BANCA D'ITALIA

EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

Fine-tuning large language models for financial markets via ontological reasoning

by Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili and Emanuel Sallinger

January 2024

Number 44

BANCA D'ITALIA
EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

Fine-tuning large language models for financial markets via ontological reasoning

by Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili and Emanuel Sallinger

Number 44 – January 2024

The papers published in the 'Markets, Infrastructures, Payment Systems' series provide information and analysis on aspects regarding the institutional duties of the Bank of Italy in relation to the monitoring of financial markets and payment systems and the development and management of the corresponding infrastructures in order to foster a better understanding of these issues and stimulate discussion among institutions, economic actors and citizens.

The views expressed in the papers are those of the authors and do not necessarily reflect those of the Bank of Italy.

The series is available online at www.bancaditalia.it.

Printed copies can be requested from the Paolo Baffi Library: richieste.pubblicazioni@bancaditalia.it.

ISSN 2724-6418 (online)
ISSN 2724-640X (print)

Banca d'Italia
Via Nazionale, 91 - 00184 Rome - Italy
+39 06 47921

Designed and printing by the Printing and Publishing Division of the Bank of Italy

FINE-TUNING LARGE LANGUAGE MODELS FOR FINANCIAL MARKETS VIA ONTOLOGICAL REASONING

Teodoro Baldazzi[1], Luigi Bellomarini[2], Stefano Ceri[3], Andrea Colombo[3], Andrea Gentili[2], and Emanuel Sallinger[4,5]

Abstract

Large Language Models (LLMs) usually undergo a pre-training process on extensive collections of generic textual data, which are often publicly accessible. Pre-training enables LLMs to grasp language grammar, understand context, and convey a sense of common knowledge. Pre-training can be likened to machine learning training: the LLM is trained to predict the next basic text unit (e.g., a word or a sequence of words) based on the sequence of previously observed units. However, despite the impressive generalization and human-like interaction capabilities shown in Natural Language Processing (NLP) tasks, pre-trained LLMs exhibit significant limitations and provide poor accuracy when applied in specialized domains. Their main limitation stems from the fact that data used in generic pre-training often lacks knowledge related to the specific domain. To address these limitations, fine-tuning techniques are often employed to refine pre-trained models using domain-specific data. Factual information is extracted from company databases to create text collections for fine-tuning purposes. However, even in this case, results tend to be unsatisfactory in complex domains, such as financial markets and finance in general.

Examining the issue from a different perspective, the Knowledge Representation and Reasoning (KRR) community has focused on producing formalisms, methods, and systems for representing complex Enterprise Knowledge. In particular, Enterprise Knowledge Graphs (EKGs) can leverage a combination of factual information in databases and business knowledge specified in a compact and formal fashion. EKGs serve the purpose of answering specific domain queries through established techniques such as ontological reasoning. Domain knowledge is represented in symbolic forms, e.g., logic-based languages, and used to draw consequential conclusions from the available data. However, while EKGs are applied successfully in many financial scenarios, they lack flexibility, common sense and linguistic orientation, essential for NLP.

This paper proposes an approach aimed at enhancing the utility of LLMs for specific applications, such as those related to financial markets. The approach involves guiding the fine-tuning process of LLMs through ontological reasoning on EKGs. In particular, we exploit the Vadalog system and its language, a state-of-the-art automated reasoning framework, to synthesize an extensive fine-tuning corpus from a logical formalization of domain knowledge in an EKG. Our contribution consists of a technique called verbalization, which transforms the set of inferences determined by ontological reasoning into a corpus for fine-tuning. We present a complete software architecture that applies verbalization to four NLP tasks: question answering, i.e., providing accurate responses in a specific domain in good prose; explanation, i.e., systematically justifying the conclusions drawn; translation, i.e., converting domain specifications into logical formalization;

1 Università Roma Tre, Italy.

2 Banca d'Italia, Italy.

3 Politecnico di Milano, Italy.

4 TU Wien, Austria.

5 University of Oxford, United Kingdom.

and description, i.e., explaining formal specifications in prose. We apply the approach and our architecture in the context of financial markets, presenting a proof of concept that highlights their advantages.

Keywords: Ontological reasoning, Large language models, Knowledge graphs.

Sintesi

I Large Language Model (LLM) vengono comunemente sottoposti a un processo di pre-training su ampie raccolte di dati testuali generici, spesso accessibili pubblicamente. Il pre-training consente agli LLM di comprendere la grammatica della lingua, capire il contesto e trasmettere un senso di conoscenza comune. Il pre-training può essere paragonato al training del machine learning: l'LLM viene allenato per prevedere la prossima unità base di testo (ad es. una parola o una sequenza di parole) in base alla sequenza di unità precedentemente osservate. Tuttavia, nonostante le impressionanti capacità di generalizzazione e di interazione simil-umana mostrata nello svolgimento task di Natural Language Processing (NLP), gli LLM pre-allenati mostrano limitazioni significative e forniscono una scarsa accuratezza quando adottati in domini specializzati. Il loro principale limite deriva dal fatto che i dati usati nel pre-training generico spesso non incorporano la conoscenza relativa allo specifico dominio. Per affrontare queste limitazioni, vengono spesso adottate tecniche di fine-tuning: partendo da un modello pre-trained, si applica un raffinamento usando dati specifici del dominio. Le informazioni fattuali vengono estratte dai database aziendali al fine di creare raccolte di testi destinate al fine-tuning del modello. Tuttavia, anche in questo caso, i risultati tendono a essere insoddisfacenti in domini complessi, come i mercati finanziari e la finanza in generale.

