Contextual Reasoning Orchestration for Enhancing Black-Box Large Language Models in Specialized Decision Support

Haoyu Cen * and Yutian Gai

Posted Date: 18 November 2025

doi: 10.20944/preprints202511.1307.v1

Keywords: large language models; decision support; financial analysis

 doi:10.20944/preprints202511.1307.v1

Article

Contextual Reasoning Orchestration for Enhancing Black-Box Large Language Models in Specialized Decision Support

Haoyu Cen * and Yutian Gai

Polytechnic Colleges, Malaysia
* Correspondence: me01084755@student.uniten.edu.my

Abstract

Recent advances in language models have greatly improved their ability to understand and generate natural language. Yet, when applied to specialized fields such as financial decision support or complex system diagnosis, they often struggle with limited domain expertise, weak logical reasoning, and unreliable performance under uncertainty. Fine-tuning these large models is typically constrained by cost, privacy, and proprietary limitations. To overcome these issues, this study introduces CRONUS: Contextual Reasoning Orchestration for Navigating Uncertain Scenarios, a framework designed to enhance general-purpose models in domain-specific and decision-intensive tasks. CRONUS employs a lightweight, trainable agent named CARA (Context-Aware Reasoning Agent) to guide the reasoning process of black-box models through structured contextual instructions. CARA is developed via a three-stage training strategy that builds domain understanding, refines reasoning path generation, and optimizes dynamic decision prompts. Experiments in financial analysis tasks show that CRONUS markedly improves reasoning depth, consistency, and robustness compared with direct model use, retrieval-augmented methods, and specialized domain models, demonstrating its effectiveness for high-stakes decision-making in complex environments.

Keywords: large language models; decision support; financial analysis

1. Introduction

The advent of large language models (LLMs) such as ChatGPT and GPT-4 has revolutionized natural language understanding and generation across a myriad of open-domain tasks [1], showcasing remarkable potential for weak to strong generalization across multi-capabilities [2]. Their remarkable ability to comprehend intricate queries and generate coherent, contextually relevant responses has positioned them as central components in various AI-driven applications. However, despite their general prowess, the deployment of these black-box LLMs in specialized vertical domains—such as high-stakes financial decision support [3], complex system fault diagnosis [4], or personalized educational pathway planning [5]—still faces significant hurdles. These domains demand not only deep, specialized domain knowledge but also multi-step logical reasoning, robust handling of uncertainty, and the ability to integrate information from diverse sources, as exemplified by challenges in autonomous navigation with SLAM systems [6–8] and sophisticated industrial control for online monitoring and parameter estimation [9–11]. Recent advances in abnormal electricity usage detection [12] further highlight the growing need for reliable, domain-aware reasoning frameworks capable of operating under uncertain and safety-critical conditions.

 doi:10.20944/preprints202511.1307.v1

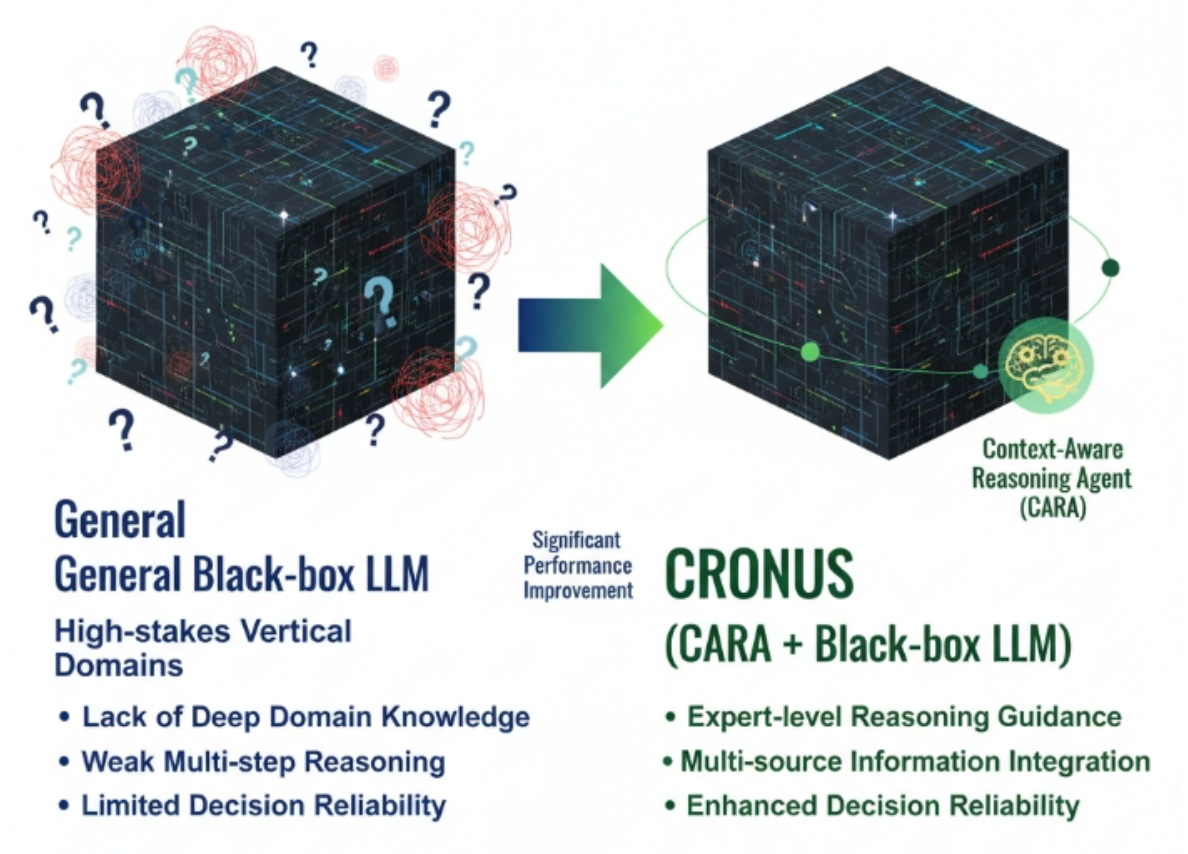

Figure 1. Bridging the Gap: CRONUS Enhances Black-box LLMs for High-stakes Vertical Domains through Contextual Reasoning Orchestration.

A primary challenge arises from the inherent limitations of general-purpose LLMs when confronted with domain-specific complexities. They often lack the requisite background knowledge, struggle with the nuanced logical inferences characteristic of expert tasks, and may exhibit insufficient reliability in making critical judgments under uncertain conditions. Furthermore, directly fine-tuning these powerful black-box LLMs is frequently impractical due to several constraints: the prohibitive computational cost, the proprietary nature of the models which often restricts access to internal parameters, and privacy concerns associated with sensitive domain data. These barriers underscore the urgent need for methods that can enhance the performance of black-box LLMs in vertical domains without requiring internal modifications or parameter access.

To address this core challenge, we propose a novel framework named **CRONUS: Contextual Reasoning Orchestration for Navigating Uncertain Scenarios**. CRONUS is designed to significantly boost the performance of black-box general-purpose LLMs in tasks demanding multi-source information integration and complex decision-making within vertical domains. Our approach introduces a lightweight, trainable **Context-Aware Reasoning Agent (CARA)** that works in conjunction with the black-box LLM. CARA is specifically trained to learn and generate expert-level contextual reasoning guidance, thereby orchestrating the black-box LLM's capabilities to better navigate high-risk, high-complexity vertical tasks without altering its internal architecture.

