Large Language Model-based suggestion of objective functions for search-based Product Line Architecture design

Willian M. Freire
State University of Maringá
Maringá, Paraná, Brazil
willianmarquesfreire@gmail.com

Murilo Boccardo
State University of Maringá
Maringá, Paraná, Brazil
ra124160@uem.br

Daniel Nouchi
State University of Maringá
Maringá, Paraná, Brazil
ra123991@uem.br

Aline M. M. M. Amaral
State University of Maringá
Maringá, Paraná, Brazil
ammmamaral@uem.br

Silvia R. Vergilio
DInf, Federal University of Paraná
Curitiba, Paraná, Brazil
silvia@inf.ufpr.br

Thiago Ferreira
University of Michigan-Flint
Flint, MI, USA
thiagod@umich.edu

Thelma E. Colanzi
State University of Maringá
Maringá, Paraná, Brazil
thelma@din.uem.br

ABSTRACT

Search-based design of *Product Line Architecture (PLA)* focuses on enhancing the design and functionality of software product lines through variability management, reuse, and optimization. A particular challenge in this area is the selection of objective functions, which significantly influence the success of the search process. Moreover, many objectives make the analysis and choice of a solution to be used harder. The literature has assigned this task to the PLA designer, i.e., the *Decision-Maker (DM)*, who does not always know all the functions and their impact on the optimization outcomes. On the other hand, recent research shows that *Large Language Models (LLMs)*, particularly the *Generative Pre-trained Transformer series (GPT)*, have obtained promising results to help in various Software Engineering (SE) tasks. Considering this fact, this work explores the integration of such LLMs, notably ChatGPT, into the search-based PLA design. By leveraging LLMs' capacity to understand/generate human-like text, we investigate their potential to assist DMs and propose an approach for suggesting objective functions, thereby simplifying and improving decision-making in PLA design optimization. Through empirical tests and qualitative feedback from domain experts, this research highlights the application of LLMs in search-based SE. The results demonstrate that integrating ChatGPT into PLA design can significantly enhance decision-making efficiency and solution quality, with a 40% reduction in time required for selecting objective functions and a 25% improvement in solution quality from the DM's point of view. This study maps out the challenges and opportunities that lie ahead in fully harnessing their potential for PLA search-based design.

KEYWORDS

Product Line Architecture, Objective Function Selection, Large Language Model.

1 INTRODUCTION

Software Product Line (SPL) engineering allows systematic reuse of software assets to develop a family of products [13], enhancing efficiency and reducing time-to-market. Adopting this paradigm brings some complexities in managing variability and designing optimal *Product Line Architectures (PLAs)*. The PLA design is a multifaceted process that balances feature modularization, reuse, variability, and extensibility. While critical for creating robust PLAs, these elements can sometimes introduce competing design considerations [17].

To deal with these complexities, the approach *MOA4PLA (Multi-Objective Approach for Product Line Architecture Design)* [5] applies *Search-Based Software Engineering (SBSE)* techniques, more particularly multi-objective evolutionary algorithms. This approach utilizes a PLA representation based on the class diagram and, for that, some specific evolution operators. The representation is realized through a metamodel that allows the architectural elements to be dynamically manipulated by the search algorithms. Critical architectural components include interfaces, operations, and relationships, marked with *Unified Modeling Language (UML)* stereotypes to denote feature associations. The variability inherent in PLAs is represented through variation points and variants.

MOA4PLA aims to optimize several architectural properties of a PLA design, which are captured by different metrics to establish the objectives to be optimized. Currently, the MOA4PLA evaluation model encompasses 20 objective functions[1], delineating indicators for feature modularization, PLA extensibility, variability, coupling, cohesion, and size dimensions. This approach is supported by the tool OPLA-Tool [11], adopted in our study.

Since these objectives can be conflicting, aggregating them into one objective solution can lead to loose information [14]. Thus, multi and many-objective algorithms are more appropriate[2]. However, some problems arise even when many-objective algorithms

[1] ACLASS, ACOMP, COE, TV, RCC, CM, DC, EC, ELEG, FM, LCC, EXT, SD, SV, TAM, WOCSCLASS, CS, LFCC, FDAC and CIBF. They are described in [23].

[2] In our work, we employed NSGA-II [7]. This choice is due to our focus on assisting DMs in choosing objective functions and because we noticed in the results that few objective functions were selected.

are used. Deb and Jain [6] state that choosing a solution gets harder because most solutions become incomparable. Many solutions are required to generate a Pareto-front[3]. This generation process takes much time and requires special visualization techniques. Thus, reducing the number of objectives used according to the designer's goals is fundamental. This makes selecting objective functions an important task to ensure the success of the search-based process and reduce effort in choosing a solution for practice.

To properly deal with the choice of the properties to be optimized, the literature has assigned this task to the PLA designer [17], i.e., the Decision-Maker (DM). However, DMs may face significant challenges in choosing appropriate objective functions. They need to know each metric and study their impact on the search process. This is magnified by the swift evolution of software requirements and the varied expertise levels among DMs.

On the other hand, recent advancements in Artificial Intelligence, especially in *Large Language Models (LLMs)* like *Generative Pre-trained Transformer (GPT)* [18], have emerged as promising tools to support users' decisions. LLMs have revolutionized the fields of natural language processing by mimicking human-like language understanding and generation capabilities [19, 22]. These advancements have extended the applications of LLMs into software engineering, notably in automating tasks such as code generation, documentation, and SPL engineering [1, 16].

Motivated by these facts, this study aims to explore the integration of ChatGPT[4] into the PLA search-based design, focusing on suggesting objective functions. Incorporating LLMs into the optimization process leverages their natural language processing capabilities to interpret, suggest, and refine optimization strategies.

By integrating LLMs, like ChatGPT, into PLA optimization processes, researchers and developers can enhance decision-making by leveraging these models' ability to interpret complex requirements and suggest objective functions [8, 24]. This integration assists decision-makers in selecting relevant objective functions, streamlining decision-making, reducing computational overhead, and improving the quality of optimized solutions.

In our exploratory study, we evaluate the impact of LLM-generated suggestions on the decision-making process, presenting new possibilities and enhancing optimization outcomes. We also derive insights from qualitative feedback provided by domain experts on the usability and effectiveness of incorporating LLMs into the OPLA-Tool. In this sense, this work not only contributes to the fields of PLA design and AI-assisted optimization but also highlights the potential of LLMs to improve decision-making in software engineering. Our main contributions are (1) an approach to integrating ChatGPT into the PLA optimization process; (2) a module that improves the usage of OPLA-Tool even for non-specialist users, based on their preferences; (3) identification of areas for improvement in using LLMs for PLA optimization; and (4) proposal of enhancements to the OPLA-Tool.