Esaminando la questione da una prospettiva diversa, la comunità di Knowledge Representation and Reasoning (KRR) si è concentrata nel produrre formalismi, metodi e sistemi per la rappresentazione di una complessa Enterprise Knowledge. In particolare, gli Enterprise Knowledge Graphs (EKGs) possono sfruttare una combinazione delle informazioni fattuali presenti nei database e della conoscenza di business specificata in modo compatto e formale. Gli EKGs servono allo scopo di rispondere ad interrogazioni specifiche di un dominio attraverso tecniche consolidate come il ragionamento ontologico. La conoscenza di dominio è rappresentata in forme simboliche, ad esempio attraverso linguaggi basati sulla logica, e utilizzata per trarre conclusioni consequenziali dai dati disponibili. Tuttavia, se da una parte gli EKG trovano applicazioni di successo in molti scenari finanziari, dall'altra, essi mancano di flessibilità, senso comune e orientamento linguistico, essenziali per l'NLP.

Questo articolo propone un approccio mirato ad aumentare l'utilità degli LLM per applicazioni specifiche, come quelle legate ai mercati finanziari. L'approccio prevede di guidare il processo di fine-tuning degli LLM attraverso il ragionamento ontologico sugli EKG. In particolare, sfruttiamo il sistema Vadalog e il suo linguaggio, un framework di ragionamento automatico allo stato dell'arte, per sintetizzare un ampio corpus di fine-tuning a partire da una formalizzazione logica della conoscenza di dominio in un EKG. Il nostro contributo consiste in una tecnica, chiamata verbalizzazione, che trasforma l'insieme delle inferenze determinate dal ragionamento ontologico in un corpus per il fine-tuning. Presentiamo un'architettura software completa che applica la verbalizzazione a quattro task di NLP: question answering, cioè fornire risposte accurate in un determinato dominio in buona prosa; explanation, cioè giustificare sistematicamente le conclusioni tratte; translation, cioè convertire specifiche di dominio in una formalizzazione logica; description, cioè spiegare in prosa delle specifiche formali. Applichiamo l'approccio e la nostra architettura nel contesto dei mercati finanziari, presentando una proof of concept che ne mette in luce i vantaggi.

CONTENTS

 This work is based on the pre-print paper arXiv:2306.10723 [cs.CL]; a derived work has been published in Rules and Reasoning – 7th International Joint Conference, RuleML+RR 2023, Oslo, Norway, September 18-20, 2023, Proceedings – ISBN 978-3-031-45071-6 on Springer.com.

1 Introduction: Context and Overview of the Approach

With the recent soar of AI-based chatbots, currently led by OpenAI's ChatGPT, the field of Natural Language Processing (NLP) and, in particular, Large Language Models (LLMs), faced a major turning point and transcended its relevance in academia and industry, steering the attention of the general public towards generative AI. While many approaches are being proposed that exploit powerful pre-trained LLMs, such as T5 [18] and GPT [16], when addressing a plethora of applied tasks, current solutions show limited effectiveness at specializing the models on knowledge-intensive domains such as financial markets.

Leveraging the experience of the *Knowledge Representation and Reasoning* (KRR) community, such domain-specific knowledge can be captured by combining factual data from corporate databases with business definitions as ontologies in *Enterprise Knowledge Graphs* (EKGs), and further augmented via *ontological reasoning*. In this paper, we build upon this domain representation and propose a novel solution to accurately specialize LLMs on NLP tasks. Although we look at the financial markets context, our techniques are domain independent.

Limits of task-specific fine-tuning. LLMs can be pre-trained on extensive datasets and, often, specialized with a *fine-tuning* process that customizes them so as to perform given NLP tasks [20], such as *question-answering*, *language translation*, *named-entity recognition*, *document summarization*, *sentiment analysis*, and more [7]. According to a very common usage pattern, general-purpose LLMs are fine-tuned for a specific NLP task based on extensive cross- or domain-generic textual corpora that are publicly available [17].

While this approach highlights good generalization capabilities and a surprising human-style interaction, the obtained models have major shortcomings in that they lack enterprise knowledge and trivially fail to solve domain-specific NLP tasks. For instance, in the financial domain, state-of-the-art yet generalist models have shown poor performance for different NLP tasks, for which, on the other hand, further fine-tuning with large additional text corpora has been proved to be helpful in improving the results, like in the case of *FinBert* [12].

Limits of domain-specific fine-tuning. Going further, recent studies are exploring the usage of factual data from enterprise databases to fine-tune LLMs and try to tackle domain-specific question-answering tasks: the factual information is leveraged to synthesize prompt-response pairs based on the data and customize the LLM in a task- and domain-specific direction. A primary example is the *SKILL project* [15], where an LLM is directly trained on factual triples derived from the translation into natural language—the so-called *verbalization*—of Wikidata (namely, the *KELM corpus* [2]) for question-answering tasks. Similarly, other approaches highlight possible improvements of accuracy in question-answering tasks, when textual information is first captured into a database, which is then verbalized and employed for fine-tuning [3].

Yet, even the combination of general-purpose knowledge of pre-trained models and the domain data still offers an accuracy that is not acceptable for core tasks in complex domains. For example, *BloombergGPT* [22] is an LLM fine-tuned on a wide range of financial data, combining internal enterprise knowledge with publicly-available datasets. The results show that the model fine-tuned for the question-answering task outperforms state-of-the-art counterparts by being able to correctly answer questions related to the financial domain. However, *BloombergGPT* has been tested only on questions whose answers are already contained in (or directly entailed by) the factual information of the input databases, either as data or meta-data (e.g., schema information). It is reasonable, in fact, that it does not have enough fine-tuning data or logical capabilities to go further.

