

Querying Large Automotive Software Models: Agentic vs. Direct LLM Approaches

Lukasz Mazur

*Chair of Robotics, Artificial Intelligence and Real-Time Systems
Technical University of Munich
Munich, Germany
lukasz.mazur@tum.de*

Nenad Petrovic

*Chair of Robotics, Artificial Intelligence and Real-Time Systems
Technical University of Munich
Munich, Germany
nenad.petrovic@tum.de*

James Pontes Miranda

*Software and Systems Engineering
Université Paris-Saclay
CEA List
Palaiseau, France
james.pontesmiranda@cea.fr*

Ansgar Radermacher

*Software and Systems Engineering
Université Paris-Saclay
CEA List
Palaiseau, France
ansgar.radermacher@cea.fr*

Robert Rasche

*Real-Time Systems, Resources,
Model Based Software Eng.
Tensor embedded GmbH
Pollenfeld, Germany
robert.rasche@tensor.gmbh*

Alois Knoll

*Chair of Robotics, Artificial Intelligence and Real-Time Systems
Technical University of Munich
Munich, Germany
knoll@in.tum.de*

Abstract—Large language models (LLMs) offer new opportunities for interacting with complex software artifacts, such as software models, through natural language. They present especially promising benefits for large software models that are difficult to grasp in their entirety, making traditional interaction and analysis approaches challenging. This paper investigates two approaches for leveraging LLMs to answer questions over software models: direct prompting, where the whole software model is provided in the context, and an agentic approach combining LLM-based agents with general-purpose file access tools. We evaluate these approaches using an Ecore metamodel designed for timing analysis and software optimization in automotive and embedded domains. Our findings show that while the agentic approach achieves accuracy comparable to direct prompting, it is significantly more efficient in terms of token usage. This efficiency makes the agentic approach particularly suitable for the automotive industry, where the large size of software models makes direct prompting infeasible, establishing LLM agents as not just a practical alternative but the only viable solution. Notably, the evaluation was conducted using small LLMs, which are more feasible to be executed locally — an essential advantage for meeting strict requirements around privacy, intellectual property protection, and regulatory compliance. Future work will investigate software models in diverse formats, explore more complex agent architectures, and extend agentic workflows to support not only querying but also modification of software models.

Index Terms—Software Model Querying, Agentic LLMs, Automotive Software, LLMs for Model-Driven Engineering

I. INTRODUCTION

Software modeling is crucial in modern software engineering, particularly in safety-critical domains like automotive

This work has received funding from the European Chips Joint Undertaking under Framework Partnership Agreement No 101139789 (HAL4SDV) including the national funding from the German Federal Ministry of Education and Research (BMBF) under grant number 16MEE0047K. The responsibility for the content of this publication lies with the authors.

development, where precise models ensure compliance and reliability [1]. However, the increasing size and complexity of these models pose significant challenges for engineers in terms of access, understanding, and manipulation [2].

Recent advancements in large language models (LLMs) offer promising avenues to assist with complex software models, leveraging their capabilities in natural language understanding and code generation [3]. Despite this potential, the direct application of LLMs to structured software models, especially with formal languages like PlantUML [4], remains largely unexplored [5]. Current research primarily focuses on unstructured text or code, overlooking the unique challenges of integrating LLMs with structured models, including format compatibility, semantic precision, and consistency. Furthermore, the efficacy of different LLM architectures, such as agentic workflows versus direct prompting, in understanding and reasoning about software models is unclear.

This study investigates and compares agentic LLM workflows and direct full-context prompting for querying software models. We evaluate these approaches on specific, concrete questions about model structure and semantics, using a representative automotive software model.

Our contributions are twofold: an empirical comparison of LLM-based agentic and direct prompting for software model question answering, and insights into the feasibility and limitations of current LLM capabilities in this context. These findings will inform future research on advanced interactions with software models in model-driven engineering.

The paper is structured as follows: Section 2 reviews related work. Section 3 details the methodology. Section 4 presents results and analysis. Section 5 discusses findings. Section 6 addresses threats to validity, and Section 7 concludes with future work.

II. RELATED WORK

A. LLMs in Software and Model-Driven Engineering

LLMs have increasingly been applied in various software engineering tasks, including code generation, bug detection, documentation, and summarization. Their ability to understand and generate natural language alongside programming languages makes them valuable assistants for developers across the software development lifecycle [3].

In model-driven engineering (MDE), LLMs are increasingly being explored for tasks such as model generation, validation, and transformation, aiming to improve automation and reduce manual effort [3]. While software models offer high-level abstractions of system designs, their structured and formal nature poses challenges for natural language models, requiring specialized input representations and reasoning strategies [6]. A variety of modeling standards are used across domains, including UML [7] and Eclipse Modeling Framework (EMF) [8] in general software engineering, and AUTOSAR [9] in the automotive industry. These standards provide precise specifications of structure and behavior, which are critical in safety-critical domains governed by standards like ISO 26262, highlighting the growing need to integrate AI-assisted techniques into rigorous development processes [10].