The paper is organized as follows: Section 2 presents the LLLM-based Objective Function Suggestion proposal. Section 3 describes the exploratory study conducted. Section 4 presents the main results. Section 5 reviews related work. Finally, Section 6 concludes the paper and outlines directions for future research.

[3]A Pareto-front is a set of solutions considered optimal. Solutions in the Pareto-front are those where no objective can be improved without worsening at least one other objective.

[4]ChatGPT is an advanced language model developed by OpenAI, based on the GPT architecture.

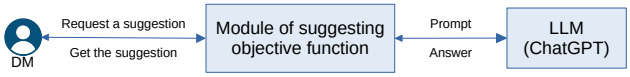

Figure 1: Module of suggesting objective functions.

2 LLLM-BASED OBJECTIVE FUNCTION SUGGESTION

This section introduces our approach to enable the suggestion of objective functions through the collaborative efforts of a DM and an LLM (ChatGPT). The main idea is to implement this approach in OPLA-Tool, which supports the search-based PLA design and allows using 20 objective functions. With many objectives. As originally available in the OPLA-Tool, the DM must understand every objective function before using the tool. With the proposed module, this understanding is downsized, as the module assists the DM in choosing the best objective functions for their use.

The proposed module aims to streamline the process of identifying suitable objective functions for the search-based PLA design, enhancing decision-making efficiency and accuracy.

System Workflow: A dynamic interaction model between the DM and the LLM is at the core of our proposal, as depicted in Figure 1. The process initiates with the DM requesting suggestions for objective functions. This request is translated into a structured prompt and dispatched to ChatGPT, which processes the query and returns a list of suggested objective functions. All the questions are accompanied by an instruction to receive the answer in a specified JSON format: `{"fns": ["..."], "suggestion": "..."}`. Here, `fns` denotes the objective functions, represented as uppercase acronyms, and `suggestion` provides detailed recommendations. This format was chosen due to its simplicity and ease of integration into existing systems, ensuring that suggestions are accessible and actionable via API (Application Programming interface).

The module's design prioritizes seamless integration with existing systems, particularly the OPLA-Tool. Adopting a standardized JSON format for responses ensures compatibility with the tool's API, facilitating a smooth workflow where suggestions can be directly applied to the optimization process. This integration is essential for the module's utility in practical scenarios.

An essential factor in the module's effectiveness is its comprehensive training strategy. The LLM is equipped with a deep understanding of the domain by leveraging a carefully curated dataset derived from authoritative sources on MOA4PLA objective functions. This preparation is crucial for enabling ChatGPT to provide recommendations that are not only accurate but also aligned with the latest developments and best practices in PLA design and optimization.

Visual Illustration. Figure 2 illustrates the sequential interactions between the DM and the LLM. This figure also showcases the use of this module within OPLA-Tool, showing a DM's journey

from selecting the "Get a suggestion" button to receiving objective function suggestions based on their preferences.

The module has an intuitive interface that facilitates the interaction between the DM and ChatGPT. By simply inputting their requirements or preferences through the "Get a suggestion" button, DMs can initiate a query for objective function suggestions. This simplicity ensures the module's accessibility to users of varying technical backgrounds.

Training Data Preparation. The LLM's proficiency in suggesting objective functions is rooted in a comprehensive training regimen, leveraging data from some important pieces of work. These sources were meticulously selected for their exhaustive coverage of MOA4PLA objective functions, providing a robust foundation for the model's recommendations. The training material, compiled into a 31-page document, encompasses various relevant topics, from PLA design principles to objective function metrics and evaluations. All the used material is available as supplementary material in our repository [10].

3 EXPLORATORY STUDY DESIGN

This section describes the study design, which **aims to** analyze the use of an LLM to suggest objective functions for search-based PLA design, **with the purpose of** assisting the DM to choose the objective functions according to their preferences and the PLA to be optimized, **from the point of view of** PLA architects, **in the context of** researchers and developers.

According to our goals, we defined the following *Research questions (RQs)*:

- **RQ1 - How accurately can ChatGPT suggest relevant objective functions for PLA design based on the decision-makers' preferences?** This RQ aims to evaluate the accuracy and relevance of ChatGPT's suggestions during the implementation validation phase. The metric used for answering RQ1 is **Precision** - The ratio of the number of relevant objective functions suggested by ChatGPT to the total number of suggestions made. This metric also includes the percentage of suggested objective functions that align with the DM's stated preferences and requirements, as assessed by the DMs themselves.
- **RQ2 - What is the impact of ChatGPT's suggestions on the efficiency and user satisfaction of the decision-making process in PLA design?** This RQ aims to assess the impact of ChatGPT's suggestions during the qualitative experiment phase, focusing on decision-making efficiency and user satisfaction. The metrics used for answering RQ2 were: **Time Efficiency** - The amount of time saved by DMs when selecting objective functions with ChatGPT's assistance compared to manual selection without the tool; **User Satisfaction** - Evaluation from DMs on the overall experience of using the tool, including ease of use, clarity of suggestions, and overall satisfaction (on a scale of 1 to 5, with 1 being very dissatisfied and 5 being very satisfied), collected through post-use surveys.

To answer each RQ, our study strategy unfolds in two phases: **Implementation Validation** for RQ1 and **Qualitative Experiment** for RQ2. Our repository [10] contains all material produced conducting both phases, including the following artifacts: objects of analysis as well as the graphs, LLM prompts, training dataset, the full questionnaire used in the qualitative analysis, detailed profiles of DMs, textual corpus with all the DMs answers, and measuring instruments. Next, we present more details on how these phases were conducted.

Figure 3 presents the design protocol steps of this work. The next subsections describe each step.

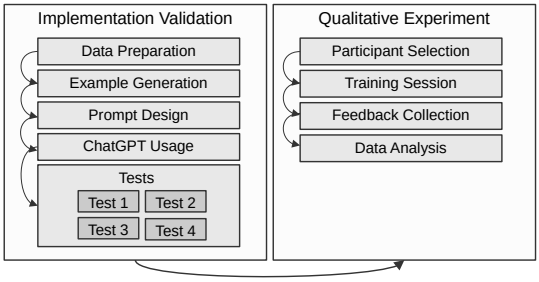

Figure 3: Design protocol steps.