The CRONUS framework is built upon three key training stages for the CARA model: *Domain Knowledge & Logic Pre-training (DKLP)*, *Contextual Reasoning Path Instruction Tuning (CRPIT)*, and *Dynamic Decision Prompt Optimization (DDPO)*. In the DKLP stage, CARA acquires foundational domain knowledge from vast unsupervised corpora relevant to the target vertical domain. Subsequently, CRPIT focuses on teaching CARA to generate detailed, multi-step reasoning paths by leveraging carefully filtered pseudo-data derived from black-box LLMs or domain experts. This filtering process ensures that only high-quality reasoning paths that demonstrably improve the black-box LLM's performance are retained. Finally, DDPO employs a non-differentiable optimization strategy, such as Bayesian

 doi:10.20944/preprints202511.1307.v1

optimization, to dynamically refine how CARA's generated reasoning paths are integrated into the prompts for the black-box LLM, ensuring optimal alignment and performance. This modular design ensures that the black-box LLM remains untouched, embodying a truly "black-box friendly" paradigm.

Our experimental evaluation focuses on complex decision support tasks within financial market analysis. We utilize a self-constructed dataset, **FinDecision-QA**, which comprises both fact-recall and intricate situational reasoning questions, alongside the existing **FinCausal** dataset [13] for causal inference in financial contexts. We evaluate CRONUS by coupling our CARA with leading black-box LLMs, including ChatGPT (GPT-3.5/GPT-4 API) and Baichuan2-13B-Chat. Performance is primarily measured by accuracy in zero-shot settings. Our results, consistent with our hypothesis, demonstrate that CRONUS significantly outperforms direct black-box LLM usage, retrieval-augmented generation (RAG) methods [14], and even existing domain-specific LLMs. Specifically, CRONUS achieves a notable improvement in tasks requiring deep situational reasoning, validating the effectiveness of our context-aware reasoning orchestration. For instance, on the FinDecision-QA dataset, CRONUS combined with ChatGPT attained an overall accuracy of 70.5%, surpassing the 66.8% of a dedicated FinLLM-13B and 64.7% of RAG-ChatGPT. This performance gain is particularly pronounced in "situational reasoning type questions," where CRONUS + ChatGPT achieved 65.2% accuracy compared to 60.5% for FinLLM-13B, highlighting CARA's ability to effectively guide complex inferences.

In summary, this paper makes the following key contributions:

- We propose **CRONUS**, a novel, black-box friendly framework that significantly enhances the performance of general-purpose LLMs in complex vertical domain decision support tasks without requiring internal model modifications.
- We introduce the **Context-Aware Reasoning Agent (CARA)**, a lightweight, domain-specific model trained through a multi-stage process (DKLP, CRPIT, DDPO) to generate high-quality contextual reasoning paths that guide black-box LLMs.
- We demonstrate the superior performance of CRONUS in the challenging domain of financial market analysis, achieving state-of-the-art results on our custom **FinDecision-QA** dataset, particularly in tasks demanding intricate situational reasoning.

2. Related Work

2.1. Enhancing and Orchestrating Black-Box Large Language Models

This section reviews approaches for enhancing black-box Large Language Models (LLMs). For prompt engineering, Ding et al. [15] proposed a prompt-learning framework to improve fine-grained entity typing. To mitigate hallucinations, Dhuliawala et al. [16] introduced HICD, a contrastive decoding strategy that manipulates attention heads as an alternative to chain-of-thought reasoning. In the realm of in-context learning (ICL), Rubin et al. [17] found that pruning demonstrations can enhance LLM performance and proposed PromptQuine, an evolutionary framework to discover effective pruning strategies. This research area also includes visual in-context learning for large vision-language models [18]. For knowledge integration, Onoe et al. [19] used box embeddings to model interdependencies between entity types, which could inform knowledge structuring in Retrieval-Augmented Generation (RAG) systems. For adapting models to sensitive applications, Turcan et al. [20] showed emotion-infused models can mitigate performance disparities when adapting stress models to underrepresented groups in low-data regimes. Similarly, the MultiFinRAG framework enhances black-box LLMs by integrating multimodal knowledge for complex financial question answering. To improve reasoning and reliability, Weng et al. [21] found that fine-tuning LLMs on verbalized confidence can elicit self-verification and improve accuracy, while other methods like 'Thread of Thought' aim to improve performance in chaotic contexts [22]. Enhancements also extend to generating executable outputs, such as using reinforcement learning to produce spreadsheet formulas [23]. Broader efforts focus on multimodal capabilities, including efficient image-text matching [24] and hierarchical 2D-3D cross-modal retrieval [25].

 doi:10.20944/preprints202511.1307.v1

2.2. Domain Adaptation and Specialized Language Models

This section explores domain adaptation strategies. One method disentangles domain-specific and domain-invariant representations to improve cross-lingual and cross-domain performance. Li et al. [13] investigated the efficacy of LLMs on financial text analytics, offering insights into their domain-specific limitations and potential. Gururangan et al. [26] introduced the DEMix layer, a modular architecture with domain-specific expert networks to enhance generalization. For knowledge-intensive tasks, Tedeschi et al. [27] proposed a method to generate high-quality, multilingual silver data for Named Entity Recognition (NER). To address model biases, Hardalov et al. [28] introduced Relative Counterfactual Contrastive Learning (RCCL) to mitigate pretrained stance bias. In medicine, Labrak et al. [29] showed that domain-specific pretraining enhances performance in endoscopic video analysis. This specialization trend includes models for cardiac disease prediction [30] and brain lesion segmentation [31]. This principle extends to diverse fields such as architecture for residential design [32] and 3D urban block generation [33], as well as ship detection in remote sensing [34], abnormal electricity usage detection [12], and creative tools for embroidery [35]. Bao et al. [36] employed a curriculum learning strategy in PLATO-2 for open-domain chatbots, offering a structured approach to model adaptation. Domain adaptation also involves critical ethical considerations like ensuring group fairness [37]. Finally, Sciavolino et al. [38] highlighted the generalization limits of dense retrievers on entity-centric questions, suggesting that specialized question encoders are key to improving adaptation.

3. Method

In this section, we detail our proposed **CRONUS: Contextual Reasoning Orchestration for Navigating Uncertain Scenarios** framework. CRONUS is specifically designed to significantly enhance the performance of black-box large language models (LLMs) when applied to complex decision support tasks within specialized vertical domains. A cornerstone of CRONUS is its **black-box friendly** nature, which means it achieves substantial performance gains without requiring any internal modifications, fine-tuning, or direct parameter access to the underlying general-purpose LLM. Instead, CRONUS introduces a lightweight, trainable **Context-Aware Reasoning Agent (CARA)** that acts as an intelligent orchestrator. CARA's primary role is to generate expert-level contextual reasoning guidance, thereby effectively steering the black-box LLM's capabilities towards more accurate and reliable outcomes in domain-specific challenges.