Traditional software modeling tools provide querying, validation, and transformation capabilities to manage these complex models. EMF and specialized automotive toolchains support model editing and consistency checks [10] [11]. The integration of LLMs into such environments has the potential to enhance these tools by enabling natural language querying, automated documentation generation, and intelligent assistance during modeling. This intersection of LLM technology with established software modeling ecosystems remains an active area of exploration [6].

B. LLM Integration with Structured Software Artifacts

Interfacing LLMs with structured software artifacts such as models, codebases, or graphs requires specialized architectural approaches. Unlike plain text, software models are inherently structured, hierarchical, and often constrained by formal metamodels, which challenges traditional LLM input paradigms [12].

One prominent technique to address these challenges is Retrieval-Augmented Generation (RAG), where external knowledge bases are queried to retrieve relevant fragments that augment the LLM's context. This approach helps mitigate token-length constraints and context fragmentation, enabling the LLM to focus on pertinent information without being overwhelmed by the entire knowledge base [13].

Beyond retrieval-based methods, tool-augmented LLM architectures enable more sophisticated workflows. These agents can perform iterative reasoning, plan multi-step queries, and invoke external tools or APIs for accessing external data sources. Agentic approaches combine LLM capabilities with domain-specific logic and memory management, leading to improved performance on complex tasks [14] [15]. As tool

usage gains popularity, emerging standards such as the Model Context Protocol (MCP) [16] and Agent2Agent Protocol (A2A) [17] are being developed to support scalable and interoperable agent ecosystems. MCP standardizes how agents connect to tools, APIs, and resources through structured inputs and outputs, while A2A facilitates dynamic, multimodal communication between agents, enabling collaboration, delegation, and shared task management.

The balance between single-shot prompting and multi-turn agentic workflows also influences system capabilities. While single-shot prompts are simpler and faster, agentic workflows provide enhanced reasoning, planning, and verification, but come with higher computational costs and increased latency [18] [19]. The choice between these architectures typically depends on the task's complexity and how well the agent's available tools align with the task's abstraction level.

C. Evaluation of Free-Text LLM Answers

Evaluating open-ended natural language answers from LLMs is challenging, especially when accuracy, semantic alignment, and factual grounding are essential. Traditional lexical matching metrics that focus on the overlap of n-grams (word sequences) between two texts, like BLEU and ROUGE, often fall short when evaluating the quality of responses to structured or knowledge-intensive queries because they primarily assess surface-level word sequence commonality [20]. While semantic similarity methods, such as bi-encoder and cross-encoder models, better capture meaning beyond lexical overlap, they still struggle with evaluating factual accuracy, logical consistency, and adherence to task-specific constraints in complex scenarios [21].

LLM-as-a-Judge methods have recently emerged as a flexible and scalable evaluation paradigm, prompting LLMs to assess the correctness, completeness, and relevance of generated answers using their own domain understanding. However, these models may still hallucinate or misjudge, limiting their reliability [22]. Human evaluation remains essential for ensuring validity, particularly for complex reasoning tasks, but it is resource-intensive and often requires expert reviewers to assess nuanced correctness, adherence to software engineering principles, and practical relevance [23].

III. METHODOLOGY

A. Research Questions

This study investigates how LLMs and agent-based architectures perform on question-answering tasks over complex software models, comparing direct full-context prompting with agentic retrieval using external tools.

RQ1: How does the accuracy of an LLM-based agent that retrieves and processes parts of a large software model using external tools compare to the accuracy of a reference LLM that processes the entire model within its context window when answering model-related questions?

This question targets the core trade-off between full-context access and modular, tool-assisted access. It aims to evaluate whether an LLM-based agent can still produce accurate and

useful answers to domain-relevant questions despite never having a complete view of the software model at once. The answer to this question has significant implications for the feasibility of using LLM agents in real-world automotive workflows where models may be large, distributed, and subject to change.

RQ2: How does the token consumption of an LLM-based agent that incrementally accesses a large software model via tool support compare to that of an LLM processing the entire model in a single pass?

This question addresses the efficiency of the two approaches in terms of token usage, which is especially critical when deploying smaller LLMs with limited context windows. Smaller models are often the only viable option for local, on-premise deployment - an increasingly important requirement in the automotive industry due to strict privacy constraints, intellectual property protection, and regulatory compliance. By analyzing token consumption, we aim to understand whether agent-based strategies enable smaller, locally hosted models to work effectively with large-scale software artifacts.

B. Direct Prompting and Agent-Based Architecture

To evaluate the ability of LLMs to answer questions about complex software models, we compare two architectures: a direct LLM call serving as a reference, and a tool-augmented agent architecture. Both setups are illustrated in Fig. 1, with example responses shown in Fig. 2 to demonstrate how each approach processes and answers a question.

1) *Reference: Direct Full-Context Prompting:* The reference architecture represents a simple, baseline setup where the entire software model is placed directly into the LLM's context window. The LLM is asked to answer questions in a zero-shot manner, with no examples or intermediate reasoning steps. This approach assumes the model fits into the context window of the evaluated LLM - an increasingly limiting factor for larger models, motivating the agentic approach.