3.1 Implementation Validation

The primary goal of this phase is to answer RQ1. For that, firstly, we developed a specialized training dataset that encapsulates a comprehensive overview of objective functions of the MOA4PLA, and then, we evaluated ChatGPT's proficiency in generating objective function suggestions that resonate with DMs' expressed preferences. This evaluation tests the model's accuracy in suggesting objective functions for optimizing PLA design.

The foundation of our study, particularly for implementation validation, involves the selection of literature to create a robust training dataset for ChatGPT. We sourced data from three fundamental works: [20] [25] [15], chosen for their comprehensive coverage of MOA4PLA objective functions. This literature compilation is the basis for training the LLM, ensuring it thoroughly understands the domain-specific nuances in PLA optimization.

To evaluate the accuracy of ChatGPT for suggesting objective functions in PLA design, we conducted a series of preparatory tests. Each test was designed with specific objectives and structured to provide insights into ChatGPT's performance in suggesting objective functions. Below, we describe each test in detail, including its structure, objectives, and expected outcomes.

Test 1 was performed to evaluate ChatGPT's ability to identify and explain correlations between different objective functions used in PLA design. A curated dataset containing descriptions and known correlations of the MOA4PLA 20 objective functions from authoritative sources was used as input data. The steps followed are: (1) Provide ChatGPT with a set of objective functions and ask it to identify and explain correlations; (2) Use a predefined prompt format to guide the model's responses; and (3) Compare ChatGPT's output with the known correlations from the dataset. As a result, ChatGPT should correctly identify significant correlations between

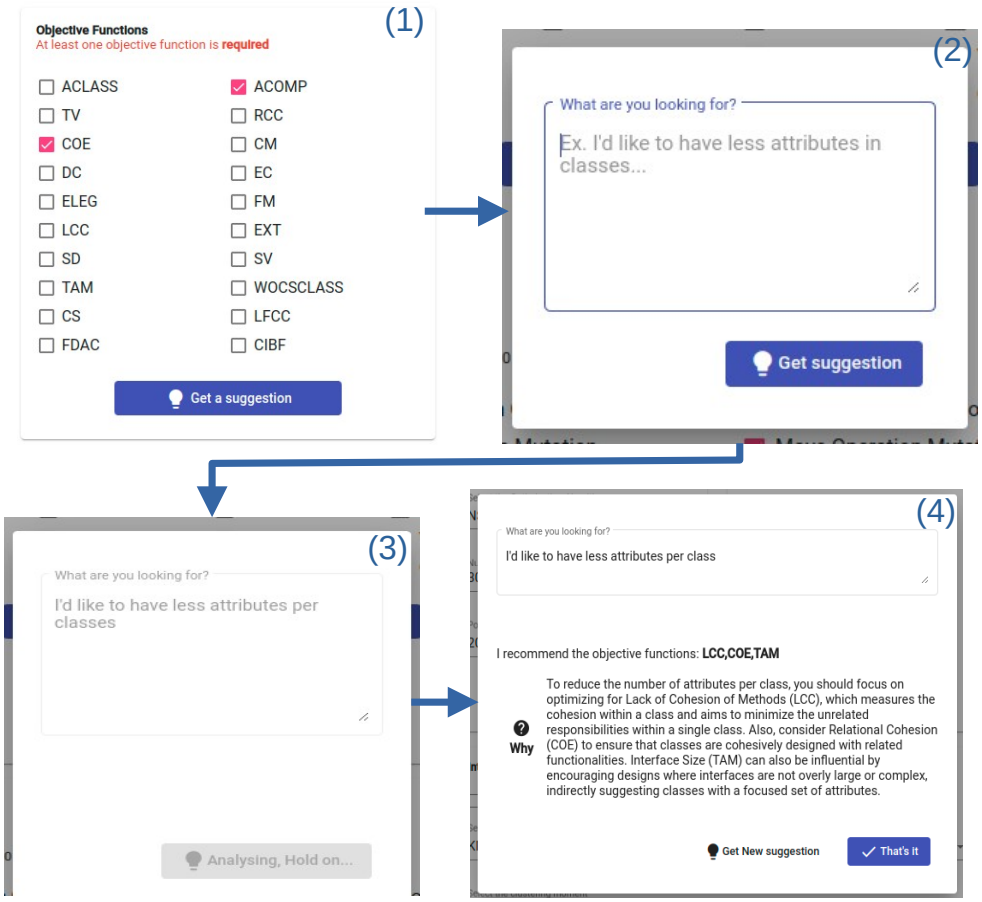

Figure 2: Running example of objective function suggestion.

pairs of objective functions and provide explanations that match the dataset's known correlations. The results demonstrate the model's understanding of the relationships between objective functions.

Test 2 was carried out to assess ChatGPT's capability to expand its domain knowledge by incorporating additional literature and providing accurate descriptions of objective functions. A dataset derived from multiple studies on MOA4PLA objective functions was used as input data. The adopted procedure includes: (1) Train ChatGPT using the expanded dataset; (2) Ask ChatGPT again to describe the 20 objective functions and their correlations with this new data; and (3) Evaluate the accuracy and completeness of the descriptions provided by ChatGPT. As an outcome, ChatGPT should demonstrate an enhanced understanding of the objective functions and provide more detailed and accurate descriptions. The results show the model's ability to assimilate and apply new information from expanded datasets.

Test 3 was performed to evaluate ChatGPT's ability to generate a comprehensive correlation table covering all 20 objective functions using the full training dataset. A combined dataset was used as input data, including all relevant information from previous tests. The procedure includes: (1) Providing ChatGPT with prompts to generate a correlation table for all objective functions; and (2) Assessing the completeness and accuracy of the generated correlation table. As an outcome, ChatGPT should produce a detailed correlation table that accurately reflects the known relationships between the objective functions. The results indicate the model's capacity for handling comprehensive data synthesis tasks.

Test 4 was performed to compare the performance and cost-effectiveness of ChatGPT-3.5 and ChatGPT-4 in generating relevant objective function suggestions. A standardized dataset created from previous tests was used as input data for both versions of ChatGPT. The procedure was: (1) Training both ChatGPT-3.5 and ChatGPT-4 using the same dataset; (2) Evaluating their suggestions' consistency, accuracy, and format of answers; and (3) Comparing the costs associated with each version. As a result, ChatGPT-4 should provide more consistent and accurate suggestions, albeit at a higher cost. Conversely, ChatGPT-3.5 may offer a cost-effective alternative with some trade-offs in performance.