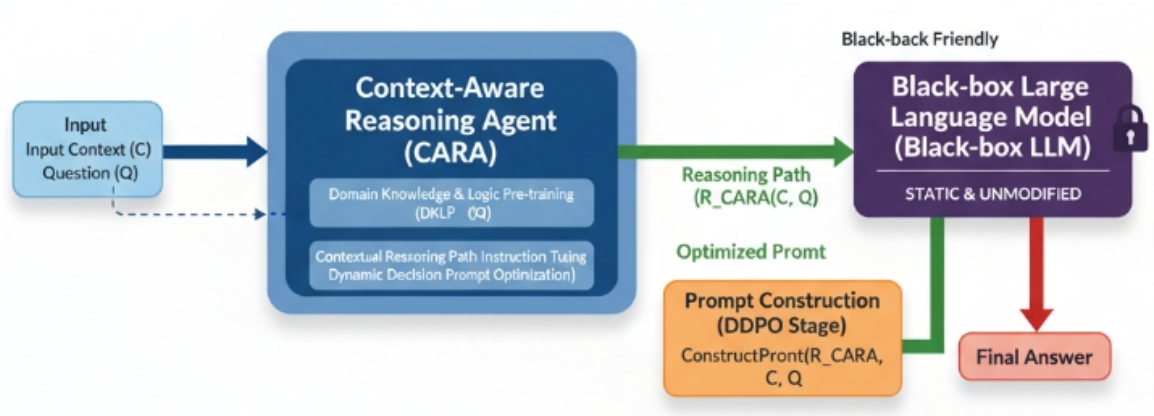

Figure 2. The CRONUS Framework leverages a lightweight Context-Aware Reasoning Agent (CARA) to orchestrate black-box Large Language Models (LLMs) by generating optimized reasoning paths for enhanced decision support.

 doi:10.20944/preprints202511.1307.v1

3.1. CRONUS Framework Overview

The CRONUS framework is architected around the dynamic interaction of two primary components: a powerful, off-the-shelf **Black-box Large Language Model (Black-box LLM)** and our novel **Context-Aware Reasoning Agent (CARA)**.

The **Black-box LLM** serves as the foundational intelligence layer. It is typically a large, pre-trained model renowned for its robust general language understanding, generation capabilities, and vast world knowledge. Crucially, this component remains static and unaltered throughout the entire CRONUS operational cycle, preserving its original integrity, proprietary nature, and foundational capabilities.

The **Context-Aware Reasoning Agent (CARA)**, on the other hand, is a much smaller, purpose-built language model. It is specifically engineered to bridge the inherent gap between the Black-box LLM's general knowledge and the intricate, often nuanced, demands of specialized vertical domains. Such demands arise not only in financial analysis but also in high-stakes industrial environments, where accurate decision-making under uncertainty is crucial, as exemplified by abnormal electricity usage detection systems [12]. CARA's core function is to analyze the given input context and question, then synthesize a structured, multi-step reasoning path. This path effectively guides the Black-box LLM by providing explicit instructions, intermediate logical steps, and relevant domain considerations. This generated guidance is then dynamically integrated into the prompt presented to the Black-box LLM, transforming a generic query into an expertly contextualized instruction.

The overall operational flow of CRONUS for a given input context C and question Q can be conceptualized as follows:

$$\text{FinalAnswer} = \text{LLM}(\text{ConstructPrompt}(R_{\text{CARA}}(C, Q), C, Q)) \tag{1}$$

where $R_{\text{CARA}}(C, Q)$ represents the reasoning path generated by CARA given C and Q, and $\text{ConstructPrompt}(\cdot)$ is the function responsible for integrating this reasoning path and the original input into a coherent prompt for the Black-box LLM.

3.2. Context-Aware Reasoning Agent (CARA)

The **Context-Aware Reasoning Agent (CARA)** is the central innovative component of the CRONUS framework. It is designed as a lightweight language model, typically with parameter counts ranging from 500 million to 2 billion, making it significantly smaller and more agile than the Black-box LLM it orchestrates. CARA's primary objective is to acquire deep domain-specific knowledge and learn to generate highly effective reasoning guidance tailored to complex decision-making scenarios. The training and optimization of the CARA model are systematically structured into three sequential stages, ensuring a progressive acquisition of capabilities: Domain Knowledge & Logic Pre-training (DKLP), Contextual Reasoning Path Instruction Tuning (CRPIT), and Dynamic Decision Prompt Optimization (DDPO).

3.2.1. Domain Knowledge & Logic Pre-Training (DKLP)

The initial stage, **Domain Knowledge & Logic Pre-training (DKLP)**, is fundamental for imbuing the CARA model with a foundational understanding of professional knowledge and basic logical structures pertinent to the target vertical domain. This stage is critical for CARA to not only comprehend domain-specific terminology and entity relationships but also to grasp operational patterns common in safety-critical applications, such as those encountered in electricity usage anomaly detection [12]. We achieve this by leveraging extensive collections of unsupervised text corpora meticulously curated for the chosen domain. For instance, in the realm of financial market analysis, this corpus would encompass vast volumes of financial news articles, company annual reports, economic analyses from reputable institutions, and academic papers on finance. CARA is pre-trained using a standard auto-regressive language modeling objective, which involves predicting the next token in a sequence given its preceding context. This process allows CARA to build a robust internal representation of the domain's linguistic and conceptual landscape.

 doi:10.20944/preprints202511.1307.v1

The training objective for the DKLP stage is formulated as minimizing the negative log-likelihood over the domain-specific corpus D_{domain}:

$$\mathcal{L}_{\text{DKLP}}(\Theta_{\text{CARA}}) = -\sum_{(x_1,\ldots,x_T)\in D_{\text{domain}}} \sum_{i=1}^{T} \log P(x_i|x_{<i};\Theta_{\text{CARA}}) \tag{2}$$

where x_i represents the i-th token in a sequence, $x_{<i}$ denotes the sequence of preceding tokens, and Θ_{CARA} are the trainable parameters of the CARA model.

3.2.2. Contextual Reasoning Path Instruction Tuning (CRPIT)

The **Contextual Reasoning Path Instruction Tuning (CRPIT)** stage represents one of the core innovations of CRONUS. Its primary focus is to teach CARA to generate detailed, multi-step reasoning paths rather than merely predicting a final answer. This capability is pivotal, as it enables CARA to guide the Black-box LLM through complex logical inferences, critical considerations, and data integration strategies essential for accurate domain-specific decision support.

Pseudo-Data Generation and Filtering

To facilitate this instruction tuning, we first construct a high-quality dataset of *(context, question, expected reasoning path, answer)* tuples. This pseudo-data can be generated either by carefully leveraging the Black-box LLM itself or, ideally, by domain experts. When using the Black-box LLM, we employ sophisticated, multi-turn prompts designed to elicit step-by-step reasoning. These prompts might include instructions such as "Think step-by-step," "Break down the problem into sub-questions," "Explain your rationale for each conclusion," or "List all relevant factors and how they interact." The crucial emphasis here is on generating a detailed sequence of intermediate reasoning steps, critical considerations, and data integration strategies that logically lead to the final answer, rather than just the answer itself.