A look beyond current solutions. Conversely, from an enterprise application perspective, it would be extremely useful to answer questions by means of intelligently combined uses of the input databases with other logic-intensive sources of knowledge (e.g., regulatory bodies, best practices, domain experts, etc.). For instance, in the context of financial markets, answering questions such as *"what factors contribute to the profitability of a market order for a particular trader?"* (**explanation**), or *"what is the behavior of this specific smart contract?"* (**description**), or *"what is the size of a market position that a trader initiated today at 4 PM?"* (**question answering**), or *"based on the available data, what is the individual contribution of each order to the overall profit of a specific trader?"* (**text-to-query translation**) would be an essential asset.

At present, all the mentioned tasks are far from being solved by off-the-shelf libraries or, directly, by most recent LLMs, and are open research. Going into the details of each of them is beyond the scope of this paper, but the motivations, which we will lay out mainly in a question-answering perspective, give the flavour of why LLMs are not enough. It is worth remarking, though, that even the translation task, for which thanks to LLMs much progress has been made in the transformation of natural language into the target query languages (say, SQL, SPARQL, etc.) [21,23] is still a largely unsolved problem, especially in the context of languages with an elaborate grammar and complex queries [9].

Ontological reasoning. From another perspective, in the Knowledge Representation and Reasoning community [11], the state-of-the-art research on ontological reasoning over EKGs makes a point of being able to offer a compact combination of factual database information (the *extensional knowledge*) and formally specified business awareness, for instance in the form of logical rules (the *intensional knowledge*), to serve *domain-specific query answering* in an accurate manner. For example, logical KGs exploiting efficient fragments of the *Datalog*$^{\pm}$ family [8] have been successfully adopted for financial applications [6].

Yet, there is an impedance mismatch between NLP and ontological reasoning, which lacks the flexibility and the language orientation to solve explanation, description, question answering, and translation tasks: queries need to be specified in KRR formalisms; all the inputs and the results are facts/n-tuples/triples; the generation of new knowledge is possible only to the extent reasoning rules capture it. Conversely, while being very good at manipulating human language, LLMs lack a comprehensive domain model, a pillar of KRR approaches.

An integrated approach. This paper strives to strengthen LLMs in their use for task- and domain-specific applications, by letting the fine-tuning process be driven by an ontological reasoning task on an EKG. We operate in the context of the VADALOG [5] system, a Datalog-based reasoning engine for EKGs, that finds many industrial applications [6]. We use VADALOG to explore the factual information derived by applying the domain rules, via the CHASE procedure [13], to the enterprise data and synthesize a fine-tuning corpus that covers the entire "reasoning space" to convey domain-specificity to the LLM. A summary of the resulting *fine-tuning pipeline*, provided in Figure 1, will guide our discussion.

More in detail, our **contributions** can be summarized as follows.

- We present a **reasoning verbalization** technique that generates sets of prompt-response pairs from ground Datalog rules. We provide the algorithm and optimize it with a *lifting technique* exploiting reasoning regularities.

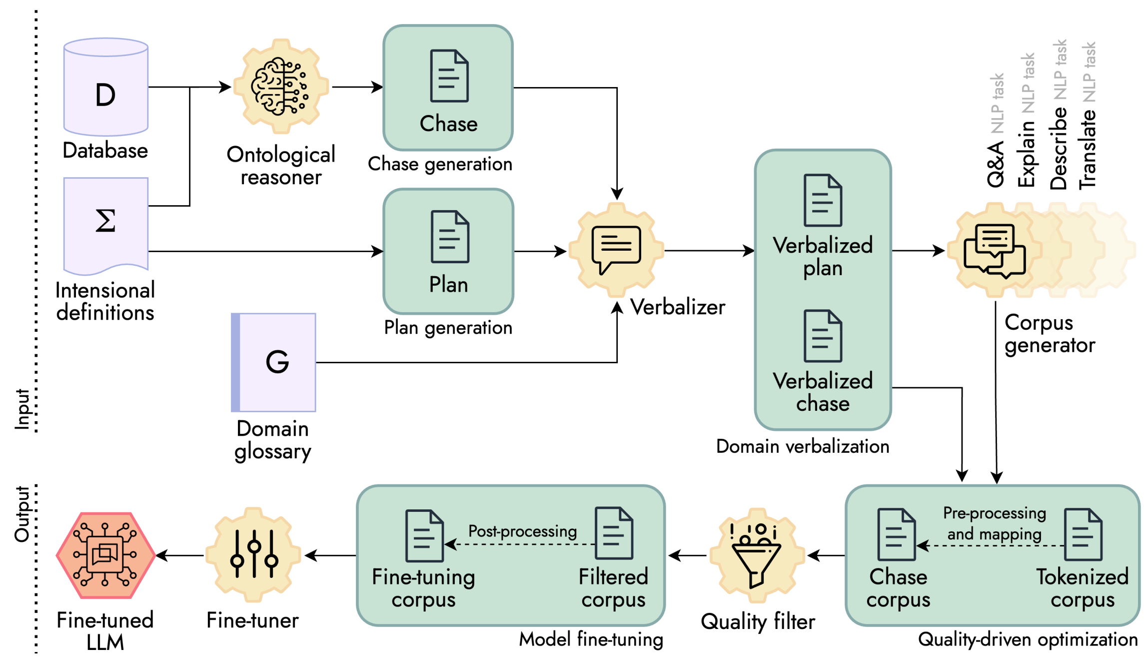

Fig. 1: Neurosymbolic pipeline for reasoning-based LLM fine-tuning.

