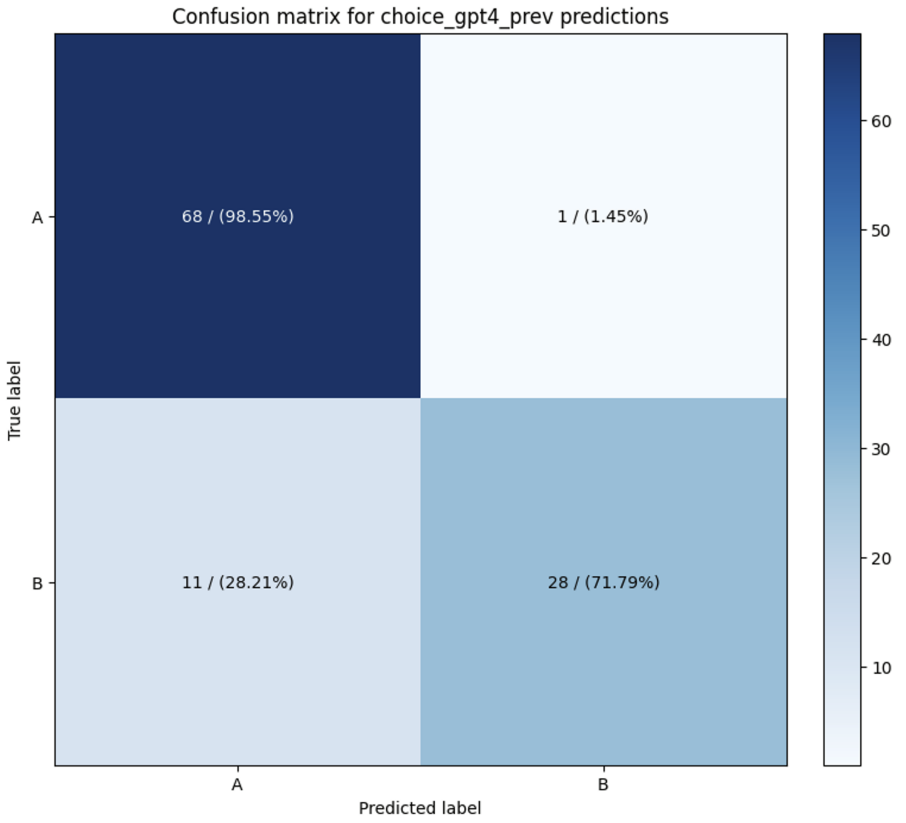

Figure 14: Confusion Matrix for gpt-4-0125-preview on Tubingen dataset

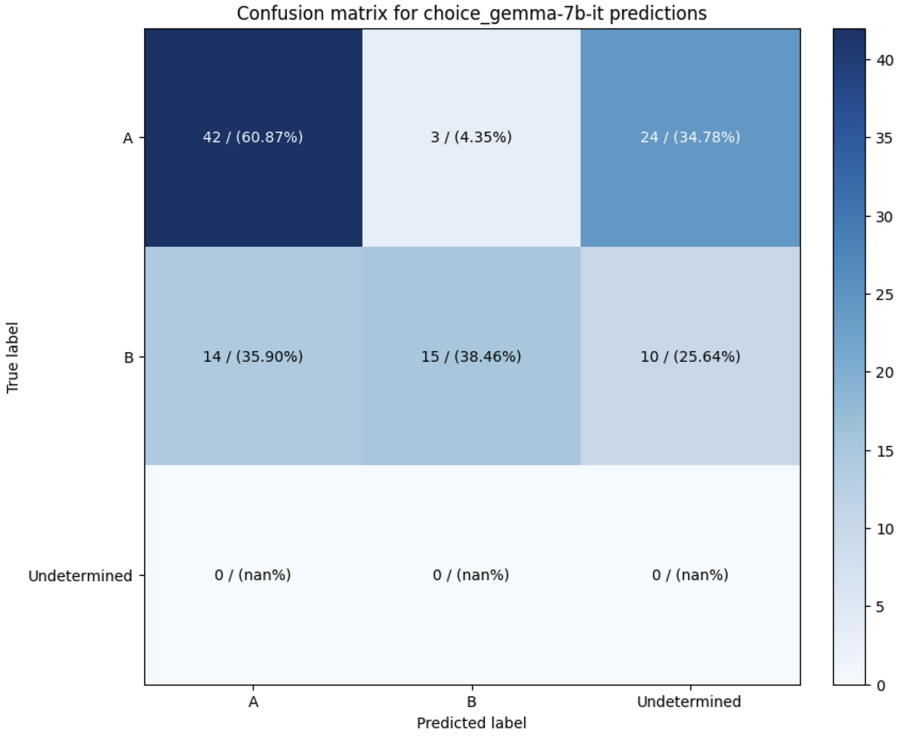

Figure 15: Confusion Matrix for gemma-7b-it on Tubingen dataset

Metric	**F-1 scores (%)**		**Overall Acuracy (%)**	**Number of Parameters (10^9)**
Model	**true label A**	**true label B**		
gpt-4-0125-preview	**92**	**82**	**89**	1760
gemma-7b-it	67	53	53	8.54
Mistral-7B-Instruct-v0.2	86	66	79	**7.24**
gemma-7b-it (fine-tuned)	**88**	**84**	**85**	8.54

Table 5: Comparison of the results of LLMs tested in this study on the Tubingen dataset based on the per category F1 score, and the overall Accuracy

however, this is less self-evident and cannot be directly attributed to the underlying attention mechanisms used on the transformer architectures.

4.1.1 Scale vs Causal Reasoning Performance

As can be observed in Tables 4.1 and 4.1 there was a significant boost in the performance of the gemma-7b-it model on both datasets as a result of instruction fine-tuning. It should be noted that both the original and the fine-tuned models were tested on exactly the same test dataset. Moreover, this dataset was not present during fine-tuning so the model has not directly observed the test data. From these results alone, it is apparent that a small fine-tuned LLM has the potential of providing significant results in reasoning tasks.

This is further evidenced by the fact the performance gains were very steep. In the CRASS dataset, the fine-tuned gemma-7b-it model showcased an overall accuracy gain of whopping 12% from 72% to 84%. As for the Tubingen dataset, the increase was even more surprising as the accuracy skyrocketed from the previous 53% to 85% inching closer to the performance of gpt-4-0125-preview. Fine-tuned gemma-7b-it even surpassed Mistral-7B-Instruct-v0.2 by 6 points where it had fallen short by 26% in comparison before fine-tuning. It is also noteworthy that the discrepancy between the rates at which the initial options were chosen by the model compared to the latter options has been reduced to a certain extent.

Another point of extreme importance is the scale of these models. As can be seen, gpt-4-0125-preview is about more than 200 times larger than both of the small-scale LLMs with over 1.7 trillion parameters compared to 7.24 and 8.54 billion parameters in the case of Mistral-7B-Instruct-v0.2 and gemma-7b-it respectively. This goes to show that scale is not everything and that task specific performance does not have to come at the expense of significant computational costs and an infrastructural overhaul.