These preparatory tests were designed to systematically evaluate ChatGPT's capabilities in different aspects of PLA design optimization. The results from these tests reasoned the subsequent qualitative experiments and overall assessment of ChatGPT's integration into the OPLA-Tool.

3.1.1 Training ChatGPT details. We employed a few-shot learning approach to suggest objective functions, using prompt engineering without fine-tuning the model. This method involved providing carefully selected examples and prompts, leveraging ChatGPT's existing knowledge for PLA design optimization. We interacted with ChatGPT-4 via the OpenAI API, utilizing its advanced capabilities through a paid subscription on a cloud-based infrastructure, ensuring efficient performance and access to all features.

(1) **Data Preparation:** As mentioned before, to make ChatGPT suggesting objective functions, we prepared a training dataset. This dataset was compiled into a 31-page document containing detailed information on each of the 20 objective functions used in the MOA4PLA. The dataset includes a description of objective functions, their correlation, examples, use cases, best practices, and guidelines.
(2) **Example Generation:** We generated a set of example queries and corresponding desired responses to illustrate how ChatGPT should interpret and suggest objective functions based on the given context. These examples were crafted to cover various possible scenarios DMs might encounter.
(3) **Prompt Design:** The prompts were designed to include specific instructions for ChatGPT, ensuring that the model's output was in the required JSON format: `{"fns": ["..."], "suggestion": "..."}` (Section 2). This format facilitates the integration with OPLA-Tool.
(4) **ChatGPT Usage:** When using LLM, we provided the examples as part of the ChatGPT prompt, enabling it to generate contextually relevant suggestions. The prompts included parameters to ensure that the responses adhered to the specified format and provided clear, actionable recommendations.

The choice of few-shot learning and the use of ChatGPT allowed us to leverage LLM's capabilities without requiring extensive computational resources or time-consuming retraining processes. This approach ensured that the model could adapt to the specific needs of PLA design optimization and provide high-quality suggestions to assist decision-makers.

Using a public version of ChatGPT involves significant privacy considerations, particularly concerning the handling of sensitive data. To address these concerns, we implemented some of the following measures. All data used in interactions with ChatGPT were anonymized. Any sensitive information that could potentially identify individuals or specific projects was removed or obfuscated before being processed by the model.

Also, only the minimum necessary information was shared with ChatGPT to generate relevant suggestions for selecting objective functions. This approach limits data exposure and reduces the risk of leaking sensitive information. The interactions with the ChatGPT were conducted over secure, encrypted channels to ensure that data transmission was protected against interception and unauthorized access. We also adhered to data protection regulations, including obtaining necessary consent and ensuring that data subjects' rights were respected.

3.2 Qualitative Experiment

Following the assessment, the study advances to a qualitative examination by engaging directly with a DM. This phase is fundamental to validate the practical applicability of ChatGPT's suggestions in real-world PLA design scenarios and to evaluate the nuanced understanding of ChatGPT in aligning its recommendations with the complex and varied DM preferences. This bifurcated approach—combining data-driven analysis with experiential insights—enables a holistic evaluation of the LLM's potential as a supportive tool in the PLA design. Through this exploration, we aim to substantiate the efficacy of employing ChatGPT in navigating the landscape of PLA optimization, thereby offering a novel way to enhance the decision-making capabilities of PLA architects.

We selected two distinct PLA designs to optimize and analyze in the qualitative experiment. The first is the academic SPL Arcade Game Maker (AGM), which encompasses three arcade games [21]. The second PLA, Electronic Tickets in Urban Transportation (BET), represents a real-world SPL developed for urban transport management [9]. Participants of this experiment were provided with two essential documents to support their involvement in the experiment [10]: (1) A questionnaire designed to capture the user profile, covering aspects such as educational level, software development activity (academic or industrial), UML knowledge, and PLA design experience, and (2) An evaluation form for assessing the PLA solutions, featuring fields for Solution identification, Score (on a scale of 1-5), and justification for their evaluation.

In evaluating the proposed objective function suggestion module, DMs were asked to provide feedback on several aspects of the tool's performance and usability. The questions focused on potential improvements for the tool or interactive module, the extent to which DMs agree that the module aids in choosing objective functions, their opinions on the response time of the algorithm for suggesting objective functions, their thoughts about the interaction mode with the module, the appropriateness of the suggested objective functions to their profile, the suitability of generated solutions to their profile including specific solution identifiers, and whether there were any specific criteria they used in their subjective analysis of the solutions. These inquiries aimed to gauge the module's effectiveness in assisting DMs with decision-making, the user-friendliness and responsiveness of the interface, and the relevance of both the suggested objective functions and the resultant optimization solutions to the specific needs and profiles of the DMs, alongside understanding the criteria DMs consider essential in evaluating such solutions.

Two individuals (DMs) were chosen to participate in the qualitative experiment based on their expertise in PLA design and UML knowledge. The first participant is a doctoral-level educator with advanced PLA design and UML knowledge. The second participant is a computer science graduate working in the industry with moderate expertise in the relevant fields. These two DMs were chosen to certify if the proposed module can capture their preferences since they understand the problem domain. Each participant underwent

a comprehensive training session, which included detailed explanations of the architectural aspects under examination and hands-on practice with the proposed module. The training was executed via Google Meet, allowing for real-time assistance from the authors.

After training, DMs use the proposed model executing the OPLA-Tool. The DMs decided to optimize only the AGM. According to DMs, this PLA is easier to analyze since it has fewer components, classes, and interfaces than BET. Also, the experiment was carried out at Google Meet, and the first author of this work was present on the call to assist the DMs in a neutral and non-influential manner, ensuring the experiment's integrity.

We employed a rigorous two-step coding process to analyze the qualitative feedback using the participant's responses systematically. This process is significant as it ensures that the analysis is thorough and comprehensive. Initially, open coding was conducted to categorize the feedback into discrete concepts. This involved reading through the responses to identify and label key ideas, issues, and suggestions mentioned by the DMs. Subsequently, axial coding was used to relate these concepts to broader themes that emerged across the feedback. Axial coding is a qualitative research technique used to identify relationships between categories and subcategories that emerge from the initial coding (open coding) process. It involves reassembling the data in new ways by connecting the identified codes to form a coherent narrative.