- We deliver such an approach in a **novel neurosymbolic architecture** that fine-tunes task-specific LLMs for a set of four relevant NLP tasks, namely, *explanation*, *description*, *question answering*, and *translation*.
- We discuss a preliminary **proof-of-concept** focussed on a particular financial market. We compare LLMs fine-tuned on both ground and chase data, thereby validating the effectiveness of our approach.

Overview. In Section 2 we present our architecture. A preliminary experimental validation is provided in Section 3. We draw our conclusions in Section 4.

2 A Neurosymbolic Pipeline to Fine-tune LLMs

The input blocks of the fine-tuning pipeline in Figure 1 are D and Σ. They are, respectively, a database of domain facts and a set of reasoning rules, capturing the business dynamics. Our rules are expressed in VADALOG. An EKG is a combination $\Sigma(D)$ of D and Σ, obtained through reasoning. The set $\Sigma(D)$ is computed via the CHASE [13]: starting from $\Sigma(D) = D$, the chase augments $\Sigma(D)$ with facts derived from the application of the rules in Σ to fixpoint.

Let us introduce our running example: a simple trading activity managed with a *smart contract* [14]. Here, D contains a log over time of buy/sell orders from the traders who invest in the smart contract as well as market information, e.g., asset prices (*Price*), or market shutdowns (*MarketClosed*).

Algorithm 1 Reasoning-based LLM Fine-tuning.

```
 1: function ReasoningFineTuning(D, Σ, G, model, nlpTask)
 2:   chase ← Vadalog.reason(D, Σ)                                   ▹ chase generation
 3:   verbChase ← ∅
 4:   for each step in chase do
 5:     stepAggrContrib ← ∅
 6:     if hasAggregate(step.getRule()) then
 7:       stepAggrContrib ← composeBack(step, chase)                  ▹ aggregates retrieval
 8:     verbStep ← verbalizeStep(step, stepAggrContrib, G)
 9:     verbChase ← verbChase ∪ {verbStep}                             ▹ chase verbalization
10:   verbPlan ← verbalizePlan(Σ.getLogicPlan())                      ▹ logic plan verbalization
11:   tokenizedCorpus ← generate(preprocess(verbPlan, nlpTask))       ▹ tok. corpus gen.
12:   chaseCorpus ← ∅
13:   for each verbStep in verbChase do                                ▹ chase mapping
14:     chasePromptResp ← map(tokenizedCorpus, verbStep)
15:     chaseCorpus ← chaseCorpus ∪ {chasePromptResp}
16:   for each pair ⟨prompt, resp⟩ in chaseCorpus do                  ▹ quality-driven optimization
17:     qualityScore ← checkQuality(⟨prompt, resp⟩, nlpTask, verbChase)
18:     if qualityScore ≤ threshold then
19:       chaseCorpus ← chaseCorpus \ {⟨prompt, resp⟩}
20:     else
21:       chaseCorpus ← chaseCorpus ∪ paraphrase(⟨prompt, resp⟩)
22:   fineTuningCorpus ← postprocess(chaseCorpus)
23:   ftModel ← fineTune(model, fineTuningCorpus)                      ▹ model fine-tuning
24:   return ftModel
```

Example 1. The following set Σ *contains the* Vadalog *rules governing the basic functioning of the market, i.e., under which conditions the orders are accepted and how profits and losses are computed.*

$$Open(x, y, t_1), \neg\, MarketClosed(t_1) \rightarrow Accepted(x, y, t_1) \tag{1}$$

$$Accepted(x, y, t_1), Price(p_1, t_1), k = y * p_1 \rightarrow Position(x, y, k, t_1) \tag{2}$$

$$Close(x, t_2), Price(p_2, t_2), Position(x, y, k, t_1), t_2 > t_1, pl = y * p_2 - k \rightarrow Return(x,pl) \tag{3}$$

If a trader x *wants to open a position (buy) on a certain asset of size* y *at time* t_1 *and the market is open at* t_1*, the order is accepted (rule* 1*). If the order by* x *is accepted and the asset price at* t_1 *is* p_1*, then* x *holds a position on the market at time* t_1 *of size* y *and of notional (total value)* k *equal to* $y * p_1$ *(rule* 2*). If, later at* t_2*, trader* x *decides to close its position (sell) and the price at* t_2 *is* p_2*, then* x *gets returns (profits or losses) from its trading activity as* $y * p_2 - k$ *(rule* 3*).*

Applying the vision we laid out to Example 1, the goal of our pipeline is fine-tuning an LLM to address *explanation*, *description*, *question answering*, and *text-to-query translation* tasks for the simple trading activity at hand. Let us follow Figure 1 and Algorithm 1 to describe the application of the pipeline to a database $D = \{$*Open*(*EGTech*,0.3,1), *Open*(*IEComp*,0.5,1), *Price*(124,1), *Price*(147,9), *Close*(*EGTech*,9), *MarketClose*(5)$\}$.

Chase generation. The first step of our pipeline builds the chase $\Sigma(D)$, that is, the expansion of D with the facts that can be derived by applying the rules of Σ (line 2, in the algorithm). Rule 1 generates the fact $Accepted(EGTech, 0.3, 1)$, as the market is not closed at time 1. Then, $Position(EGTech, 0.3, 37.2, 1)$ is derived via rule 2. Finally, as trader *EGTech* closes the position, i.e., sells the asset, at time 9 and the price goes up to 147$, then *EGTech* gets a profit of 6.9$.