The LLM prompt in this setup utilizes the Ecore format [8] for the input model, primarily because it's the native format of the tool used in our experiments. This choice streamlines the process by avoiding the need to translate the software model, even though formats like PlantUML or Mermaid [24] could also represent class relationships. Ecore, however, provides a more detailed and inherently tool-compatible representation. Consequently, this XML-based format requires more tokens to describe the software model than simpler, text-based syntaxes. Additionally, the prompt explicitly directs the LLM to analyze class relationships thoroughly and to rely exclusively on the provided content, assuming it is a complete and self-contained definition.

2) *ReAct Architecture: Agent with File Tools:* The second approach uses a tool-augmented agent architecture based on the ReAct framework [25], where the LLM can choose to issue tool calls if it determines that the answer cannot be derived directly. The context window initially contains only the path to the directory with the software model files, not the model

content itself. This setup allows the agent to selectively retrieve and reason over relevant parts of the model on demand.

The agent is equipped with a set of tools adapted from the SWE-agent framework [26], which simulates an IDE-like interaction by exposing a windowed view over source files. Table I lists the tools available to the agent.

Although some tools (`list_directory`, `find_file`, and `search_directory`) are more relevant for models distributed across multiple files, they were retained for consistency and to enable the reuse of the agent with other software model datasets. For this experiment, all interactions involved a single file.

The windowing parameters for file tools were selected empirically to balance efficiency and context completeness. A window size of 50 lines with an overlap of 2 lines was used. This configuration proved effective for the selected software model, as most type definitions could be captured within one or two windows, thereby minimizing the need for extensive scrolling. At the same time, the window size was kept small enough to maintain low token usage and avoid exceeding the LLM's context limits during multi-step reasoning.

The agent was allowed up to 100 iterations per question to avoid infinite loops while preserving enough flexibility for complex queries.

The agent prompt followed a similar structure to the reference setup, stating that the software model is in Ecore format and emphasizing that the model is complete.

C. LLM Selection and Configuration

This study evaluates LLMs that support tool usage - a capability essential for agent-based workflows. While direct prompting can, in principle, use any LLM, we employed the same tool-capable models for both approaches to ensure a fair comparison. In line with the privacy and deployment constraints typical in the automotive industry, we concentrated on smaller LLMs that are more feasible to run locally or on secure, private infrastructure. All selected models belong to the latest two generations of LLM development, offering enhanced reasoning capabilities and extended context lengths. We selected GPT-4o mini, GPT-4.1 mini, and o4-mini to represent OpenAI's current lightweight models with progressively improving architecture and performance characteristics, with o4-mini offering the strongest reasoning performance within this group. To broaden the evaluation, we also included Gemini 2.5 Flash Preview, a fast, tool-capable model from a different vendor, to assess cross-vendor performance and introduce architectural diversity. Table II summarizes the selected models.

In all agent configurations, the temperature parameter was set to 0.0 to ensure deterministic behavior and maximize answer reproducibility. The only exception is o4-mini, where the temperature is fixed at 1.0.

D. Question Dataset Design

As no existing datasets are tailored for evaluating LLMs on software model understanding, we constructed a small, custom dataset based on an open-source metamodel in Ecore format

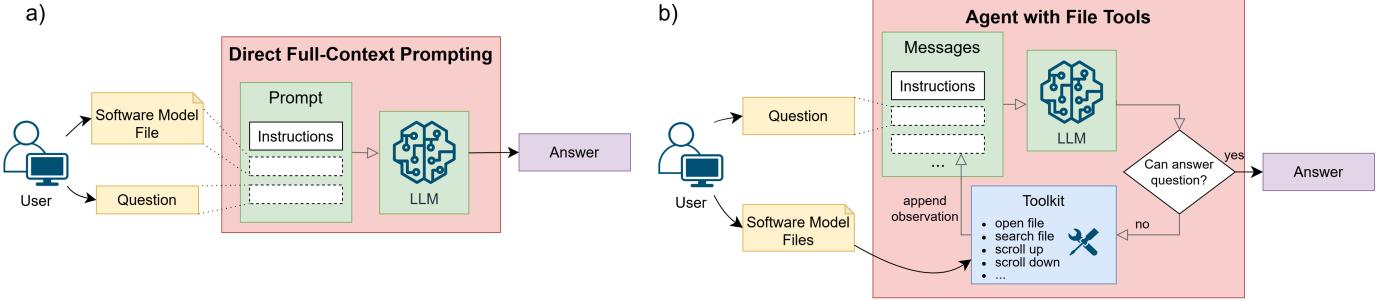


Fig. 1. Evaluation setups for software model question answering. (a) Direct LLM call with full model in prompt. (b) Agent using file-access tools to retrieve relevant model parts before answering.