Themes represent patterns and insights across the coded data, allowing researchers to construct meaningful narratives or theories from their analysis. A theme is more than a collection of similar codes; it represents a significant pattern in the data related to the research questions and has an underlying idea or concept that holds the data together. The coding method allowed us to synthesize the data into significant patterns and insights relevant to the optimization tool's design and functionality.

Finally, to evaluate the impact of ChatGPT on the time efficiency of selecting objective functions, we collected data on time spent by decision-makers (DMs) during the qualitative experiment. DMs were conducted to select objective functions twice: (Round 1) once without and (Round 2) once with the assistance of ChatGPT. We recorded the time taken for each session using a stopwatch. The DMs were asked to choose objective functions to optimize feature modularization, class coupling, and cohesion in the PLA. The best observed in literature are ACLASS, COE, and FM since this objective function measures these aspects [15] [20]. In Round 1, when selecting objective functions without ChatGPT, DMs need to read the description of each objective function to choose it correctly for the problem. We collected the time spent and asked them to choose objective functions again, but now assisted by ChatGPT. The average time spent in Round 1 was compared to the time spent in Round 2 to quantify if there was a reduction in time when using ChatGPT. For that, we calculated the proportion of the time spent in Round 1 divided by time spent in Round 2. Considering that the time spent in Round 1 was smaller than Round 2, the formula used to calculate the proportion was `"1 - (Round 1 time - Round 2 time)"`. The result of this formula is a value between 0 and 1, indicating the decreasing of time spent in selecting objective functions.

The qualitative experiment aimed to understand the practical utility of ChatGPT's suggestions in real-world PLA design scenarios, a key aspect that underscores the relevance of the research. It also aimed to identify areas for improvement in the ChatGPT integration and the OPLA-Tool interface, and gather insights into the decision-making processes of DMs and how they are influenced by the suggestions provided by ChatGPT.

4 RESULTS AND DISCUSSION

This section delineates the findings from both phases of our exploratory study to answer our RQs. Detailed analyses and additional information are accessible as a complementary material [10].

4.1 Implementation Validation

This phase comprises four distinct tests designed to evaluate ChatGPT's capabilities in various scenarios pertinent to search-based PLA design. The few-shot learning approach facilitated through the OpenAI API, enabled us to train ChatGPT effectively for this task.

Test 1 was performed to check the correlation analysis of objective functions. ChatGPT-4 accurately identified significant correlations, both positive and negative, between pairs of objective functions, echoing the findings reported in the literature. For example, it correctly identified that increasing modularization (measured by objective function FM) often leads to decreased class coupling (measured by objective function ACLASS), a known positive correlation. Similarly, it was identified that increasing cohesion (measured by objective function COE) might conflict with class coupling, which is a known negative correlation. Notably, the LLM provided detailed explanations for the correlation between objective functions such as ACLASS (Class Coupling) and ACOMP (Component Coupling), underscoring its potential as a tool for guiding SPL architects in their optimization efforts.

During the questions we made to the LLM, the model was requested to give some suggestions of prompts considering what we wanted, and these prompts were used to formulate the questions to ChatGPT-4. The LLM's ability to suggest precise prompts for further inquiries was instrumental in refining our questions and ensuring clarity and relevance in the responses obtained. This capability demonstrates the LLM's value in facilitating more effective communication with DMs, particularly in complex domains such as PLA optimization.

In an illustrative query, we presented the LLM with a scenario involving the AGM PLA class "Velocity" and its attributes, seeking recommendations for objective functions to enhance the search-based design. ChatGPT-4's suggestions were relevant and accompanied by a rationale grounded in PLA design principles, showing the model's ability to provide actionable insights.

A critical aspect of our investigation was evaluating the LLM's compatibility with the OPLA-Tool's requirements, specifically generating responses in a predefined JSON format for API communication. ChatGPT-4 demonstrated a 100% success rate in adhering to this format, indicating its potential for seamless integration into automated PLA optimization workflows.

An outcome of Test 1 was the LLM's generation of a correlation table directly in LaTeX format (see example presented in Table 1). This is a significant achievement as it showcases the LLM's utility in automating aspects of research documentation. The correlations'

accuracy, as verified against the original study, highlights ChatGPT-4's precision in data interpretation and presentation.

The negative correlation between TAM (Total Attributes per Module) and ACOMP (Component Coupling) is significant because it accurately reflects the well-established understanding in PLA design that increasing the number of attributes per module often leads to reduced component coupling. This correlation aligns with expert knowledge, thereby validating ChatGPT's ability to capture essential design principles in its suggestions.

Table 1: Correlation Between Objective Functions as Generated by ChatGPT-4.

Function	ACOMP	ACLASS	COE	TAM
ACOMP	-	Strong	Moderate	Negative
ACLASS	Strong	-	-	-
COE	Moderate	-	-	-
TAM	Negative	-	-	-

Test 2 was performed to expand the Domain Knowledge with Additional Literature. ChatGPT-4 demonstrated an advanced capability to distill complex information from the literature, effectively describing 17 of the 20 objective functions. This indicated the model's ability to integrate and apply new information to provide relevant insights.

However, generating a comprehensive correlation table with all 20 objective functions proved challenging. This difficulty arose due to the model's token constraints, which limit the amount of information ChatGPT can process and output in a single response. To put it simply, the model can only handle a certain amount of data at a time, and when it's asked to process more than it can handle, it struggles to provide a comprehensive output. As a result, while the model successfully described individual objective functions and their correlations, creating a comprehensive table covering all correlations exceeded its current capabilities. This limitation highlights a significant challenge in using LLMs for tasks that require extensive data synthesis and detailed outputs, suggesting the need for strategies to manage and segment information more effectively when dealing with complex, data-intensive tasks.

Attempts at detailed correlation analysis encountered limitations in data synthesis and presentation, with the LLM struggling to produce a full correlation table accurately with all 20 objective functions. Expanded training data in **Test 3** highlighted challenges in excess of data and inconsistent correlations, necessitating strategic adjustments for more targeted analysis. **Test 3** expanded the scope of ChatGPT-4's training to encompass the entirety of content from three main studies. This approach evaluated the model's capacity for synthesizing a broad array of information into a coherent correlation table covering all identified objective functions. From this test, we highlight two critical insights: the accuracy of correlation predictions by the LLM is highly dependent on the specificity of the input context, particularly the PLA version under consideration; the challenges encountered underscore the necessity for strategic data management and question formulation to navigate the limitations of LLM short-term memory. The variability observed in correlation strength across different PLA versions further emphasizes the nuanced nature of objective function interdependence and the essential role of precise and context-dependent inquiries to extract insights from LLMs for PLA optimization.