Domain verbalization. Whenever a VADALOG rule is involved in the CHASE, it is translated into pure text with a deterministic transformation, based on the *select-project-join* semantics, which looks up a *glossary* $\boldsymbol{G}$ of atom descriptions. When rules involve aggregation functions, allowed in VADALOG, the process is less straightforward and involves unfolding a chain of chase activations altogether [1] (line 7). At the end of this phase, we are in hold of a "*since-then* closure" of our domain, that focuses on what can be obtained by activating the intensional knowledge of Σ. From another perspective, Σ can be seen as an *attention* mechanism, to select the fragment of $\boldsymbol{D}$ that one wants to verbalize. For instance, with respect to our running example, the chase step $Open(EGTech, 0.3, 1), \neg MarketClose(1) \rightarrow Accepted(EGTech, 0.3, 1)$ (rule 1) is verbalized as: *Since the trader EGTech at time 1 sends an order to open a position of size 0.3, and it is not true that 1 is a time when the market is closed, then the order of size 0.3 by EGTech is accepted at time 1.*

Fine-tuning corpus generation. With the basic verbalization available, we are now ready to generate the fine-tuning corpus. We consider the corpus generation itself as a text manipulation task and exploit the effectiveness of powerful pre-trained LLMs [7], such as GPT-3, to synthesize a finite set of possible prompt-response pairs. Here we have two goals: 1) minimising the number of "calls" to the LLM, for cost- and time-efficiency reasons; 2) avoiding any ground value (coming from the EKG) being disclosed to the LLM, for data protection reasons. We leverage the regularity of logical languages and resort to a *lifting technique*. We build a *logic plan* out of Σ (line 10). A plan is the equivalent in our context of a database execution plan and can be seen as the dependency graph of the rules of Σ, where nodes represent rules and edges stand for head-body dependencies. The plan is then verbalized, obtaining a text with tokens as placeholders for rule variables. Finally, a tokenized fine-tuning corpus is generated from the plan, after minor pre-processing (line 11). The form of the prompts depends on the task. Now, for each verbalized chase step, we look up the corresponding verbalized portion of the plan and instantiate its tokens (lines 13-15). Note that no invocations to the *corpus generator* are needed in this phase. Figure 2 exemplifies the generation process in our example domain.

<table>
<tr><th>Plan</th><th>Tokenized corpus</th><th>Verbalized chase step</th><th>Chase corpus</th></tr>
<tr><td rowspan="3">$Open(x, y, t_1)$,
$\neg MarketClosed(t_1)$
$\rightarrow Accepted(x, y, t_1)$</td><td>Q1: What is the size of the order accepted sent by trader <x>?
A1: The size is <y></td><td rowspan="3">Since the trader EGTech at time 1 sends an order to open a position of size 0.3, and it is not true that 1 is a time when the market is closed, then the order of size 0.3 by EGTech is accepted at time 1.</td><td>Q1: What is the size of the order accepted sent by trader EGTech?
A1: The size is 0.3</td></tr>
<tr><td>Q2: Why was the order sent by trader <x> at time <t_1> accepted?
A2: Because at time <t_1> the market was open</td><td>Q2: Why was the order sent by trader EGTech at time 1 accepted?
A2: Because at time 1 the market was open</td></tr>
<tr><td>...</td><td>...</td></tr>
<tr><td rowspan="2">$Open(x, y, t_1)$,
$\neg MarketClosed(t_1)$
$\rightarrow Accepted(x, y, t_1)$
$\downarrow$
$Accepted(x, y, t_1)$,
$Price(p_1, t_1), k = y * p_1$
$\rightarrow Position(x, y, k, t_1)$</td><td>Q1: When did <$x$> send an order to open a position with notional <k>?
A1: The order to open that position was sent at time <t_1></td><td rowspan="2">Since the trader EGTech at time 1 sends an order to open a position of size 0.3, and it is not true that 1 is a time when the market is closed, then the order of size 0.3 by EGTech is accepted at time 1. Since the order of size 0.3 by EGTech is accepted and the price is 124 at time 1, then EGTech holds a position of size 0.3 and notional 37.2 at time 1.</td><td>Q1: When did EGTech send an order to open a position with notional 37.2?
A1: The order to open that position was sent at time 1</td></tr>
<tr><td>...</td><td>...</td></tr>
</table>

Fig. 2: From plans to fine-tuning corpus, in our running example.

Quality-driven optimization. The corpus undergoes a quality check where each pair is filtered according to an NLP-based scoring model in terms of specificity, plausibility, absence of bias, and other user-defined criteria. The filtered-in pairs are enhanced via *NLP paraphrasing* to improve generalization and finally cleansed with additional post-processing procedures (lines 16-22).

Model fine-tuning. The refined corpus is injected into an LLM for task- and domain-specific fine-tuning (line 23). In the case of Q&A, the model operates in a *closed-book* approach, that is, it learns to map questions to the corresponding answers without extracting them from an input context, but rather encapsulating the knowledge of the domain into its internal parameters and weights [19]. The resulting specialized model is provided to the user via API, and will act as a natural language interface to the EKG and the ontological reasoning at its foundation in a neurosymbolic fashion.

3 Preliminary Validation via Proof-of-Concept

We implemented our fine-tuning pipeline in VADALOG. While a full-scale evaluation of our architecture is beyond the scope of this short work, in this section, we propose a conceptual validation of the approach, by briefly showing some executions of the pipeline, with a focus on the question answering task. For the proof-of-concept, we made use of a *T5-large* [10] model and considered the same domain as in Example 1. To obtain a dataset that could be visually inspected to informally assess the quality of the textual answers given by an LLM fine-tuned with our pipeline, we performed a kind of *ablation study.*

For randomly chosen sets of sample questions, for the NLP tasks of interest, we compared the answers provided by an LLM fine-tuned only with ground facts (*T5-large-ground*) and one fine-tuned with our pipeline (*T5-large-chase*). Both models were trained for 10 epochs and with the same hyperparameters. The fine-tuning corpora and the models are made available [4].