TABLE I
FILE TOOLING CAPABILITIES IN AGENT-BASED ARCHITECTURE

Tool Name	Description
list_directory	Lists all files in a directory in a tree-style format.
find_file	Finds all files with a given name or pattern.
search_directory	Searches for a term across all files in the directory.
open_file	Opens a file at a given path. Optionally scrolls to a specified line.
scroll_up	Scrolls the open file window up by a fixed number of lines.
scroll_down	Scrolls the open file window down by a fixed number of lines.
go_to_line	Moves the window view to a specific line number in the open file.
search_file	Searches for a term within the open file (or a specified file).

TABLE II
LANGUAGE MODELS USED IN THE EXPERIMENTS

Model Name	Release Date	Input Token Limit	Output Token Limit	Key Characteristics
GPT-4o mini	2024-07-18	128,000	16,384	Compact multimodal model for efficient reasoning.
GPT-4.1 mini	2025-04-14	1,047,576	32,768	High-performance multimodal reasoning and coding.
o4-mini	2025-04-16	200,000	100,000	Lightweight model for chain-of-thought reasoning.
Gemini 2.5 Flash Preview	2025-04-17	1,048,576	65,536	Fast, multilingual reasoning and summarization.

from INCHRON’s am2inc repository [27]. The am2inc tool facilitates the conversion of AMALTHEA models [28] into the INCHRON format, which is then utilized by chronSUITE 3.X [29]. AMALTHEA provides an open and unified model platform for multi-core software and hardware development. Its subsequent evolution into Eclipse APP4MC [28] has significantly broadened its adoption, establishing it as an essential tool for analyzing and optimizing multi-core embedded systems, particularly within automotive applications. The am2inc conversion process is crucial because it filters out extraneous information irrelevant to timing analysis, thereby streamlining the model for chronSIM’s specific focus on timing challenges [30].

The selected metamodel root.ecore is substantial, defining 384 classes and consisting of 13,572 lines. When tokenized using the OpenAI tokenizer for GPT-4o, this translates to approximately 118,000 tokens. This token count highlights that while the model is considerably large and complex, it remains small enough to fit entirely within the context windows of modern LLMs - a crucial requirement for our experiments. Consequently, the selected INCHRON model serves as a representative example of real-world models used in the automotive industry for detailed timing analysis and optimization.

A total of 20 questions were manually created to cover different reasoning and retrieval patterns, grouped into four categories:

- 1) Direct class inspection (5 questions):
Require identifying only the fields (names and types) of a specified class, without considering inheritance.
- 2) Single inheritance chain (5 questions):
Require collecting fields from a class and its base classes along a single inheritance path.
- 3) Multiple inheritance chains (5 questions):
Require traversing and aggregating fields across multiple inheritance paths.
- 4) Semantic queries (5 questions):
Require understanding the model structure and semantics beyond keyword matching.

This dataset enables controlled evaluation of both retrieval-based and reasoning-intensive tasks, simulating realistic software model comprehension needs in automotive software engineering workflows.

E. Evaluation Criteria

This study evaluates the performance of LLMs and agent-based architectures in answering questions about a large software model, using factual correctness as the primary metric.

Question:

What are the fields of Frequency class? Provide names and data types of only this class, don't list fields of base classes.

Answer:

The Frequency class has the following fields defined directly:

1. name: value
 - data type: EDouble (double)
 - lowerBound: 1
 - defaultValueLiteral: 1000
2. name: unit
 - data type: FrequencyUnit (enumeration)
 - lowerBound: 1
 - defaultValueLiteral: MHz

(b)**Tool:** open_file root.ecore**Observation:**

... first 50 lines of the file ...

Tool: search_file Frequency root.ecore**Observation:**

Found 8 matches for "Frequency" in root.ecore:
...

Line 2830: name="Frequency">
...

Tool: goto 2830**Observation:**

File: root.ecore (13572 lines total)
...
2828: <eClassifiers
2829: xsi:type="ecore:EClass"
2830: name="Frequency">
...

Answer:

The Frequency class has the following fields:

1. name: value
 - type: EDouble (double precision floating point)
 - default value: 1000
 - description: The numerical value of the quantity.
2. name: unit
 - type: FrequencyUnit (enumeration)
 - default value: MHz
 - description: The unit of the quantity.

Fig. 2. Example responses to a model-related question using (a) Direct Full-Context Prompting and (b) Agent with File Tools. The agent's directory-related tool call is omitted for clarity.

Factual correctness measures whether the statements in an LLM-generated answer are objectively true with respect to the software model [31]. It ensures that the output neither hallucinates, omits, nor misrepresents key information. In this context, the software model defines the complete scope of truth, including components, interfaces, dependencies, and structural or behavioral constructs described in the Ecore format.

Due to the number and structure of questions, as well as the number of evaluated setups, full human evaluation was deemed impractical. Instead, LLMs were used as proxies for scoring (LLM-as-a-Judge), with manual spot-checking performed for a randomly selected subset of answers to ensure reliability. Two complementary scoring methods were used:

a) Per-answer scoring: Each generated answer was compared against a reference answer using a binary judgment, classified as either correct or incorrect. The accuracy metric was then computed as the ratio of correct answers to total questions. Evaluations were performed using GPT-4.1 mini with a structured prompt taken from the LangSmith Hub [32]. The output format was constrained to structured integer responses, where a score of 1 indicated correctness and 0 indicated incorrectness.

b) Per-fact scoring: To capture more nuanced factual alignment, we employed the FactualCorrectness metric provided by Ragas version 0.2.15 [33]. This metric decomposes the model output and the reference into individual claims and applies natural language inference to identify factual overlap. The same LLM model, GPT-4.1 mini, is used here to perform the inference.