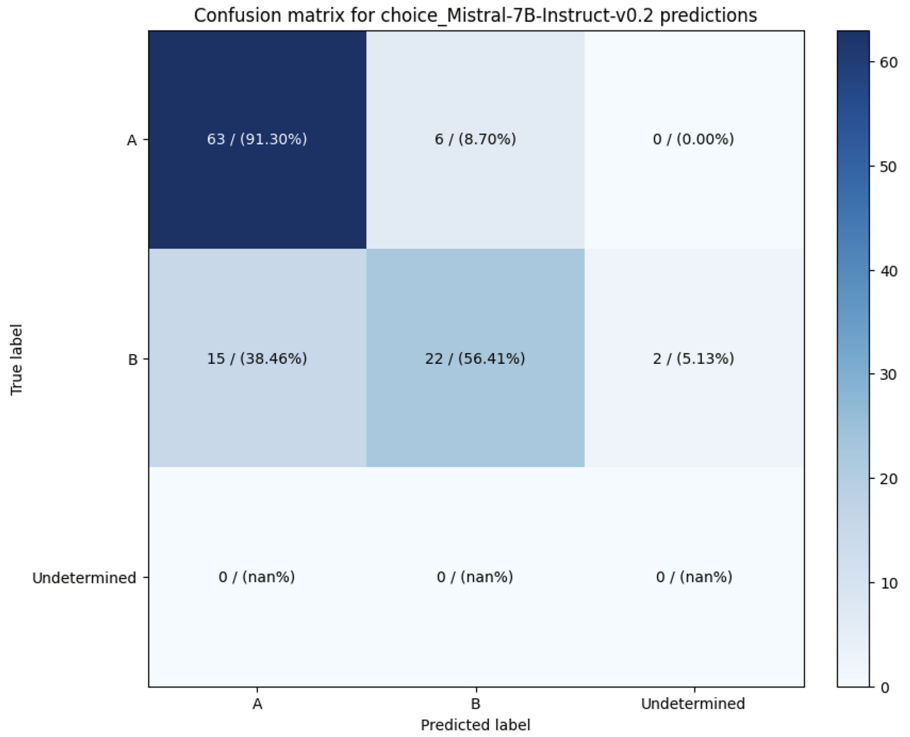

Figure 16: Confusion Matrix for Mistral-7B-Instruct-v0.2 on Tubingen dataset

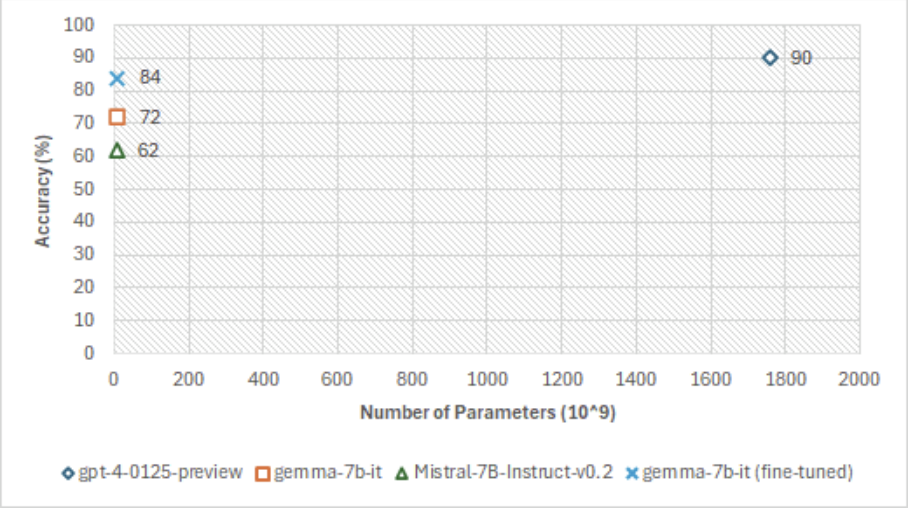

Figure 17: Scale vs Accuracy of the tested LLMs on the CRASS Dataset

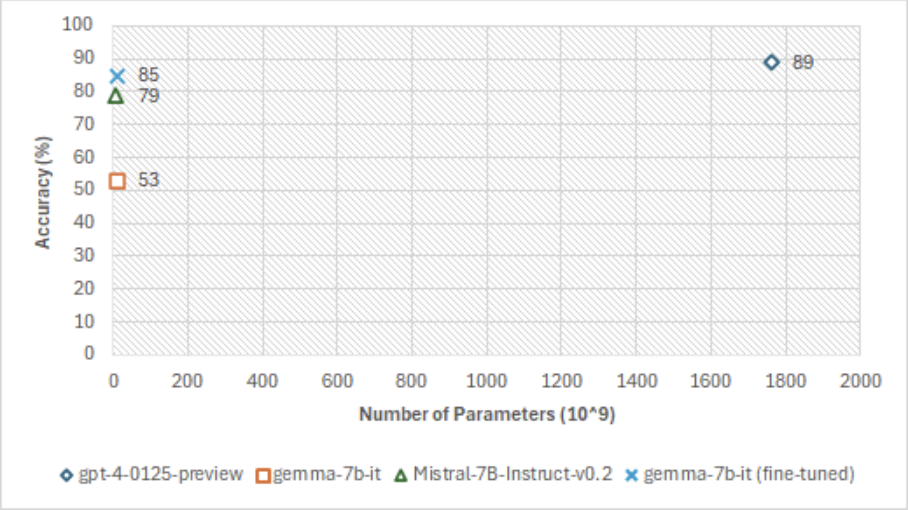

Figure 18: Scale vs Accuracy of the tested LLMs on the Tubingen Dataset

5 Conclusion and Future Work

5.1 Conclusion

In this thesis, small-scale Large Language Models were investigated with respect to their utility in causal reasoning tasks, with dataset augmentation and instruction fine-tuning as the primary contributions into the body of already available research. The CRASS and Tübingen datasets were initially augmented with GPT-4 by integrating a broader range of cause-effect relationships and counterfactual reasoning examples. An essential aspect of the study was the creation of equivalent versions of the datasets in the Azerbaijani language using Google's MultiLingual BERT and manually developed changes. This work addressed a language modelling related gap for the Azerbaijani language and laid the foundation for future models specifically for this language.

The instruction fine-tuning step used LoRA to optimize the gemma-7b-it model and minimize the computational intensity. Moreover, the use of Cosine Annealing Learning Scheduler, Cross Entropy Loss, and advanced NVIDIA GPUs ensured that the fine-tuning process was of high-quality and captured the required details. The training cycles were conducted on 2 data center grade high bandwidth Nvidia V100 GPUs with each at 16GB of RAM capacity and separately on a single consumer-grade Nvidia RTX 3090 with 24GB of RAM. Regarding the results, it is possible to note the extensive enhancements in the causal reasoning tasks observed on the fine-tuned model. Specifically, the fine-tuned gemma-7b-it model demonstrated 85% accuracy in the Tübingen dataset and 84% in the CRASS benchmark where we observed improvements of about 12 and 32 points, respectively, compared to the base model. These