Figure 3 visually reports the comparison. Questions a and b are the baseline, as they can be answered by facts in D. Apart from a less refined write-up, the LLMs show the same output behaviour. On the other hand, in questions c, d, and f T5-large-ground is outperformed by T5-large-chase, which succeeds in answering about events related to trader *EGTech*. Actually, the corresponding facts derive from $\Sigma(D)$, which is not considered in the ground fine-tuning. Similarly, the answer to question e by T5-large-ground is incomplete and only T5-large-chase is able to use the specific domain knowledge from rule 1 of Example 1. Question f still pertains to a case observed by the models and both the answers are correct, with T5-large-chase providing a thorough explanation rather than a black-box response, as seen with T5-large-ground. Finally, with questions *g-i*, we further challenge our LLMs and ask them to engage in more abstract or hypothetical considerations. These queries gradually depart from the ground data and instead rely on the business setting. As a consequence of the lack of domain knowledge, T5-large-ground provides factually incorrect answers, while T5-large-chase accomplishes the task.

4 Conclusion

According to a recent work [24] appeared in the European Chapter of the Association for Computational Linguistics, pre-trained language models cannot yet perform deductive reasoning: they are still unable to generalize logical rules and, even when rules are provided, LLMs tend to forget previously inferred facts. While no extensive comparison between transformer architectures and reasoning approaches has been conducted yet, our work showed that LLM performance for domain-specific NLP tasks can be visibly improved by

	User question	T5-large-ground	T5-large-chase
a.	What is the size of the position that EGTech opens at time 1?	0.3	The size is 0.3
b.	What is the notional of EGTech's position at time 1?	37.2$	The notional is 37.2$, equal to price 124$ × size 0.3
c.	Why is the order by EGTech at time 1 accepted?	As it is a trade	Because the market at time 1 is not closed
d.	How can EGTech hold a position on the market at time 1 of notional 3$?	0.3	Since EGTech's position of size 0.3 was accepted and the price at time 1 is 10
e.	How does a trader get their position accepted?	By placing an order	By opening the position at a time in which the market is not closed
f.	How does EGTech profits from its trading activity?	EGTech has a profit of 6.9$	By opening a position of size 0.3 at time 1, having the position accepted with an initial price of 124$ and a notional of 37.2$ at time 1, and then closing it at time 9 with a final price of 147$, thus having a profit of 6.9$
g.	Trader Charlie wants to open a position at time 1 with a notional value of $10,000. If the asset price at 1 is $50, what size of the position does he need to open?	Charlie needs to open a position	The notional value (10,000$) is given by the size of the position multiplied by the asset price (50$) at time 1, so trader Charlie needs to derive the value reversing this formula
h.	If Trader Mike decides to close his position of size 9 at time 2 with a return of $3,000 and the price at that time is $20, what is the notional value?	The notional value at time 2 is 3,000$	The notional value can be obtained knowing that the returns of 3,000$ are computed as size of the position of 9 × 20$, which is the price at time 2 minus the notional value
i.	At time 1, Trader Sarah opens a position by buying 200 shares of a company's stock. If the stock price at 1 is $60 per share, what is the notional value of Sarah's position?	The notional value of Sarah's position is 60$.	Trader Sarah opens a position at time 1 by buying 200 shares of the stock at 60$ per share, resulting in a notional value of 60$ × 200

Fig. 3: Proof-of-concept for our fine-tuning pipeline.

producing a fine-tuning corpus as a byproduct of ontological reasoning. We capitalized on our experience in deductive reasoning to offer a first step towards a neuro-symbolic platform for reasoning on enterprise knowledge graphs.

Acknowledgements. The work on this paper was partially supported by the Vienna Science and Technology Fund (WWTF) grant VRG18-013.

References

1. Afrati, F.N., Gergatsoulis, M., Toni, F.: Linearisability on datalog programs. Theor. Comput. Sci. **308**(1-3), 199–226 (2003)
2. Agarwal, O., Ge, H., Shakeri, S., Al-Rfou, R.: Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. arXiv preprint arXiv:2010.12688 (2020)
3. Andrus, B.R., Nasiri, Y., Cui, S., Cullen, B., Fulda, N.: Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. AAAI **36**(10), 10436–10444 (2022)
4. Baldazzi, T., Bellomarini, L., Ceri, S., Colombo, A., Gentili, A., Sallinger, E.: Material. `https://bit.ly/44249b5`, accessed: 2023-06-17
5. Bellomarini, L., Benedetto, D., Gottlob, G., Sallinger, E.: Vadalog: A modern architecture for automated reasoning with large knowledge graphs. IS **105** (2022)
6. Bellomarini, L., Fakhoury, D., Gottlob, G., Sallinger, E.: Knowledge graphs and enterprise AI: the promise of an enabling technology. In: ICDE. pp. 26–37 (2019)
7. Brown, T., et al.: Language models are few-shot learners. In: NeurIPS. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
8. Calì, A., Gottlob, G., Lukasiewicz, T.: A general datalog-based framework for tractable query answering over ontologies. J. Web Semant. **14**, 57–83 (2012)
9. Fu, H., Liu, C., Wu, B., Li, F., Tan, J., Sun, J.: Catsql: Towards real world natural language to sql applications. VLDB **16**(6), 1534–1547 (2023)
10. Google: T5 large. `https://huggingface.co/t5-large`, accessed: 2023-06-17
11. Krötzsch, M., Thost, V.: Ontologies for knowledge graphs: Breaking the rules. In: ISWC (1). LNCS, vol. 9981, pp. 376–392 (2016)
12. Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J.: Finbert: A pre-trained financial language representation model for financial text mining. In: IJCAI 2020 (2021)