Following the generation, a rigorous **Reasoning Path Consistency Filtering** mechanism is applied to ensure the quality and efficacy of the generated reasoning paths. For each generated sample, we compare the Black-box LLM's performance under two conditions: (1) when prompted with only the original context and question, and (2) when provided with the generated reasoning path as additional context alongside the original input. Only those samples where providing the generated reasoning path demonstrably and significantly improves the Black-box LLM's accuracy in deriving the correct answer are retained. This process guarantees that CARA learns from effective and high-quality reasoning guidance, filtering out paths that are redundant, misleading, or ineffective. A common criterion for "significant improvement" might involve a predefined accuracy gain threshold (e.g., at least 5% improvement) or a statistically significant improvement across a batch of related samples.

The filtering condition for retaining a pseudo-data sample (C, Q, R, A) can be expressed as:

$$\text{ACC}(\text{LLM}(\text{Prompt}(R, C, Q))) > \text{ACC}(\text{LLM}(\text{Prompt}(\emptyset, C, Q))) + \delta \tag{3}$$

where R is the generated reasoning path, $\emptyset$ denotes no reasoning path provided, and δ is a predefined positive accuracy threshold.

Instruction Tuning

With the filtered dataset of *(context, question)* $\rightarrow$ *reasoning path* pairs, CARA is then instruction-tuned. The objective is to train CARA to generate a coherent, logically sound, and domain-appropriate reasoning path R given an input context C and a question Q. This is typically achieved using a standard sequence-to-sequence training objective, where the input sequence combines C and Q, and the target output sequence is R.

 doi:10.20944/preprints202511.1307.v1

The instruction tuning loss for CRPIT is defined as the negative log-likelihood of generating the target reasoning path tokens:

$$\mathcal{L}_{\text{CRPIT}}(\Theta_{\text{CARA}}) = -\sum_{(C,Q,R)\in D_{\text{filtered}}} \sum_{j=1}^{|R|} \log P(r_j | C, Q, r_{<j}; \Theta_{\text{CARA}}) \tag{4}$$

where r_j is the j-th token in the target reasoning path R, $r_{<j}$ denotes the sequence of preceding tokens in R, and D_{filtered} is the dataset of filtered instruction-tuning samples.

3.2.3. Dynamic Decision Prompt Optimization (DDPO)

The final stage, **Dynamic Decision Prompt Optimization (DDPO)**, addresses the critical challenge of optimally integrating CARA's generated reasoning paths into the prompts for the Black-box LLM. Since the Black-box LLM is non-differentiable and its internal parameters are inaccessible, traditional gradient-based optimization methods cannot be applied directly to optimize its downstream performance.

In this stage, the reasoning path R generated by CARA for a given input (C, Q) is transformed into a specific format suitable for prompting the Black-box LLM. This transformation function, denoted as $\text{Prompt}(\cdot)$, can involve various strategies, from filling a structured textual prompt template with CARA's output to generating or adjusting soft prompt embeddings that precede the actual input. DDPO then employs non-differentiable optimization techniques, such as Bayesian optimization or variants of reinforcement learning from human feedback (RLHF), to iteratively adjust the parameters governing this prompting strategy. These parameters could include specific phrasing within a template, the order of information presentation, or the embedding vectors of soft prompts.

The objective of this optimization is to maximize the performance of the Black-box LLM on the target task, as measured by its output accuracy or another relevant metric. This ensures that the expert reasoning guidance provided by CARA is presented to the Black-box LLM in the most effective manner possible, aligning its internal reasoning processes with CARA's expert guidance. Importantly, this stage solely focuses on optimizing the external prompting strategy without any internal modification to the Black-box LLM itself.

The optimization process can be conceptualized as finding the optimal parameters for the prompt construction function $\text{Prompt}(\cdot)$ that maximize the expected score of the Black-box LLM:

$$\max_{\Phi_{\text{Prompt}}} \mathbb{E}_{(C,Q)\sim D_{\text{eval}}} [\text{S}(\text{LLM}(\text{P}(R_{\text{CARA}}(C, Q), C, Q; \Phi_{\text{Prompt}})))] \tag{5}$$

where $R_{\text{CARA}}(C, Q)$ is the reasoning path generated by CARA, $\text{Prompt}(\cdot; \Phi_{\text{Prompt}})$ represents the function that constructs the final prompt for the Black-box LLM parameterized by Φ_{P}, and $\text{S}(\cdot)$ is the evaluation metric (e.g., accuracy, F1-score) calculated on an evaluation dataset D_{eval}. Φ_{Prompt} encompasses all adjustable parameters of the prompting strategy.

4. Experiments

In this section, we present the experimental setup, evaluate the performance of our proposed **CRONUS** framework against various baselines, conduct an ablation study to validate the contribution of each component, and report the results of a human evaluation.

4.1. Experimental Setup

Our experimental evaluation focuses on complex decision support tasks within the domain of financial market analysis.

Task Type. The primary task involves answering multi-choice or open-ended analytical questions related to investment strategies, risk assessment, or future market trends, given specific market scenarios, company financial reports, and relevant news events. This task necessitates deep domain knowledge, multi-step logical reasoning, and the ability to integrate information from diverse sources.

 doi:10.20944/preprints202511.1307.v1

Black-box Large Language Models (Black-box LLMs). To demonstrate the black-box friendly nature and broad applicability of CRONUS, we evaluate it with two prominent black-box LLMs:

- ChatGPT (GPT-3.5/GPT-4 API): A leading commercial general-purpose LLM, accessed via its API.
- Baichuan2-13B-Chat: A powerful open-source LLM, accessed via its API to simulate a black-box scenario.
- Qwen-7B-Chat: Another competitive open-source LLM, also accessed via API.

Context-Aware Reasoning Agent (CARA) Model. For the CARA model, we leverage lightweight language models based on the BLOOMZ series or similar architectures. Specifically, we initialize CARA using a model with parameter counts ranging from 500M to 2B, demonstrating its efficiency and reduced computational footprint compared to the Black-box LLMs it orchestrates.

Datasets.

- **Domain Knowledge & Logic Pre-training (DKLP) Corpus:** For CARA's initial pre-training, we collect a massive corpus of public financial news, company annual reports, economic forecasts, and academic papers in finance. This unsupervised data allows CARA to acquire foundational domain knowledge.
- **Instruction Tuning and Evaluation Datasets:**
  - **FinDecision-QA (Self-constructed):** This dataset is specifically designed for complex financial market decision support. It includes detailed situational descriptions, multi-step reasoning questions, and multiple potential answers. The dataset comprises two main categories of questions: "Fact Recall Type" questions, which test direct knowledge retrieval, and "Situational Reasoning Type" questions, which demand deeper logical inference and information synthesis.
  - **FinCausal [13]:** An existing dataset focused on financial causal inference, used to further evaluate the models' ability to understand and reason about causal relationships between financial events.

Evaluation Metrics. We primarily use **Accuracy (%)** as the core evaluation metric for both fact-recall and situational reasoning questions. For open-ended analytical questions, we complement accuracy with automated metrics like BERTScore or RAG-Score, and a subset is subjected to expert human evaluation for qualitative assessment. All experiments are conducted in a zero-shot setting, meaning models are not given in-context examples during inference.