rates were quite close to those of the notably larger GPT-4 model with the latter model scoring 90% accuracy on the CRASS dataset while having 89% acuracy on the Tubingen dataset. Therefore, it is possible to conclude that with datasets of high quality and the precise instruction, small-scale LLMs might be used effectively in various fields. They can bring significant benefits in scalability and efficiency, especially when quantization is applied, supporting numerous areas requiring complex problem-solving and extensive decision-making capacities.

5.2 Future Work

To further push the capabilities and expand the practical impact of small-scale LLMs in causal reasoning and other AI tasks, a few crucial directions for future research are presented. First, a broader spectrum of languages and dialects could be included in the dataset augmentation process to enable causal reasoning capable small-scale LLMs in resource limited languages as well.

Secondly, even though an extensive exploration of optimization procedures for LLM training was conducted, further attempts at more efficient or more advanced methodologies that move beyond the capabilities of LoRA and QLoRA, for instance, could improve the training speed and model efficiency further. It would also be important to look into the effectiveness of other training parameters such as the impact of the loss function for model fine-tuning. This, unfortunately, was not possible during the duration of this study due to computational resource limitations

Moreover, deploying small-scale LLMs into real-life scenarios would be the next essential research direction. In pilot studies with healthcare, finance, or legal services, the data might be collected for the model development and fine-tuning purposes, where the focus is on context perception and reasoning within human-AI interaction settings on problem-

solving tasks. This would ensure availability of a feedback loop and allow collection of the data on small-scale LLMs' performance under real use-case scenario.