Due to the operational costs associated with ChatGPT-4 (~5 times ChatGPT-3.5), **Test 4** aimed to assess the viability of utilizing the more economical ChatGPT-3.5 version for suggesting objective functions in PLA optimization. This evaluation focused on comparing the cost-benefit ratio and consistency of responses between the two model versions. **Test 4** revealed the cost advantages of utilizing ChatGPT-3.5. However, the decision to proceed with ChatGPT-4 was driven by its superior consistency in the answers in the JSON format. While ChatGPT-3.5 offers a more economical option, its challenges in maintaining response consistency and adhering to specified formatting requirements hinder the efficiency and effectiveness of the objective function suggestion module in PLA optimization contexts. The choice to utilize ChatGPT-4 acknowledges the trade-off between operational costs and the quality of output, underlining the importance of reliability and precision in the decision-making support provided to DMs. This decision impacts the work by emphasizing the necessity to balance financial considerations against the need for a robust tool that can accurately align with DMs' complex and varied preferences in real-world PLA design scenarios. The limitations observed when using ChatGPT-3.5 underscore the need for future research to explore more cost-effective solutions without compromising the quality of service, potentially through model optimization or leveraging advancements in LLM technologies.

The relevance of ChatGPT's suggestions, achieved through few-shot learning, was evaluated by asking DMs to rate the alignment of each suggested objective function with their preferences. Out of 200 suggestions made by ChatGPT, 170 were deemed relevant, resulting in an alignment rate of 85%. This high alignment rate indicates that most of ChatGPT's suggestions aligned with what the DMs considered important and relevant for their PLA solutions. The precision metric, calculated as the ratio of relevant suggestions to the total number of suggestions, was 0.85. This means that 85% of the time, ChatGPT's suggestions were judged as useful and appropriate by the DMs. This qualitative feedback highlights the tool's ability to understand the DMs' requirements and provide recommendations that reflect their preferences and needs, thus demonstrating the model supports decision-making in PLA design.

The relevance of ChatGPT's suggestions, achieved through few-shot learning, was evaluated by asking DMs to rate the alignment of each suggested objective function with their preferences. Out of 200 suggestions made by ChatGPT, 170 were deemed relevant, resulting in an alignment rate of 85%. The precision metric indicated that ChatGPT provided relevant suggestions with a precision of 0.85. This qualitative feedback highlights the tool's ability to understand and reflect the DM requirements.

Answer to RQ1: To address this question, our investigation demonstrated that ChatGPT could indeed recommend objective functions that align well with DM preferences. Through targeted tests, ChatGPT showcased its proficiency in grasping the nuances of PLA design requirements and generating suggestions that not only met but occasionally exceeded the expectations of DMs. Its ability to parse complex design scenarios and propose relevant, preference-aligned objective functions underlines the potential of LLMs to

significantly contribute to more personalized and effective PLA optimization strategies.

4.2 Qualitative Experiment

This subsection presents a qualitative analysis of the feedback provided by the experiment participants. As presented in Table 2, during the coding process, several major themes were identified through open and axial coding, reflecting the core areas of DM feedback and concerns. Tables presenting the themes' relationship are in our repository [10]. Some speeches from DMs are presented below.

Regarding **Theme i**, the evaluation of PLA by two DMs with distinct profiles (academic and industry) demonstrates the multifaceted nature of PLA evaluation. As mentioned by DM 1, *"can assist the DM with no experience in choosing objective functions"*. DM 1, rooted in academia, emphasizes the importance of PLA design, feature descriptions, and thorough documentation, advocating for a balanced approach that considers technical and explanatory aspects. On the other hand, DM 2, from the industry sector, evaluated the module as *"Fast, normal, as expected"*. This DM focused on practical elements such as reducing abstraction layers, enhancing modularity, and optimizing objective functions, highlighting the need for efficiency and adaptability in software design.

Analyzing **Theme ii**, we observe that several recommendations were formulated by the DMs, such as: (1) to implement tooltips or question marks next to critical features and configuration items to assist DMs; (2) to develop more detailed and interactive training modules that include practical examples and simulations to illustrate the tool's capabilities; (3) to redesign the DM interface to be more intuitive, with clear labels and fewer abstraction layers; and (4) to allow greater customization of the optimization process, enabling DMs to tailor the tool to their specific needs and preferences.

These recommendations can be observed when DM 2 highlights, *"It would be interesting to improve interaction when the user wants to refine a previous prompt, perhaps presenting a short history of the message exchange (e.g., last 3)"*. This DM stresses the importance of DM-centric tool improvements for better usability. Furthermore, DM 1 emphasizes that *"It is important to place a question button in all fields (including chat) to inform the meaning of each resource"*.

Considering **Theme iii**, the feedback from DMs 1 and 2 provides insightful reflections on the proposed module performance and its impact on their decision-making. DM 1's experience highlights the module's capability to deliver insights that potentially exceed those from individuals without prior experience in analyzing optimized solutions, emphasizing the module's adeptness in suggesting relevant objective functions and discerning the nuanced differences between the correlation of objective functions and features. DM 1 stated *"I was surprised. The module brought answers that, from my point of view, surpass the suggestion of a trained user with no experience in analyzing optimized solutions"* and *"The chat managed to understand the difference between the correlation of objective functions and the correlation of features"*.

Furthermore, DM 2's feedback reinforces the module's value in mirroring the analytical depth associated with experienced users of the OPLA-Tool. The alignment of the module's suggestions with solutions is characteristic of an experienced evaluator's approach, which validates the module's efficacy in capturing and applying complex evaluative criteria similar to those employed by seasoned practitioners in PLA optimization. DM 2 said that the suggestions *"corresponded to the solution of a user with experience in evaluating solutions of OPLA-Tool."*

Regarding **Theme iv**, the feedback on adaptability and customization from DMs underscores a crucial need for the proposed module to be highly flexible and user-centric, catering to their preferences and needs. DMs highlighted the importance of incorporating customizable objective functions and interactive aids for enhanced navigation, reflecting a strong desire for a tool that is not only technically proficient but also intuitively aligned with the diverse styles of its users. A notable suggestion from DM 2 about implementing *"Wizard context cleaning"* to allow fresh interactions with the LLM without the influence of past dialogues further emphasizes the demand for a system that can dynamically adjust to new contexts and user needs. DM 1 also suggested that the module could *"resolve any possible confusion"*. Such improvements are essential for elevating the tool's usability, making it a more effective and personalized aid in the complex process of PLA optimization.