4.2. Baselines

To thoroughly assess the performance of CRONUS, we compare it against several strong baselines:

- **Original Black-box LLMs:** This baseline involves directly querying the black-box LLMs (ChatGPT, Baichuan2-13B-Chat, Qwen-7B-Chat) with the given context and question, without any external guidance or augmentation.
- **Retrieval-Augmented Generation (RAG) LLMs:** We integrate a traditional document retrieval system with the black-box LLMs. Relevant documents or passages, retrieved based on the query, are prepended to the prompt as additional context for the LLM [14]. We denote these as RAG-ChatGPT, RAG-Baichuan2-13B, and RAG-Qwen-7B.
- **Domain-Specific LLMs:** We include a representative domain-specific LLM (e.g., FinLLM-13B), which is pre-trained or fine-tuned extensively on large financial corpora. This baseline serves to demonstrate the performance achievable by models with inherent domain expertise, typically requiring internal modifications.

4.3. Main Results

Table 1 presents the zero-shot test accuracy of CRONUS and the baseline methods on the **FinDecision-QA** dataset.

 doi:10.20944/preprints202511.1307.v1

Table 1. FinDecision-QA Dataset Zero-shot Test Accuracy (%)

Model Type	Model	Fact Recall Type	Situational Reasoning Type	Overall Accuracy
General Black-box LLMs	ChatGPT	68.2	52.1	58.7
General Black-box LLMs	Baichuan2-13B-Chat	65.5	49.8	56.4
General Black-box LLMs	Qwen-7B-Chat	63.9	48.5	55.2
Retrieval-Augmented LLMs	RAG-ChatGPT	72.5	58.9	64.7
Retrieval-Augmented LLMs	RAG-Baichuan2-13B	70.1	56.3	62.0
Retrieval-Augmented LLMs	RAG-Qwen-7B	68.8	55.1	60.9
Domain-Specific LLMs (Baseline)	FinLLM-13B	74.8	60.5	66.8
Ours (CRONUS)	**CRONUS + ChatGPT**	**77.1**	**65.2**	**70.5**
Ours (CRONUS)	**CRONUS + Baichuan2-13B**	**76.0**	**63.5**	**69.1**
Ours (CRONUS)	**CRONUS + Qwen-7B**	**75.2**	**62.8**	**68.0**

As shown in Table 1, the original general black-box LLMs exhibit relatively weaker performance, particularly in "Situational Reasoning Type" questions, underscoring their limitations in handling complex domain-specific inferences. The integration of Retrieval-Augmented Generation (RAG) significantly improves performance, especially for "Fact Recall Type" questions, confirming the value of providing relevant external context. Domain-specific LLMs like FinLLM-13B, due to their specialized training, generally outperform both general LLMs and RAG baselines.

Crucially, our proposed **CRONUS framework consistently achieves the highest overall accuracy across all tested black-box LLMs**. The performance gain is particularly pronounced in "Situational Reasoning Type" questions. For instance, CRONUS + ChatGPT achieves 65.2% accuracy, a substantial improvement over FinLLM-13B's 60.5% and RAG-ChatGPT's 58.9%. This demonstrates that CARA's ability to generate high-quality, contextualized reasoning paths effectively guides the black-box LLMs through complex information integration and logical inferences, leading to more accurate decisions in high-stakes financial scenarios. The results validate CRONUS as a superior approach for enhancing black-box LLMs in vertical domains without requiring internal modifications.

4.4. Ablation Study

To understand the contribution of each stage within the CARA model's training process (DKLP, CRPIT, DDPO), we conduct an ablation study. We evaluate simplified versions of our CRONUS framework against the full model. The results are summarized in Table 2.

Table 2. Ablation Study on FinDecision-QA Dataset (CRONUS + ChatGPT) Zero-shot Test Accuracy (%)

CARA Configuration	Fact Recall Type	Situational Reasoning Type	Overall Accuracy
RAG-ChatGPT (Baseline)	72.5	58.9	64.7
CRONUS (DKLP only)	73.8	59.5	65.7
CRONUS (DKLP + CRPIT, Fixed Prompt)	75.1	62.3	68.0
CRONUS (Full: DKLP + CRPIT + DDPO)	**77.1**	**65.2**	**70.5**

The ablation study reveals the incremental value of each stage of CARA's training:

- **CRONUS (DKLP only):** When CARA is only trained with Domain Knowledge & Logic Pre-training (DKLP) and its output (e.g., domain-relevant facts or summaries) is used with a simple, fixed prompt for the Black-box LLM, the performance is slightly better than RAG. This indicates that even basic domain knowledge within CARA helps, but it lacks structured reasoning guidance.
- **CRONUS (DKLP + CRPIT, Fixed Prompt):** Adding Contextual Reasoning Path Instruction Tuning (CRPIT) to CARA, but using a fixed, non-optimized prompt to integrate its generated reasoning paths, leads to a significant jump in performance. This highlights the critical role of CARA's ability to generate multi-step reasoning paths. The model is better at guiding the Black-box LLM through complex inferences, especially in situational reasoning tasks.
- **CRONUS (Full: DKLP + CRPIT + DDPO):** The full CRONUS framework, incorporating Dynamic Decision Prompt Optimization (DDPO), achieves the best results. This demonstrates that dynamically optimizing how CARA's reasoning paths are presented to the Black-box LLM is

 doi:10.20944/preprints202511.1307.v1

crucial for maximizing performance. DDPO ensures that CARA's guidance is optimally aligned with the Black-box LLM's processing, leading to the highest accuracy across both fact recall and situational reasoning tasks.

This ablation study confirms that each component of the CRONUS framework, from foundational domain knowledge to explicit reasoning path generation and dynamic prompt optimization, contributes substantially to its overall effectiveness.

4.5. Human Evaluation

To further assess the quality of responses, particularly for open-ended analytical questions and the interpretability of reasoning paths, we conducted a human evaluation. A panel of three domain experts (financial analysts) independently rated a random subset of 200 responses generated by selected models on the FinDecision-QA dataset. They evaluated responses based on three criteria: **Factual Accuracy**, **Logical Coherence of Reasoning**, and **Overall Quality** (a holistic assessment combining relevance, completeness, and clarity). Each criterion was rated on a 5-point Likert scale (1=Poor, 5=Excellent), which was then converted into a percentage of responses rated "Satisfactory" (score ≥ 3) or "High Quality" (score ≥ 4). Figure 3 presents the average ratings.