The evaluation returns precision, recall, and F1 scores, defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$F_1 \text{ score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where true positives (TP) refer to claims in the response that also appear in the reference. False positives (FP) are claims present in the response but not in the reference, while false negatives (FN) are claims in the reference that are missing from the response.

This scoring method enables more granular insight into partially correct responses and helps differentiate between omission and hallucination errors.

IV. RESULTS AND ANALYSIS

This section presents the experimental results of direct prompting and agent-based LLM architectures for querying software models. We evaluated these methods using multiple LLM variants on a 20-question dataset, focusing on answer correctness and token efficiency.

A. Correctness Evaluation

Table III presents the factual correctness results of all evaluated architectures and LLMs. Overall, the direct full-context prompting setup performed better or comparably for several models. Notably, Gemini 2.5 Flash and GPT-4o mini achieved significantly higher accuracy in the reference setup compared to the agent-based configuration. Gemini 2.5 Flash reached 80% accuracy in the direct prompting, but only 40% when operating within the ReAct-based tool-augmented agent. Similarly, GPT-4o mini's performance dropped sharply from 45% to 10%. This suggests that these models may struggle with tool-based reasoning, maintaining coherence across tool calls, or operating effectively within constrained, windowed views of the software model.

In contrast, GPT-4.1 mini and o4-mini maintained high accuracy (90%) across both setups. Their performance was especially strong in the agent with file tools architecture, indicating better capabilities for orchestrating tool use and reasoning over partial views of the model. These results support the hypothesis that agent-based setups place greater demands on a model’s reasoning and memory capabilities, which smaller or less capable LLMs may not handle well.

Another key observation is the precision-recall trade-off visible across models. GPT-4.1 mini in the direct prompting, for instance, achieved high recall (0.86) but low precision (0.44), implying a tendency to include more relevant information at the cost of introducing inaccuracies.

Finally, the relatively high standard deviations in precision, recall, and F1 score across models (up to ± 0.36) suggest considerable variability in model performance depending on the specific question type. This highlights the importance of comprehensive evaluation across diverse question categories when assessing LLM effectiveness in model understanding tasks.

B. Token Usage Comparison

To better understand the computational footprint and processing patterns of the evaluated setups, we analyzed token usage across all models and architectures. The results are summarized in Table IV.

A striking, though not surprising, difference is observed in the number of prompt tokens between the direct full-context prompting and the agent with file tools. Direct prompting relies on injecting the entire software model directly into the prompt, resulting in high prompt token counts - typically ranging from 118,000 to over 137,000 tokens on average, depending on the tokenizer used. In contrast, the agent setup loads context incrementally via file tools, resulting in dramatically lower prompt usage—only about 640–780 tokens depending on the model. This demonstrates that agent-based designs are far more efficient in managing input context.

In terms of completion tokens, which represent the size of the final answer, the values are consistently small across all setups. However, they are slightly higher in the reference setup, likely because the model has access to the entire software model upfront and can produce more detailed responses. The agent-based setups produce shorter completions, possibly due to more focused context windows available at response time.

Reasoning token count offers insight into how models internally process queries. GPT-4.1 mini and GPT-4o mini show no reasoning tokens, suggesting they produce answers without explicit intermediate steps. In contrast, Gemini 2.5 Flash and o4-mini exhibit substantial built-in reasoning. Both models use significantly fewer reasoning tokens in the agent configuration than in direct prompting, likely due to the modular, tool-driven context limiting the need for extended internal processing.

Overall, these findings confirm that agent-based approaches are significantly more efficient in token usage, especially for large software models, while still enabling internal reasoning in models designed for that purpose. They also highlight

that reasoning behaviors vary greatly depending on model capabilities.

C. Failure Case Analysis

Two main categories of failures were observed during the evaluation: agent errors, which prevented the model from producing an answer, and incorrect answers, where the model made a factually inaccurate response.

The only error type observed in the agent-based architecture was a recursive error, occurring when the number of tool-use iterations exceeded the predefined recursion limit of 100. These errors prevented the agent from producing any answer, leading to a failed attempt. Across all evaluations, recursive errors occurred 15 times when using the agent with file tools. Two of these cases were attributed to the o4-mini model. In both instances, the agent repeatedly opened the same file and interleaved file searches with opening the file at different lines, continuing this behavior until the iteration limit was reached. Some search terms appeared arbitrary or incorrectly chosen, which may point to a misunderstanding of Ecore syntax. The remaining 13 failures were observed with GPT-4o mini. In twelve of these, the agent opened a file and continuously scrolled down until it hit the iteration cap. In one case, the agent alternated between scrolling down and up while occasionally invoking `find_file` with class names, again without arriving at a coherent answer. All such recursive errors were treated as incorrect responses and were therefore assigned a score of 0.0 in the precision, recall, and F1 score metrics.