13. Maier, D., Mendelzon, A.O., Sagiv, Y.: Testing implications of data dependencies. ACM TODS **4**(4), 455–469 (1979)
14. Mohanta, B.K., Panda, S.S., Jena, D.: An overview of smart contract and use cases in blockchain technology. In: ICCCNT. pp. 1–4 (2018)
15. Moiseev, F., Dong, Z., Alfonseca, E., Jaggi, M.: SKILL: Structured knowledge infusion for large language models. In: ACL 2022. pp. 1581–1588 (2022)
16. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
17. Rae, J.W., et al.: Scaling language models: Methods, analysis & insights from training gopher. arXiv preprint arXiv:2112.11446 (2021)
18. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020)
19. Roberts, A., Raffel, C., Shazeer, N.: How much knowledge can you pack into the parameters of a language model? In: EMNLP (1). pp. 5418–5426. ACL (2020)
20. Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T.: Transfer learning in natural language processing. In: NAACL: Tutorials. pp. 15–18 (2019)
21. Wang, B., et al.: Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. arXiv preprint arXiv:1911.04942 (2019)
22. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D.S., Mann, G.: Bloomberggpt: A large language model for finance. CoRR **abs/2303.17564** (2023)
23. Yin, X., Gromann, D., Rudolph, S.: Neural machine translating from natural language to sparql. Future Generation Computer Systems **117**, 510–519 (2021)
24. Yuan, Z., Hu, S., Vulic, I., Korhonen, A., Meng, Z.: Can pretrained language models (yet) reason deductively? In: EACL. pp. 1439–1454 (2023)

Papers published in the 'Markets, Infrastructures, Payment Systems' series

n. 1 TIPS - TARGET Instant Payment Settlement – The Pan-European Infrastructure for the Settlement of Instant Paymentsi, *by Massimiliano Renzetti, Serena Bernardini, Giuseppe Marino, Luca Mibelli, Laura Ricciardi and Giovanni M. Sabelli* (Institutional Issues)

n. 2 Real-Time Gross Settlement systems: breaking the wall of scalability and high availability, *by Mauro Arcese, Domenico Di Giulio and Vitangelo Lasorella* (Research Papers)

n. 3 Green Bonds: the Sovereign Issuers' Perspective, *by Raffaele Doronzo, Vittorio Siracusa and Stefano Antonelli* (Research Papers)

n. 4 T2S - TARGET2-Securities – The pan-European platform for the settlement of securities in central bank money, *by Cristina Mastropasqua, Alessandro Intonti, Michael Jennings, Clara Mandolini, Massimo Maniero, Stefano Vespucci and Diego Toma* (Institutional Issues)

n. 5 The carbon footprint of the Target Instant Payment Settlement (TIPS) system: a comparative analysis with Bitcoin and other infrastructures, *by Pietro Tiberi* (Research Papers)

n. 6 Proposal for a common categorisation of IT incidents, *by Autorité de Contrôle Prudentiel et de Résolution, Banca d'Italia, Commissione Nazionale per le Società e la Borsa, Deutsche Bundesbank, European Central Bank, Federal Reserve Board, Financial Conduct Authority, Ministero dell'Economia e delle Finanze, Prudential Regulation Authority, U.S. Treasury* (Institutional Issues)

n. 7 Inside the black box: tools for understanding cash circulation, *by Luca Baldo, Elisa Bonifacio, Marco Brandi, Michelina Lo Russo, Gianluca Maddaloni, Andrea Nobili, Giorgia Rocco, Gabriele Sene and Massimo Valentini* (Research Papers)

n. 8 The impact of the pandemic on the use of payment instruments in Italy, *by Guerino Ardizzi, Alessandro Gambini, Andrea Nobili, Emanuele Pimpini and Giorgia Rocco* (Research Papers) (in Italian)

n. 9 TARGET2 – The European system for large-value payments settlement, *by Paolo Bramini, Matteo Coletti, Francesco Di Stasio, Pierfrancesco Molina, Vittorio Schina and Massimo Valentini* (Institutional Issues) (in Italian)

n. 10 A digital euro: a contribution to the discussion on technical design choices, *by Emanuele Urbinati, Alessia Belsito, Daniele Cani, Angela Caporrini, Marco Capotosto, Simone Folino, Giuseppe Galano, Giancarlo Goretti, Gabriele Marcelli, Pietro Tiberi and Alessia Vita* (Institutional Issues)

n. 11 From SMP to PEPP: a further look at the risk endogeneity of the Central Bank, *by Marco Fruzzetti, Giulio Gariano, Gerardo Palazzo and Antonio Scalia* (Research Papers)

n. 12 TLTROs and collateral availability in Italy, *by Annino Agnes, Paola Antilici and Gianluca Mosconi* (Research Papers) (in Italian)

n. 13 Overview of central banks' in-house credit assessment systems in the euro area, *by* Laura *Auria, Markus Bingmer, Carlos Mateo Caicedo Graciano, Clémence Charavel, Sergio Gavilá, Alessandra Iannamorelli, Aviram Levy, Alfredo Maldonado, Florian Resch, Anna Maria Rossi and Stephan Sauer* (Institutional Issues)

n. 14 The strategic allocation and sustainability of central banks' investment, *by Davide Di Zio, Marco Fanari, Simone Letta, Tommaso Perez and Giovanni Secondin* (Research Papers) (in Italian)

n. 15 Climate and environmental risks: measuring the exposure of investments, *by Ivan Faiella*, *Enrico Bernardini*, *Johnny Di Giampaolo*, *Marco Fruzzetti*, *Simone Letta, Raffaele Loffredo and Davide Nasti* (Research Papers)