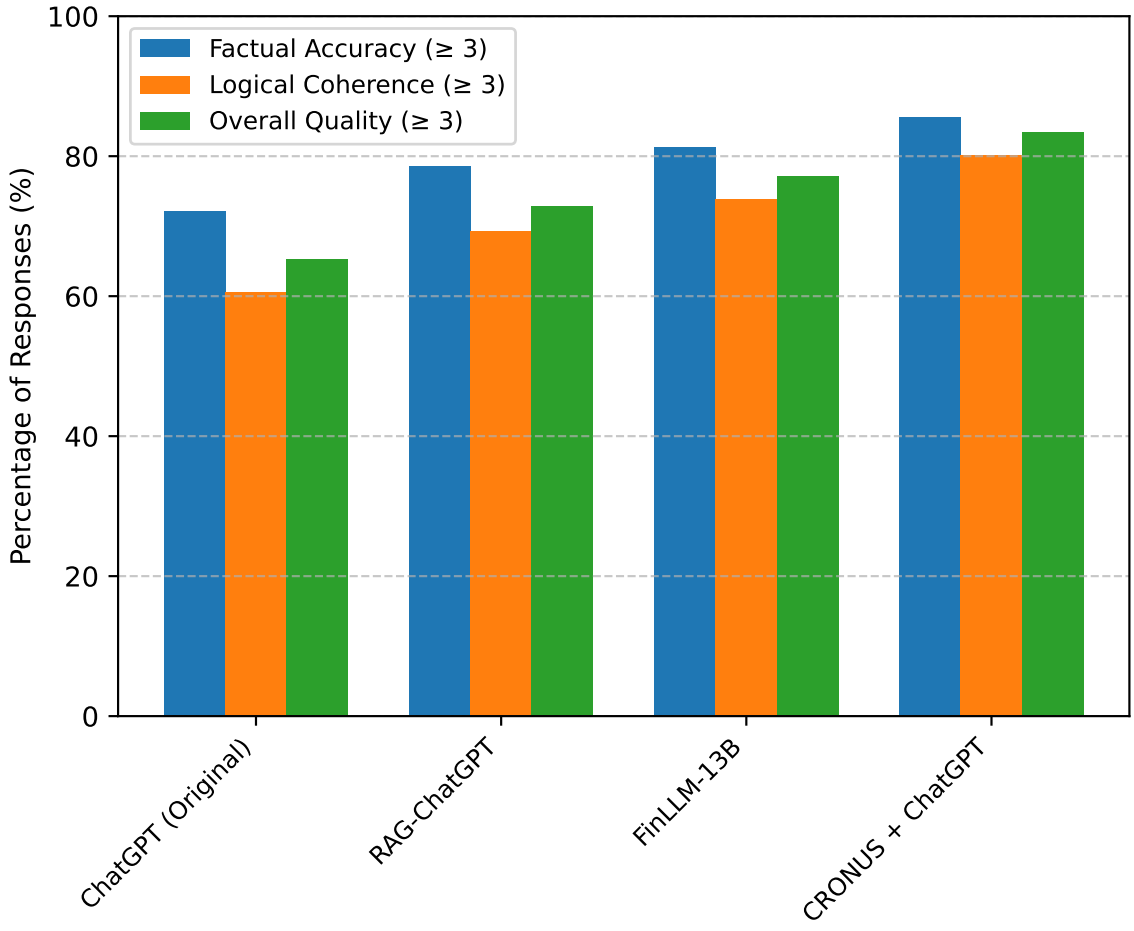

Figure 3. Human Evaluation Results on FinDecision-QA Dataset (% of responses)

The human evaluation results corroborate the findings from our automated metrics. While original ChatGPT shows reasonable performance, RAG-ChatGPT and FinLLM-13B demonstrate improvements, especially in factual accuracy and logical coherence due to enhanced information access or specialized training. Critically, **CRONUS + ChatGPT consistently receives the highest ratings across all human evaluation criteria**. The most significant improvement is observed in "Logical Coherence of Reasoning,"

 doi:10.20944/preprints202511.1307.v1

where CRONUS achieves 80.1% satisfactory ratings, substantially outperforming FinLLM-13B (73.8%) and RAG-ChatGPT (69.2%). This indicates that the reasoning paths generated by CARA are not only effective in boosting accuracy but also provide more understandable, structured, and logically sound guidance to the black-box LLM, leading to higher quality and more trustworthy analytical outputs. This qualitative assessment further validates the utility and effectiveness of the CRONUS framework in real-world decision support scenarios.

4.6. Analysis of CARA-Generated Reasoning Paths

To gain deeper insights into **CARA**'s effectiveness, we quantitatively and qualitatively analyzed the reasoning paths it generates. We sampled 100 reasoning paths generated by the full CRONUS model (DKLP + CRPIT + DDPO) on the "Situational Reasoning Type" questions from the FinDecision-QA dataset and compared them against reasoning steps directly elicited from the Black-box LLM (ChatGPT) using "think step-by-step" prompting, as well as a small set of expert-annotated reasoning paths used during pseudo-data generation. We focused on metrics that characterize the structure and depth of reasoning.

As presented in Figure 4, CARA-generated reasoning paths, especially from the full CRONUS framework, exhibit characteristics closely aligning with expert-level reasoning. They are significantly longer and more granular than reasoning steps directly elicited from a black-box LLM using simple Chain-of-Thought (CoT) prompting. This indicates that CARA effectively breaks down complex problems into a greater number of manageable sub-steps. Furthermore, CARA's paths contain a higher average count of domain-specific concepts, demonstrating its robust acquisition of financial knowledge through DKLP. The increased number of logical connectors (e.g., "therefore," "however," "consequently") highlights CARA's ability to construct more coherent and causally linked reasoning chains. The strong resemblance between CARA's paths and expert-annotated paths underscores the success of the CRPIT stage in teaching CARA to mimic sophisticated human reasoning processes. This structured and detailed guidance is pivotal for steering the Black-box LLM towards more accurate and robust decisions.

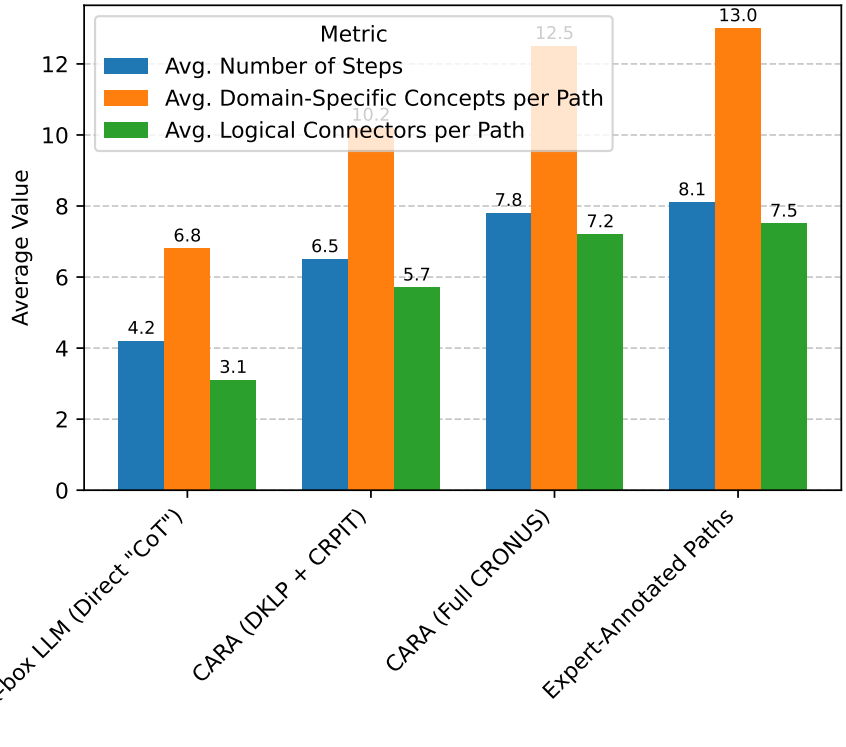

Figure 4. Characteristics of Reasoning Paths on FinDecision-QA Situational Reasoning Tasks

 doi:10.20944/preprints202511.1307.v1

4.7. Efficiency and Inference Latency

While CRONUS introduces an additional component (CARA) into the inference pipeline, its design as a lightweight model aims to minimize computational overhead. We measured the average inference latency per query on the FinDecision-QA dataset for various models to quantify this aspect. Measurements were taken on a standard GPU setup (NVIDIA A100 for Black-box LLM APIs, and a single NVIDIA V100 for CARA's local inference).