As mentioned in Section 3, we trained the DMs on using the OPLA-Tool and asked them to assign a score between 1 and 5 to this training. Feedback regarding the need for training was highly positive, with DM 1 rating its importance with the maximum score (5). Their suggestions for improving the tool usage stand in a clearer explanation of its functionalities, specifically the purpose and impact of each button, and the introduction of interactive elements such as question marks for instant guidance. DM 2 emphasized the importance of thorough training for the OPLA-Tool, rating this need with the maximum score (5). The DM's feedback demonstrates the need for better clarity in describing objective functions and a deeper explanation of their effects on optimization. It highlights a need for DMs to fully grasp the implications of their choices when using the tool.

The qualitative analysis revealed strengths and areas for improvement in the OPLA-Tool. To support further research and replication, we have made all relevant materials, including datasets and prompts, available in [10]. This transparency allows the research community to build on our work and apply methodologies similar to those of other software engineering contexts.

The integration of ChatGPT significantly reduced the time required for DMs to select objective functions. This reduction was calculated by comparing the time spent by DMs on this task with and without the assistance of ChatGPT. Table 3 summarizes the average time spent selecting objective functions without the assistance of ChatGPT (Round 1) and with it (Round 2) for both DMs. The second and third columns present the time spent in minutes for rounds 1 and 2, respectively; the last column presents the proportion of reduction and the percentage of the time spent for selecting objective functions. On average, DMs spent 30 minutes manually selecting objective functions, whereas ChatGPT's assistance reduced the time to 18 minutes, resulting in an average reduction of 40%. User feedback was positive, with DMs rating their overall experience as 4 out of 5, citing ease of use and the clarity of suggestions as major benefits.

Table 2: Major themes.

ID	Theme	Meaning
i	DM Experience and Evaluation Criteria	Feedback highlighted the diverse perspectives of DMs from academic and industry, emphasizing the need for a tool that balances technical depth with practical applicability
ii	Tool Usability and Interaction Design	The necessity for enhancing the OPLA-Tool DM's interface to make it more DM-friendly, with specific suggestions for interactive guidance and clearer functionalities
iii	Training Effectiveness	The importance of comprehensive training materials was underscored, including the need for practical examples and clearer guidance on the tool's use
iv	Adaptability and Customization	DMs expressed a desire for the tool to be more adaptable to individual needs, suggesting features like customizable objective functions and interactive aids for easier navigation

Table 3: Time Reduction in Selecting Objective Functions with ChatGPT

DM	Round 1	Round 2	Reduction in time
DM 1	28	16	0.428 (42.8%)
DM 2	32	20	0.375 (37.5%)
Average	30	18	0.401 (40.1%)

Answer to RQ2: Concerning RQ2, the study's outcomes highlight ChatGPT's substantial impact on streamlining and enriching the decision-making workflow. Feedback from domain experts confirmed that the integration of ChatGPT into the PLA design process simplified the selection of objective functions and enhanced the overall quality of decision-making. By providing insightful, data-driven recommendations, ChatGPT allowed DMs to optimize PLAs with greater confidence, marking a clear advancement in using LLMs to bolster decision-making capabilities in selecting objective functions.

4.3 Impact on the Optimization Process

The integration of ChatGPT into the PLA design process significantly influenced the search/optimization process. Decision-makers, particularly those with limited technical expertise, often struggle with understanding and selecting appropriate objective functions. ChatGPT's targeted suggestions reduce this cognitive load, making the decision-making process clearer and more straightforward. In our qualitative feedback, experts noted increased confidence when using the tool, as it provided clear rationales for each suggested objective function.

4.4 Threats to Validity

This study faces some validity threats that could impact the interpretation and generalization of its findings. To ensure *internal validity*, a diverse range of objective functions was selected to represent various design challenges, though the use of a single LLM model may limit the breadth of our insights. *External validity* is addressed by acknowledging that the specific LLM and PLAs used may not encompass all possible scenarios encountered in software engineering, suggesting the need for further research with additional models and PLA. *Construct validity* was reinforced through a qualitative experiment, but the small number of participants and limited scenarios restrict the generalizability of the results. Additionally, relying on DM evaluations to determine the relevance of ChatGPT's suggestions introduces the potential for bias. To mitigate this, we ensured that DMs were experienced in PLA design, but future studies should include more objective metrics or involve multiple DMs to cross-validate the relevance of the suggestions. *Conclusion validity* was supported by multiple tests to assess the consistency of LLM responses, yet the variation in LLM performance could affect the reliability of conclusions. Overall, while measures were taken to mitigate these threats, future studies should expand on the diversity of PLAs and LLMs to enhance the robustness and applicability of the findings.

5 RELATED WORK

The evolution of PLA optimization techniques has been substantial, moving from manual, expertise-driven processes to automated, algorithm-based approaches. Early works, such as those by Clements and Northrop [4], laid the groundwork for understanding the PLA complexities. The work of Benavides et al. [2] focuses leveraging computational models to address these complexities, optimizing SPL aspects from feature selection to architectural configuration.

In parallel, developing and applying LLMs in software engineering have opened new avenues for automating tasks that traditionally require deep domain knowledge. Works by Devlin et al. [8] and Brown et al. [3] have demonstrated LLMs' capabilities in understanding and generating human-like text, suggesting their potential utility in interpreting software engineering tasks and providing relevant recommendations or solutions.

Attempts to integrate machine learning into PLA optimization are familiar but have gained momentum with the advent of more sophisticated models. The work of Acher et al [1] adopts Generative languages for reengineering variants into SPLs. Many studies have explored using SBSE to optimize different SPL tasks [12]. However, we have not found a study with goals similar to ours. The integration of LLMs, such as ChatGPT, aims to enhance decision-making in SPL optimization through advanced natural language processing capabilities [16] [24]. This work builds upon these foundations, exploring the relation between PLA optimization techniques and the

latest advancements in LLMs to propose a novel, efficient pathway for optimizing SPLs.