n. 16 Cross-Currency Settlement of Instant Payments in a Multi-Currency Clearing and Settlement Mechanism, *by Massimiliano Renzetti*, *Fabrizio Dinacci and Ann Börestam* (Research Papers)

n. 17 What's ahead for euro money market benchmarks?, *by Daniela Della Gatta* (Institutional Issues) (in Italian)

n. 18 Cyber resilience per la continuità di servizio del sistema finanziario, *by Boris Giannetto and Antonino Fazio* (Institutional Issues) (in Italian)

n. 19 Cross-Currency Settlement of Instant Payments in a Cross-Platform Context: a Proof of Concept, *by Massimiliano Renzetti, Andrea Dimartina, Riccardo Mancini, Giovanni Sabelli, Francesco Di Stasio, Carlo Palmers, Faisal Alhijawi, Erol Kaya, Christophe Piccarelle, Stuart Butler, Jwallant Vasani, Giancarlo Esposito, Alberto Tiberino and Manfredi Caracausi* (Research Papers)

n. 20 Flash crashes on sovereign bond markets – EU evidence, *by Antoine Bouveret, Martin Haferkorn, Gaetano Marseglia and Onofrio Panzarino* (Research Papers)

n. 21 Report on the payment attitudes of consumers in Italy: results from ECB surveys, *by Gabriele Coletti, Alberto Di Iorio, Emanuele Pimpini and Giorgia Rocco* (Institutional Issues)

n. 22 When financial innovation and sustainable finance meet: Sustainability-Linked Bonds, *by Paola Antilici, Gianluca Mosconi and Luigi Russo* (Institutional Issues) (in Italian)

n. 23 Business models and pricing strategies in the market for ATM withdrawals, *by Guerino Ardizzi and Massimiliano Cologgi* (Research Papers)

n. 24 Press news and social media in credit risk assessment: the experience of Banca d'Italia's In-house Credit Assessment System, *by Giulio Gariano and Gianluca Viggiano* (Research Papers)

n. 25 The bonfire of banknotes, *by Michele Manna* (Research Papers)

n. 26 Integrating DLTs with market infrastructures: analysis and proof-of-concept for secure DvP between TIPS and DLT platforms, *by Rosario La Rocca, Riccardo Mancini, Marco Benedetti, Matteo Caruso, Stefano Cossu, Giuseppe Galano, Simone Mancini, Gabriele Marcelli, Piero Martella, Matteo Nardelli and Ciro Oliviero* (Research Papers)

n. 27 Statistical and forecasting use of electronic payment transactions: collaboration between Bank of Italy and Istat, *by* Guerino Ardizzi and Alessandra Righi (Institutional Issues) (in Italian)

n. 28 TIPS: a zero-downtime platform powered *by* automation, by *Gianluca Caricato, Marco Capotosto, Silvio Orsini and Pietro Tiberi* (Research Papers)

n. 29 TARGET2 analytical tools for regulatory compliance, *by Marc Glowka, Alexander Müller, Livia Polo Friz, Sara Testi, Massimo Valentini and Stefano Vespucci* (Institutional Issues)

n. 30 The security of retail payment instruments: evidence from supervisory data, *by Massimiliano Cologgi* (Research Papers)

n. 31 Open Banking in the payment system: infrastructural evolution, innovation and security, supervisory and oversight practices, *by Roberto Pellitteri, Ravenio Parrini, Carlo Cafarotti and Benedetto Andrea De Vendictis* (Institutional Issues) (in Italian)

n. 32 Banks' liquidity transformation rate: determinants and impact on lending, *by Raffaele Lenzi, Stefano Nobili, Filippo Perazzoli and Rosario Romeo* (Research Papers)

n. 33 Investor behavior under market stress: evidence from the Italian sovereign bond market, *by Onofrio Panzarino* (Research Papers)

n. 34 Siamese neural networks for detecting banknote printing defects, *by Katia Boria, Andrea Luciani, Sabina Marchetti and Marco Viticoli* (Research Papers) (in Italian)

n. 35 Quantum safe payment systems, *by Elena Bucciol and Pietro Tiberi*

n. 36 Investigating the determinants of corporate bond credit spreads in the euro area, *by Simone Letta and Pasquale Mirante*

n. 37 Smart Derivative Contracts in DatalogMTL, *by Andrea Colombo, Luigi Bellomarini, Stefano Ceri and Eleonora Laurenza*

n. 38 Making it through the (crypto) winter: facts, figures and policy issues, *by Guerino Ardizzi, Marco Bevilacqua, Emanuela Cerrato and Alberto Di Iorio*

n. 39 The Emissions Trading System of the European Union (EU ETS), *by Mauro Bufano, Fabio Capasso, Johnny Di Giampaolo and Nicola Pellegrini* (in Italian)

n. 40 Banknote migration and the estimation of circulation in euro area countries: the italian case, *by Claudio Doria, Gianluca Maddaloni, Giuseppina Marocchi, Ferdinando Sasso, Luca Serrai and Simonetta Zappa* (in Italian)

n. 41 Assessing credit risk sensitivity to climate and energy shocks, *by Stefano Di Virgilio, Ivan Faiella, Alessandro Mistretta and Simone Narizzano*

n. 42 Report on the payment attitudes of consumers in italy: results from the ecb space 2022 survey, *by Gabriele Coletti, Alberto Di Iorio, Emanuele Pimpini and Giorgia Rocco*

n. 43 A service architecture for an enhanced Cyber Threat Intelligence capability and its value for the cyber resilience of Financial Market Infrastructures, *by Giuseppe Amato, Simone Ciccarone, Pasquale Digregorio and Giuseppe Natalucci*