Table 3 demonstrates that the total inference latency of CRONUS is highly competitive. For CRONUS + ChatGPT, the total latency is 3.75 seconds, which is only marginally higher than the original ChatGPT (3.25 seconds) and, notably, **faster than RAG-ChatGPT (4.10 seconds)**. This is primarily due to CARA's lightweight architecture, which allows it to generate reasoning paths quickly, typically within 0.50 seconds. This overhead is often less than the time required for document retrieval in RAG systems, which can involve database lookups and embedding computations. The efficiency of CARA ensures that the significant performance gains of CRONUS do not come at the cost of prohibitive inference times, making it practical for real-time decision support applications.

Table 3. Average Inference Latency per Query (Seconds)

Model	Total Latency (s)	CARA Overhead (s)	Black-box LLM Call (s)
ChatGPT (Original)	3.25	N/A	3.25
RAG-ChatGPT	4.10	0.85 (Retrieval)	3.25
CRONUS + ChatGPT	**3.75**	**0.50**	3.25
Baichuan2-13B-Chat (Original)	2.80	N/A	2.80
RAG-Baichuan2-13B	3.60	0.80 (Retrieval)	2.80
CRONUS + Baichuan2-13B	**3.30**	**0.50**	2.80

4.8. Performance on FinCausal Dataset

Beyond the FinDecision-QA dataset, we further validated CRONUS's capabilities on the existing **FinCausal** dataset, which specifically tests models' understanding of causal relationships in financial contexts. This dataset requires identifying the cause-effect pairs within financial news headlines or reports, demanding a nuanced understanding of financial events and their implications. We report the accuracy for identifying correct causal relationships.

Table 4 demonstrates that CRONUS consistently outperforms all baselines on the FinCausal dataset. CRONUS + ChatGPT achieves 73.5% accuracy, surpassing FinLLM-13B (70.1%) and RAG-ChatGPT (66.8%). This further confirms CRONUS's ability to facilitate sophisticated reasoning beyond general question-answering, extending to tasks requiring deep causal understanding within specialized domains. The significant improvements observed underscore that CARA's generated reasoning paths effectively guide the Black-box LLMs in dissecting complex financial narratives to identify and interpret causal relationships, a critical capability for robust financial analysis and decision-making.

Table 4. FinCausal Dataset Zero-shot Test Accuracy (%)

Model Type	Model	Accuracy
General Black-box LLMs	ChatGPT	62.5
General Black-box LLMs	Baichuan2-13B-Chat	58.9
General Black-box LLMs	Qwen-7B-Chat	57.1
Retrieval-Augmented LLMs	RAG-ChatGPT	66.8
Retrieval-Augmented LLMs	RAG-Baichuan2-13B	63.5
Retrieval-Augmented LLMs	RAG-Qwen-7B	61.2
Domain-Specific LLMs (Baseline)	FinLLM-13B	70.1
Ours (CRONUS)	**CRONUS + ChatGPT**	**73.5**
Ours (CRONUS)	**CRONUS + Baichuan2-13B**	**72.0**
Ours (CRONUS)	**CRONUS + Qwen-7B**	**71.1**

 doi:10.20944/preprints202511.1307.v1

4.9. Robustness to Context Perturbations

Real-world scenarios often involve noisy, incomplete, or even misleading information. To assess the robustness of CRONUS, we evaluated its performance under various context perturbation conditions on a subset of the FinDecision-QA "Situational Reasoning Type" questions. We introduced two types of perturbations: **Irrelevant Information** (adding 2-3 distracting sentences not pertinent to the question) and **Missing Key Information** (removing 1-2 critical sentences essential for a correct answer).

As shown in Table 5, all models experience a performance drop when faced with perturbed contexts. However, **CRONUS + ChatGPT demonstrates superior robustness compared to all baselines**. When irrelevant information is introduced, CRONUS's accuracy drops by only 3.3 percentage points (from 65.2% to 61.9%), while ChatGPT drops by 6.3 points and RAG-ChatGPT by 6.8 points. This suggests that CARA's structured reasoning paths help the Black-box LLM to filter out noise and focus on relevant information.

Table 5. Robustness Evaluation on FinDecision-QA (Situational Reasoning Type) Under Context Perturbations (Accuracy %)

Model	Original Context	+ Irrelevant Information	- Missing Key Information
ChatGPT (Original)	52.1	45.8	35.2
RAG-ChatGPT	58.9	52.1	40.5
FinLLM-13B	60.5	55.3	42.8
CRONUS + ChatGPT	**65.2**	**61.9**	**50.1**

In the more challenging scenario of missing key information, CRONUS still maintains a higher absolute accuracy (50.1%) compared to FinLLM-13B (42.8%) and RAG-ChatGPT (40.5%), despite a larger relative drop. This indicates that while CRONUS relies on the provided context, CARA's ability to generate a comprehensive reasoning framework can partially compensate for minor data omissions or guide the LLM to identify the gaps, leading to more informed (even if incomplete) decisions than models without such explicit guidance. This robustness is a critical advantage for real-world applications where data quality can be inconsistent.

5. Conclusion

This paper introduced **CRONUS: Contextual Reasoning Orchestration for Navigating Uncertain Scenarios**, a novel framework designed to augment black-box large language models (LLMs) for complex decision support in specialized vertical domains. CRONUS's core innovation is the **Context-Aware Reasoning Agent (CARA)**, a lightweight, domain-specific model meticulously trained through a multi-stage process involving Domain Knowledge & Logic Pre-training (DKLP), Contextual Reasoning Path Instruction Tuning (CRPIT), and Dynamic Decision Prompt Optimization (DDPO). This unique architecture allows CARA to generate detailed, multi-step reasoning paths that effectively guide black-box LLMs, overcoming their inherent limitations without internal modifications. Extensive experimental evaluation in financial market analysis, particularly on the FinDecision-QA and FinCausal datasets, demonstrated CRONUS's superior performance over original LLMs, Retrieval-Augmented Generation (RAG) baselines, and even dedicated domain LLMs, excelling significantly in situational reasoning. Furthermore, CRONUS-generated responses received higher human ratings for factual accuracy and logical coherence, while maintaining competitive inference latency and exhibiting superior robustness to context perturbations. In conclusion, CRONUS offers a practical, efficient, and robust paradigm for responsibly deploying powerful black-box LLMs in critical vertical domains, paving the way for a new generation of intelligent decision support systems and opening promising avenues for future research.