6 CONCLUSION

This research underscores the importance of DM participation in the optimization process, emphasizing how LLMs can reduce the cognitive load by assisting the DM in choosing objective functions. The study presents a novel investigation into integrating LLMs, specifically ChatGPT, for the context of search-based PLA design. It demonstrated that LLMs could significantly enhance decision-making by providing customized objective function suggestions, facilitating more effective PLA optimization. The integration of ChatGPT can make optimization tools more user-friendly and intelligent, enabling software architects to navigate complex optimization scenarios more efficiently.

Domain experts' feedback underscores the potential of LLMs to improve tool usability and effectiveness. While challenges such as response consistency and the need for context-specific inputs were identified, they also highlight areas for future improvement. These insights underscore the transformative potential of incorporating AI into software engineering, suggesting a move towards more adaptable and accessible optimization strategies.

In this context, the study enriches the domains of PLA and AI-enhanced optimization by showcasing how LLM can notably enhance decision-making processes within software engineering, particularly the SBSE area. Through a qualitative evaluation, this research delineates the advantages and identifies opportunities for refinement in deploying LLMs for search-based PLA design. Drawing from these insights, enhancements to the OPLA-Tool are suggested to boost its user-friendliness, the efficacy of training programs, and its comprehensive value in aiding software architects in their work.

Future works include refining LLMs' ability to process diverse optimization contexts of software engineering, integrating LLMs with other PLA optimization components, and expanding their application within SPL Engineering. The approach and ideas evaluated in this work can also be investigated in different SBSE areas, such as search-based refactoring and testing.

ACKNOWLEDGMENTS

This work is supported by CNPq grant 404027/2023-7 and CAPES - Finance Code 001.

REFERENCES

[1] Mathieu Acher and Jabier Martinez. 2023. Generative AI for Reengineering Variants into Software Product Lines: An Experience Report. In *Proceedings of the 27th ACM International Systems and Software Product Line Conference-Volume B*. 57–66.

[2] David Benavides, Sergio Segura, and Antonio Ruiz-Cortés. 2010. Automated analysis of feature models 20 years later: A literature review. *Information Systems* 35, 6 (2010), 615–636.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Paul Clements and Linda Northrop. 2002. *Software Product Lines: Practices and Patterns*. Addison-Wesley Professional.

[5] Thelma Elita Colanzi, Silvia Regina Vergilio, Itana Gimenes, and Willian Nalepa Oizumi. 2014. A search-based approach for software product line design. In *Proceedings of the 18th International Software Product Line Conference (SPLC)*, Vol. 1. 237–241. https://doi.org/10.1145/2648511.2648537

[6] Kalyanmoy Deb and Himanshu Jain. 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on Evolutionary Computation* 18, 4 (Aug 2014), 577–601.

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and Tamt Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. 4171–4186.

[9] Paula M. Donegan and Paulo C. Masiero. 2007. Design Issues in a Component-based Software Product Line.. In *SBCARS*. Citeseer, 3–16.

[10] Willian M. Freire, Murilo Boccardo, Daniel Nouchi, Aline M. M. M. Amaral, Silvia R. Vergilio, Thiago Ferreira, and Thelma E. Colanzi. 2024. Complementary Material. https://doi.org/10.6084/m9.figshare.25556157

[11] Willian Marques Freire, Mamoru Massago, Arthur Cattaneo Zavadski, Aline Maria Malachini Miotto Amaral, and Thelma Elita Colanzi. 2020. OPLA-Tool v2.0: a Tool for Product Line Architecture Design Optimization. In *34th Brazilian Symposium on Software Engineering (SBES)*. Association for Computing Machinery, New York, NY, USA, 818–823.

[12] Mark Harman, S. Afshin Mansouri, and Yuanyuan Zhang. 2012. Search-based software engineering: Trends, techniques and applications. *ACM Computing Surveys (CSUR)* 45, 1 (2012), 11.

[13] Frank van der Linden, Klaus Schmid, and Rommes Eelco. 2007. The product line engineering approach. In *Software Product Lines in Action*. Springer, 3–20. https://doi.org/10.1007/978-3-540-71437-8_1

[14] Mohamed Wiem Mkaouer, Marouane Kessentini, Slim Bechikh, Kalyanmoy Deb, and Mel Ó Cinnéide. 2014. High dimensional search-based software engineering: Finding tradeoffs among 15 objectives for automating software refactoring using NSGA-III. In *Proceedings of the 16th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO'14)*. ACM, Vancouver, Canada, 1263–1270.

[15] Luiz Fernando Okada. 2023. *Guidelines to Support OPLA-Tool Adoption for Novice Users*. Undergraduate Monograph. State University of Maringá, Maringá-PR, Brazil. Advisor(s) Thelma Elita Colanzi Lopes.

[16] Ipek Ozkaya. 2023. Application of large language models to software engineering tasks: Opportunities, risks, and implications. *IEEE Software* 40, 3 (2023), 4–8.

[17] Klaus Pohl, Günter Böckle, and van Der L Frank J. 2005. *Software product line engineering: foundations, principles and techniques* (1 ed.). Springer Science & Business Media.

[18] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1, 8 (2019). https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[20] Marcelo C. B. Santos, Thelma E. Colanzi, Aline M. M. M. Amaral, and Edson OliveiraJr. 2017. Preliminary Study on the Correlation of Objective Functions to Optimize Product-Line Architectures. In *Proceedings of Brazilian Symposium on Software Components, Architectures, and Reuse*. SBCARS'17, Fortaleza-CE, Brazil, 10.

[21] SEI. 2009. *Software Engineering Institute - The Arcade Game Maker Pedagogical Product Line*. https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=485941. Accessed in 2018 August.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).

[23] Yenisei Delgado Verdecia, Thelma Elita Colanzi, Silvia Regina Vergilio, and Marcelo C. Santos. 2017. An Enhanced Evaluation Model for Search-based Product Line Architecture Design. In *20th Ib. Conference on Software Engineering (CIbSE)*. CIbSE, San Jose, Costa Rica, 155–168.

[24] Xiaofei Wang, David Lo, Xin Xia, Shuai Li, and Jianling Sun. 2021. A Survey on Natural Language Processing for Software Engineering. *IEEE Transactions on Software Engineering* (2021).

[25] Lucas Hideki Yamanaka. 2023. Study on the correlation of objective functions to optimize software product line architecture. *Journal of Software Engineering Research* (2023). In Portuguese.