The main observation from the incorrect answers is that the dominant failure mode across models - especially in tasks involving class field and inheritance chain extraction - is incomplete or prematurely terminated inheritance traversal. Many agents stop after one or two inheritance levels, failing to follow the full chain and include all inherited fields, which leads to missing attributes. Another frequent issue is confusion or inconsistency in attributing fields to the correct base class, often misplacing a field’s origin in the hierarchy. Some answers are penalized due to minor factual inaccuracies, such as incorrect data types or multiplicity (e.g., treating a field as a list where it’s not explicitly typed as such in the model). In cases where class membership questions are involved (e.g., bus ports, event chain requirements), models often overgeneralize by including additional classes with semantically similar names or nearby definitions, which, while arguably plausible, deviate from the ground truth. Notably, models using file tools often exhibit partial exploration behaviors, such as navigating to a class but not scrolling to see all fields, or identifying a base class but not recursively inspecting further base classes up the inheritance chain. These patterns suggest that incorrect answers often arise not from fundamental misunderstanding but from shallow or incomplete retrieval and reasoning processes.

V. DISCUSSION

Our results reveal several essential insights. First, the reference architecture generally yields higher or comparable factual

TABLE III

FACTUAL CORRECTNESS ACROSS 20 QUESTIONS FOR EVALUATED SETUPs: ACCURACY (%), PRECISION, RECALL, F1 (MEAN \pm STD).

Architecture	LLM Model	Correct	Incorrect	Accuracy (%)	Precision (\pm SD)	Recall (\pm SD)	F1 score (\pm SD)
Direct Full-Context Prompting	Gemini 2.5 Flash	16	4	80.0	0.70 ± 0.29	0.91 ± 0.25	0.75 ± 0.27
	GPT-4.1 mini	16	4	80.0	0.44 ± 0.23	0.86 ± 0.30	0.56 ± 0.23
	GPT-4o mini	9	11	45.0	0.53 ± 0.23	0.79 ± 0.28	0.61 ± 0.22
	o4-mini	18	2	90.0	0.59 ± 0.23	0.90 ± 0.24	0.66 ± 0.22
Agent with File Tools	Gemini 2.5 Flash	8	12	40.0	0.59 ± 0.36	0.68 ± 0.35	0.58 ± 0.36
	GPT-4.1 mini	18	2	90.0	0.49 ± 0.23	0.95 ± 0.11	0.61 ± 0.19
	GPT-4o mini	2	18	10.0	0.19 ± 0.31	0.21 ± 0.35	0.20 ± 0.32
	o4-mini	18	2	90.0	0.48 ± 0.36	0.85 ± 0.37	0.57 ± 0.35

TABLE IV

TOKEN USAGE ACROSS EVALUATED SETUPs. MEAN VALUES WITH STANDARD DEVIATIONS ARE REPORTED.

Architecture	LLM Model	Prompt Tokens (\pm SD)	Completion Tokens (\pm SD)	Reasoning Tokens (\pm SD)	Total Tokens (\pm SD)
Direct Full-Context Prompting	Gemini 2.5 Flash	137461.4 ± 6.1	174.7 ± 138.8	598.0 ± 328.8	138234.1 ± 404.4
	GPT-4.1 mini	118776.0 ± 5.8	210.0 ± 129.4	0.0 ± 0.0	118986.0 ± 131.5
	GPT-4o mini	118776.0 ± 5.8	223.7 ± 164.0	0.0 ± 0.0	118999.7 ± 166.5
	o4-mini	118775.0 ± 5.8	150.3 ± 95.1	563.2 ± 291.9	119488.5 ± 355.2
Agent with File Tools	Gemini 2.5 Flash	783.3 ± 6.1	17.6 ± 6.5	282.5 ± 60.6	1083.4 ± 65.5
	GPT-4.1 mini	639.3 ± 5.7	17.7 ± 1.4	0.0 ± 0.0	656.9 ± 5.8
	GPT-4o mini	642.6 ± 2.1	30.9 ± 4.3	0.0 ± 0.0	673.4 ± 3.6
	o4-mini	641.8 ± 6.0	53.4 ± 9.4	60.4 ± 40.9	755.7 ± 45.4

correctness for most models when the full model fits into the LLM’s context window. This outcome is intuitive since the model has complete visibility of the entire software structure at once, enabling holistic reasoning and direct fact retrieval. Notably, models such as Gemini 2.5 Flash and GPT-4o mini exhibited a significant drop in accuracy when constrained to the agent setup. This suggests that certain LLM variants may struggle with maintaining state coherence across multiple tool calls or with reasoning effectively over fragmented, partial views of the model. Moreover, analysis of incorrect answers shows that the models frequently failed to choose the most effective tool (`search_file`) for the task, instead relying on a less efficient option (`scroll_down`).

