Generative AI for Reengineering Variants into Software Product Lines: An Experience Report
Mathieu Acher, Jabier Martinez

To cite this version:

HAL Id: hal-04160693
https://inria.hal.science/hal-04160693
Submitted on 12 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Generative AI for Reengineering Variants into Software Product Lines: An Experience Report

Mathieu Acher
Univ Rennes, IRISA, Inria, CNRS, IUF
Rennes, France
mathieu.acher@irisa.fr

Jabier Martinez
Tecnalia, Basque Research and Technology Alliance (BRTA)
Pessac, France
jabier.martinez@tecnalia.com

ABSTRACT
The migration and reengineering of existing variants into a software product line (SPL) is an error-prone and time-consuming activity. Many extractive approaches have been proposed, spanning different activities from feature identification and naming to the synthesis of reusable artefacts. In this paper, we explore how large language model (LLM)-based assistants can support domain analysts and developers. We revisit four illustrative cases of the literature where the challenge is to migrate variants written in different formalism (UML class diagrams, Java, GraphML, statecharts). We systematically report on our experience with ChatGPT-4, describing our strategy to prompt LLMs and documenting positive aspects but also failures. We compare the use of LLMs with state-of-the-art approach, BUT4Reuse. While LLMs offer potential in assisting domain analysts and developers in transitioning software variants into SPLs, their intrinsic stochastic nature and restricted ability to manage large variants or complex structures necessitate a semi-automatic approach, complete with careful review, to counteract inaccuracies.

ACM Reference Format:

1 INTRODUCTION
Reengineering variants into a Software Product Line (SPL) is a complex problem [40]. Several industrial cases have been reported in the scientific literature [41] making evident that it is a global and challenging issue for companies seeking to embrace systematic reuse and an overall increase in the quality of their product families.

Numerous automatic techniques have been proposed to analyze the variants, identify the common and variable parts, and linking these parts to features that can be leveraged by an SPL platform [40]. This is related to the n-way merging problem with its compare, match, and compose steps [54]. There are also frameworks which objective is to integrate and seamlessly chain different techniques and to allow the analysis of variants implemented with different programming or modelling languages [44].

Generative AI approaches can be considered a disruptive technology that is questioning how different software engineering tasks are being performed so far. It is important to understand the advantages and limitations to both leverage them and avoiding incorrect expectations. The question that motivates this preliminary analysis is to understand to which extent generative AI technologies can help in the tasks of reengineering variants into SPLs. We defend the idea that large language models (LLMs) can be leveraged to support the migration of variants into SPLs.

LLMs are capable of tackling problems from linguistics, maths, or commonsense reasoning. BERT [23], GPT-3 [13], PaLM [18], to name a few, are scaling to support a variety of tasks such as text generation, question-answering, or text classification [24, 28, 33, 57]. In software engineering, code assistants based on LLMs have been proposed like Alphacode, CodeParrot or Codex [7, 12, 17, 19, 37, 38] and are now deployed at scale for supporting programmers, such as GitHub Copilot [29]. Based on prompts (see green boxes hereafter), composed of an instruction written in natural language that may include existing code, LLMs produce programs and possibly explanations (red boxes hereafter).

Our contribution is to show how LLMs can be concretely and originally used for synthesizing an SPL out of variants. We mainly used and report on sessions with ChatGPT-4 [51], using the Web chat interface of OpenAI. To the best of our knowledge, generative AI and LLMs have not been considered in this context. Machine learning has been applied in SPL and variability engineering, with different use cases [3, 30, 32, 34, 35, 48, 49, 56, 59–64]. The use of natural language processing (NLP) within SPL engineering has caught attention, e.g., for domain analysis and requirements engineering [2, 10, 14–16, 22, 27, 31, 47, 50, 52, 53], but not through LLMs. An exception is a recent paper accepted at SPLC 2023 [1] about programming variability with LLM. However, the problem of synthesizing an SPL out of variants is not considered.

Our objective is also to call the SPL community to start the discussion and conduct more experiments on this topic. We publicly release the material used for the experience report in this repository: https://github.com/acherm/variantsGPT/

2 BACKGROUND
Reengineering variants into an SPL: Clone-and-own is a reuse practice that consists of taking advantage of a similar existing product to adapt it to new requirements through the addition or removal of features. This ad-hoc reuse is accepted in industrial settings because it is simple, available and independent of the assigned developers [25]. However, in the long term, this practice to create variants is not appropriate as the high number of variants and features makes maintenance and management of the whole
product family very difficult. At this point, an SPL approach, in which features are explicitly defined to represent the variability and commonalities between assets needs to be adopted. According to an industrial survey [11], more than fifty percent of the companies that adopt an SPL start from a set of existing variants previously created through ad-hoc reuse techniques.