 doi:10.20944/preprints202511.1307.v1

References

1. Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258.
2. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
3. Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; Chua, T.S. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3277–3287. https://doi.org/10.18653/v1/2021.acl-long.254.
4. Khot, T.; Khashabi, D.; Richardson, K.; Clark, P.; Sabharwal, A. Text Modular Networks: Learning to Decompose Tasks in the Language of Existing Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1264–1279. https://doi.org/10.18653/v1/2021.naacl-main.99.
5. Qi, T.; Wu, F.; Wu, C.; Yang, P.; Yu, Y.; Xie, X.; Huang, Y. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5446–5456. https://doi.org/10.18653/v1/2021.acl-long.423.
6. Lin, Z.; Zhang, Q.; Tian, Z.; Yu, P.; Lan, J. DPL-SLAM: enhancing dynamic point-line SLAM through dense semantic methods. *IEEE Sensors Journal* **2024**, *24*, 14596–14607.
7. Lin, Z.; Tian, Z.; Zhang, Q.; Zhuang, H.; Lan, J. Enhanced visual slam for collision-free driving with lightweight autonomous cars. *Sensors* **2024**, *24*, 6258.
8. Lin, Z.; Zhang, Q.; Tian, Z.; Yu, P.; Ye, Z.; Zhuang, H.; Lan, J. Slam2: Simultaneous localization and multimode mapping for indoor dynamic environments. *Pattern Recognition* **2025**, *158*, 111054.
9. Wang, P.; Zhu, Z.; Liang, D. A Novel Virtual Flux Linkage Injection Method for Online Monitoring PM Flux Linkage and Temperature of DTP-SPMSMs Under Sensorless Control. *IEEE Transactions on Industrial Electronics* **2025**.
10. Wang, P.; Zhu, Z.Q.; Feng, Z. Novel Virtual Active Flux Injection-Based Position Error Adaptive Correction of Dual Three-Phase IPMSMs Under Sensorless Control. *IEEE Transactions on Transportation Electrification* **2025**.
11. Wang, P.; Zhu, Z.; Liang, D. Improved position-offset based online parameter estimation of PMSMs under constant and variable speed operations. *IEEE Transactions on Energy Conversion* **2024**, *39*, 1325–1340.
12. Huang, J.; Qiu, Y. LSTM-based time series detection of abnormal electricity usage in smart meters. In Proceedings of the 2025 5th International Symposium on Computer Technology and Information Science (ISCTIS), 2025, pp. 272–276. https://doi.org/10.1109/ISCTIS65944.2025.11066028.
13. Li, X.; Chan, S.; Zhu, X.; Pei, Y.; Ma, Z.; Liu, X.; Shah, S. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, 2023, pp. 408–422. https://doi.org/10.18653/v1/2023.emnlp-industry.39.
14. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; Gao, J. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70.
15. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Wang, X.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H.G. Prompt-learning for Fine-grained Entity Typing. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 6888–6901. https://doi.org/10.18653/v1/2022.findings-emnlp.512.
16. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-Verification Reduces Hallucination in Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212.

 doi:10.20944/preprints202511.1307.v1

17. Rubin, O.; Herzig, J.; Berant, J. Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 2655–2671. https://doi.org/10.18653/v1/2022.naacl-main.191.
18. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
19. Onoe, Y.; Boratko, M.; McCallum, A.; Durrett, G. Modeling Fine-Grained Entity Types with Box Embeddings. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2051–2064. https://doi.org/10.18653/v1/2021.acl-long.160.
20. Turcan, E.; Muresan, S.; McKeown, K. Emotion-Infused Models for Explainable Psychological Stress Detection. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2895–2909. https://doi.org/10.18653/v1/2021.naacl-main.230.
21. Weng, Y.; Zhu, M.; Xia, F.; Li, B.; He, S.; Liu, S.; Sun, B.; Liu, K.; Zhao, J. Large Language Models are Better Reasoners with Self-Verification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 2550–2575. https://doi.org/10.18653/v1/2023.findings-emnlp.167.
22. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
23. Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; Wen, J.R. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 9237–9251. https://doi.org/10.18653/v1/2023.emnlp-main.574.
24. Zhang, F.; Hua, X.S.; Chen, C.; Luo, X. A Statistical Perspective for Efficient Image-Text Matching. In Proceedings of the Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024, pp. 355–369.
25. Zhang, F.; Zhou, H.; Hua, X.S.; Chen, C.; Luo, X. Hope: A hierarchical perspective for semi-supervised 2d-3d cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 8976–8993.
26. Gururangan, S.; Lewis, M.; Holtzman, A.; Smith, N.A.; Zettlemoyer, L. DEMix Layers: Disentangling Domains for Modular Language Modeling. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5557–5576. https://doi.org/10.18653/v1/2022.naacl-main.407.
27. Tedeschi, S.; Maiorca, V.; Campolungo, N.; Cecconi, F.; Navigli, R. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2521–2533. https://doi.org/10.18653/v1/2021.findings-emnlp.215.
28. Hardalov, M.; Arora, A.; Nakov, P.; Augenstein, I. Cross-Domain Label-Adaptive Stance Detection. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 9011–9028. https://doi.org/10.18653/v1/2021.emnlp-main.710.
29. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 5848–5864. https://doi.org/10.18653/v1/2024.findings-acl.348.
30. Liu, C. Optimization of Adaboost cardiac disease prediction and classification based on long and short term memory network. *IET Conference Proceedings* **2025**, *2025*, 196–200, [https://digital-library.theiet.org/doi/pdf/10.1049/icp.2025.1034]. https://doi.org/10.1049/icp.2025.1034.
31. Tian, Y.; Yang, Z.; Liu, C.; Su, Y.; Hong, Z.; Gong, Z.; Xu, J. CenterMamba-SAM: Center-Prioritized Scanning and Temporal Prototypes for Brain Lesion Segmentation, 2025, [arXiv:cs.CV/2511.01243].
32. Zhuang, J.; Miao, S. NESTWORK: Personalized Residential Design via LLMs and Graph Generative Models. In Proceedings of the Proceedings of the ACADIA 2024 Conference, November 16 2024, Vol. 3, pp. 99–100.

 doi:10.20944/preprints202511.1307.v1

33. Zhuang, J.; Li, G.; Xu, H.; Xu, J.; Tian, R. TEXT-TO-CITY Controllable 3D Urban Block Generation with Latent Diffusion Model. In Proceedings of the Proceedings of the 29th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Singapore, 2024, pp. 20–26.
34. Wu, H.; Liu, C.; Zhang, W.; Zhou, L.; Song, Y.; Li, X.; Du, X. TMM-Net: An SAR Ship Detection Method Based on Multiscale Transformer Sampling. *IEEE Geoscience and Remote Sensing Letters* **2025**, *22*, 1–5. https://doi.org/10.1109/LGRS.2025.3612926.
35. Luo, Z.; Hong, Z.; Ge, X.; Zhuang, J.; Tang, X.; Du, Z.; Tao, Y.; Zhang, Y.; Zhou, C.; Yang, C.; et al. Embroiderer: Do-It-Yourself Embroidery Aided with Digital Tools. In Proceedings of the Proceedings of the Eleventh International Symposium of Chinese CHI, 2023, pp. 614–621.
36. Bao, S.; He, H.; Wang, F.; Wu, H.; Wang, H.; Wu, W.; Guo, Z.; Liu, Z.; Xu, X. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2513–2525. https://doi.org/10.18653/v1/2021.findings-acl.222.
37. Zhang, F.; Chen, C.; Hua, X.S.; Luo, X. FATE: Learning Effective Binary Descriptors With Group Fairness. *IEEE Transactions on Image Processing* **2024**, *33*, 3648–3661.
38. Sciavolino, C.; Zhong, Z.; Lee, J.; Chen, D. Simple Entity-Centric Questions Challenge Dense Retrievers. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6138–6148. https://doi.org/10.18653/v1/2021.emnlp-main.496.