In contrast, GPT-4.1 mini and o4-mini maintained robust accuracy (90%) across both architectures, highlighting the importance of advanced reasoning and memory capabilities for effective tool-assisted model comprehension. Their success in the agent setup demonstrates that an LLM capable of efficiently orchestrating the retrieval and synthesis of partial information can overcome the limitations imposed by lacking full context. However, analysis of incorrect answers suggests that accuracy could be further improved by integrating tools specifically designed for browsing software models, which better address the challenges of partial exploration, rather than relying solely on general-purpose file access tools. This insight has important practical implications for industrial workflows, where software models often exceed the size limits of even the largest available context windows.

The observed precision-recall trade-offs provide important insights. As expected, high recall generally corresponds to higher overall accuracy in most cases, indicating that models tend to capture relevant information effectively. However, overall precision remains relatively low across models, suggesting a common tendency to include extra or occasionally

inaccurate details. Striking the right balance between recall and precision is especially critical in engineering domains, where accuracy is paramount and any hallucinations or errors can mislead developers or compromise system safety.

From an efficiency perspective, agent-based architectures dramatically reduce prompt token usage by avoiding full ingestion of large software models. This efficiency gain is crucial for enabling deployment of LLMs with smaller context windows on-premise, addressing automotive industry constraints around privacy, intellectual property protection, and regulatory compliance. Although completion tokens (the size of answers) were similar across setups, the reduction in prompt size directly translates into lower computational costs, faster response times, and the feasibility of more frequent or iterative querying.

VI. THREATS TO VALIDITY

a) Single Model and Format: The study used a single software model in Ecore format, limited to a single file. This constrains generalizability, as other models, formats, or multi-file setups may present different challenges.

b) Limited and Narrow Question Scope: The evaluation was based on 20 questions focused solely on structural, low-level aspects (e.g., class fields). It did not assess behavioral properties or higher-abstraction queries.

c) Answer Validation Limitations: Extra statements generated by the LLM beyond the expected answers were not verified for correctness, which may have impacted precision and recall metrics. Additionally, the evaluation relied on LLM-based scoring methods, potentially introducing inaccuracies in answer correctness determination.

d) Fixed Agent Parameters: The agent used a fixed configuration (e.g., window size, iteration limits), which may have limited adaptability or performance across different tasks.

VII. CONCLUSIONS AND FUTURE WORK

This study compared direct prompting and agent-based approaches for answering questions over software models using LLMs. The agentic approach proved effective with smaller LLMs, demonstrating that compact models possess strong tool usage and reasoning capabilities. This improvement makes sophisticated, retrieval-augmented agent workflows feasible on resource-constrained setups. The results highlight the potential of such structured approaches for interacting with large software artifacts efficiently and accurately.

Future work will explore several directions. First, the approaches will be evaluated in a broader variety of software models and formats to assess generalizability. More complex agent architectures will be investigated to improve reasoning capabilities and task coordination, leveraging emerging standards such as MCP and A2A to enhance tool integration and agent collaboration. The question set will be expanded, including queries at different abstraction levels to probe deeper understanding. Additionally, efforts will be made to automatically generate model instances from metamodels and specifications, and to support textual modifications of software models while preserving consistency.