Creation of multidisciplinary collaborative networks of researchers interested in development of small-scale LLMs is another point that needs to be considered as a future direction in this domain of research. Such a network would help combine efforts in overcoming large and complex challenges, for example, by creating and sharing datasets of sufficient size and quality that pertain to different reasoning tasks and tackle the issues with causal reasoning in LLMs more efficiently.

Thus, in future studies, the tested small-scale LLMs and their hypothetical variations built upon the existing research, should be studied in larger, more complex projects to enable a continuous advancment of this type of small, efficient and widely adoptable models.

6 Appendix

Data Processing and Augmentation for Causal Reasoning Datasets

```
# Preprocessing of data
def preprocess_options(row):
    options = row[4:8].values
    options = [opt for opt in options if pd.notna(opt)]  # Remove any empty options
    correct = row[4]
    random.shuffle(options)  # Shuffle the options for randomness
    choices = "ABCD"
    options_dict = dict(zip(choices, options))
    correct_choice = [k for k, v in options_dict.items() if v == correct][0]
    return (*options_dict.values(), correct_choice)

# Generating fine-tuning data
for index, row in crass_df.iterrows():
    user_prompt = f"""SYSTEM: You are a helpful assistant for counterfactual reasoning.
    USER: {row["Premise"]} {row["QCC"]}
    A: {row["A"]}
    B: {row["B"]}
    C: {row["C"]}
    """
    if pd.notna(row["D"]):
        user_prompt += f"D: {row['D']}\n"
    user_prompt += """Let's work this out in a step by step way to be sure that we have the right answer.
                      Then provide your final answer within the tags, <Answer>A/B/C/D</Answer>.
                   """

    response = openai.Completion.create(
        engine="gpt-4-1106-preview", prompt=user_prompt, max_tokens=50
    )
```

Figure 19: Preprocessing and Fine-tuning data generation

Model Fine-tuning and Evaluation

```python
# Function to generate data augmentation prompts
def generate_augmented_data(prompt, model="gpt-4-1106-preview"):
    response = client.chat.completions.create(
        model=model,
        messages=[
            {"role": "system", "content": "You are a helpful assistant."},
            {"role": "user", "content": prompt}
        ]
    )
    return response.choices[0].message.content

# Example prompt and processing
example_prompt = "Who won the world series in 2020?"
print(generate_augmented_data(example_prompt))

# Process responses to extract structured data
def process_responses(response_text):
    question_pattern = r"\[Q\]([\s\S]*?)\[A\]"
    answer_pattern = r"\[A\]([\s\S]*?)(?:\[Q\]|$)"
    questions = re.findall(question_pattern, response_text)
    answers = re.findall(answer_pattern, response_text)
    return questions, answers
```

Figure 20: Prompt Engineering for Causal Reasoning

```python
# Define the dataset and model parameters
DATASET = "crass"
model_name = "toghrultahirov/gemma-1.1-7b-it-CausalFT-merged"

# Initialize tokenizer and model
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    add_lora=True,
    lora_alpha=16,
    lora_r=8
)

# Load dataset
df = pd.read_csv(os.path.join(data_path, DATASET, "results.csv"))
df.head()

# Fine-tuning setup
training_args = TrainingArguments(
    output_dir='./results',          # output directory
    num_train_epochs=3,              # number of training epochs
    per_device_train_batch_size=16,  # batch size for training
    per_device_eval_batch_size=64,   # batch size for evaluation
    warmup_steps=500,                # number of warmup steps for learning rate scheduler
    weight_decay=0.01,               # strength of weight decay
    logging_dir='./logs',            # directory for storing logs
    logging_steps=10,
)

# Define a simple training loop
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=df,  # your training dataset
)

# Start training
trainer.train()
```

Figure 21: Model Fine-tuning using LoRA