REFERENCES

- [1] M. Broy, I. H. Kruger, A. Pretschner, and C. Salzmann, “Engineering automotive software,” *Proc. IEEE*, vol. 95, no. 2, pp. 356–373, Feb. 2007. [Online]. Available: <https://ieeexplore.ieee.org/document/4142919>
- [2] A. Haghightkhah, A. Banijamali, O.-P. Pakanen, M. Oivo, and P. Kuuvaja, “Automotive software engineering: A systematic mapping study,” *J. Syst. Softw.*, vol. 128, pp. 25–55, Jun. 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121217300560>
- [3] A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, and J. M. Zhang, “Large language models for software engineering: Survey and open problems,” in *Proc. IEEE/ACM Int. Conf. on Softw. Eng. (ICSE)*, May 2023, pp. 31–53. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10449667>
- [4] “PlantUML.” [Online]. Available: <https://plantuml.com/>
- [5] F. Muff and H.-G. Fill, “Limitations of ChatGPT in conceptual modeling: insights from experiments in metamodeling,” in *Modellierung 2024 - Workshopband*. Gesellschaft für Informatik e.V., Mar. 2024. [Online]. Available: <https://dl.gi.de/handle/20.500.12116/43782>
- [6] J. Di Rocco, D. Di Ruscio, C. Di Sipio, P. T. Nguyen, and R. Rubei, “On the use of large language models in model-driven engineering,” *Softw. Syst. Model.*, vol. 24, no. 3, pp. 923–948, Jun. 2025. [Online]. Available: <https://doi.org/10.1007/s10270-025-01263-8>
- [7] “Unified Modeling Language (UML) Specification Version 2.5.1.” [Online]. Available: <https://www.omg.org/spec/UML/>
- [8] “Eclipse Modeling Framework (EMF).” [Online]. Available: <https://eclipse.dev/emf/>
- [9] “AUTOSAR (Automotive Open System Architecture).” [Online]. Available: <https://www.autosar.org/>
- [10] K. Vinoth Kannan, “Model-based automotive software development,” in *Automotive Embedded Systems: Key Technologies, Innovations, and Applications*, M. Kathiresan and R. Neelavani, Eds. Cham: Springer International Publishing, Apr. 2021, pp. 71–87. [Online]. Available: https://doi.org/10.1007/978-3-030-59897-6_4
- [11] D. Mosquera, M. Ruiz, O. Pastor, and J. Spielberger, “Understanding the landscape of software modelling assistants for MDSE tools: A systematic mapping,” *Inf. Softw. Technol.*, vol. 173, p. 107492, Sep. 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950584924000971>
- [12] Y. Sui, M. Zhou, M. Zhou, S. Han, and D. Zhang, “Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study,” in *Proc. 17th ACM Int. Conf. Web Search Data Mining*, ser. WSDM ’24. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 645–654. [Online]. Available: <https://dl.acm.org/doi/10.1145/3616855.3635752>
- [13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, “Retrieval-augmented generation for large language models: A survey,” Mar. 2024, arXiv:2312.10997. [Online]. Available: <http://arxiv.org/abs/2312.10997>
- [14] T. Sumers, S. Yao, K. Narasimhan, and T. Griffiths, “Cognitive architectures for language agents,” *Trans. Mach. Learn. Res.*, Oct. 2023. [Online]. Available: <https://openreview.net/forum?id=1i6ZCVflQJ>
- [15] B. Paranjape, S. Lundberg, S. Singh, H. Hajishirzi, L. Zettlemoyer, and M. T. Ribeiro, “ART: Automatic multi-step reasoning and tool-use for large language models,” Mar. 2023, arXiv:2303.09014. [Online]. Available: <http://arxiv.org/abs/2303.09014>
- [16] “Model Context Protocol.” [Online]. Available: <https://modelcontextprotocol.io/>
- [17] “Agent2Agent Protocol (A2A).” [Online]. Available: <https://google-a2a.github.io/A2A/>
- [18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, “Chain-of-Thought prompting elicits reasoning in large language models,” in *Advances Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., Nov. 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abc4-Paper-Conference.pdf
- [19] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Sep. 2022. [Online]. Available: <https://openreview.net/forum?id=1PL1NIMMrw>
- [20] E. Kamalloo, N. Dziri, C. Clarke, and D. Rafiei, “Evaluating open-domain question answering in the era of large language models,” in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 5591–5606. [Online]. Available: <https://aclanthology.org/2023.acl-long.307/>
- [21] F. Mustafazade and P. F. Ebbinghaus, “Evaluation of semantic answer similarity metrics,” *Int. J. Natural Lang. Comput.*, vol. 11, no. 3, pp. 43–57, Jun. 2022. [Online]. Available: <http://arxiv.org/abs/2206.12664>
- [22] R. Wang, J. Guo, C. Gao, G. Fan, C. Y. Chong, and X. Xia, “Can LLMs replace human evaluators? An empirical study of LLM-as-a-Judge in software engineering,” Feb. 2025, arXiv:2502.06193. [Online]. Available: <http://arxiv.org/abs/2502.06193>
- [23] S. Badshah and H. Sajjad, “Reference-guided verdict: LLMs-as-Judges in automatic evaluation of free-form text,” Aug. 2024, arXiv:2408.09235. [Online]. Available: <http://arxiv.org/abs/2408.09235>
- [24] “Mermaid.” [Online]. Available: <https://mermaid.js.org>
- [25] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, “ReAct: Synergizing reasoning and acting in language models,” in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Sep. 2023. [Online]. Available: https://openreview.net/forum?id=WE_vluYUL-X
- [26] J. Yang, C. E. Jimenez, A. Wetzig, K. Lieret, S. Yao, K. Narasimhan, and O. Press, “SWE-agent: Agent-computer interfaces enable automated software engineering,” *Advances Neural Inf. Process. Syst.*, vol. 37, pp. 50 528–50 652, Dec. 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a7c947568c1b1328ccc5230172e1e7c-Abstract-Conference.html
- [27] “INCHRON’s am2inc Ecore model.” [Online]. Available: <https://github.com/inchron/am2inc/blob/main/EcoreModels/root.ecore>
- [28] “Eclipse APP4MC.” [Online]. Available: <https://eclipse.dev/app4mc/>
- [29] “chronSUITE from INCHRON.” [Online]. Available: <https://www.inchron.com/chronsuite/>
- [30] “INCHRON’s APP4MC / Amalthea Importer.” [Online]. Available: <https://www.inchron.com/amalthea/>
- [31] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, and P. Nakov, “Factuality of large language models: A survey,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 19 519–19 529. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.1088/>
- [32] “LangSmith Hub Prompt: Evaluation for RAG answer accuracy vs a reference.” [Online]. Available: <https://smith.langchain.com/hub/langchain-ai/rag-answer-vs-reference>
- [33] “Ragas.” [Online]. Available: <https://docs.ragas.io/en/v0.2.15/>