The process where existing variants are leveraged to adopt an SPL is known as extractive SPL adoption [36]. One of the first steps consists in identifying and locating the features spread among the variants. If the features are not completely known in advance, feature identification and naming must be performed. Feature location techniques rely on heuristics that are sensitive to multiple factors, particularly the quantity and diversity of existing variants [42, 45], as well as the presence of changes not propagated to the rest of variants [26]. These factors directly impact the effectiveness and metrics such as accuracy, precision, and recall. After that, in many cases, the aim is to create a “single-copy” representation of the variants, i.e., a configurable code base containing all the possible features with feature annotations. This maximal single-copy system is also known as the 150% representation, and it is used during the SPL derivation process to automatically create variants through the removal of the undesired features. N-way merging of variants to create 150% representations is a common practice to adopt an SPL using annotative approaches [55].

In order to establish a variability model, it is necessary to undertake the process of discovering feature constraints and then synthesize an understandable abstraction for domain experts. This analysis phase of the reengineering process [4] is very relevant, but in this work we focus on the detection and transformation phases.

**Tool support with BUT4Reuse: Bottom-Up Technologies for Reuse (BUT4Reuse) [44] is a framework to support users during the technical activities of extractive SPL adoption such as feature identification, naming, and location, and the construction of reusable assets for SPL implementation (e.g., creation of 150% representations). Its main advantages are its capacity to be generic and extensible, the significant set of available extensions, and a level of abstraction that allows to seamlessly chain different activities compared to tools that exclusively focus on one of them and most of the time with a unique technique. The framework is generic in the sense that it supports different programming and modelling languages through the concept of Adapters. There are Adapters already available for widely-used languages such as Java, C++, EMF-based models including UML, languages to describe graphs, and new Adapters can be integrated to support the types of artefacts of interest for each case. BUT4Reuse is extensible in the sense that different techniques can be integrated for all the activities, from feature identification, naming, and location, to feature constraints discovery, feature model synthesis, visualization etc. The list and description of the 17 currently available adapters and 23 extensions is available [44].

3 RESEARCH METHODOLOGY

In the remainder of the paper, we consider five cases, involving different variants, kinds of artifacts, and reengineering scenarios. For each case, we report on sessions with ChatGPT-4 [51]. We used the Web chat interface of OpenAI, not the API. The hyperparameters such as temperature cannot be controlled in this setting, and we had to rely on default values.

We started a new chat for each session and case. We have also repeated some sessions for the same case with the same prompt or slight variations of the prompts. The intent was to investigate the robustness of LLM or a potential missing opportunity. In any case, we restarted a new, fresh session between each repetition of a case or between each case. The prompts, written in a green box with ‘(User)’ hereafter, are reported in the different sections and there was no additional context. The interested reader can find more details about sessions in the repository https://github.com/acherm/variantsGPT/.

The session per case involves different interactions and prompts until the presented results were judged satisfactory. Within a session, we, as users, asked for different tasks such as:

- domain analysis, summary of commonalities and differences, as plain languages or as tables, identification of features
- synthesis of an integrated, 150% model or code (and possibly a visualization)
- synthesis of a template-based generator
- refactoring of models or code
- corrections of code, model, or explanations that were perceived as inaccurate or simply unusable.

Hence, sessions were highly interactive. The first author of the paper wrote the prompts and made the interactive sessions. The two authors have reviewed and assessed the different sessions, throughout and after the experience.

4 CASE 1 AND 2: BANKING SYSTEMS VARIANTS IN JAVA AND UML

We start with Case 1 and 2. Our first challenge involves three variants for different banking systems used in previous works [5, 6, 65, 66]. In this scenario, we have created three model variants through ad-hoc reuse with variations on the limit of bank withdrawal, the consortium entity and currency conversion. The challenge is to factor out the different variants and re-engineer them as an SPL. We consider two cases: (1) the first case involves variants implemented in Java; (2) the second case involved variants implemented in UML class diagrams. Figure 1 shows a visual representation of the UML class diagrams, using PlantUML notation. Listing 1 provides an excerpt of the Java code corresponding to the first variant of Figure 1a.

4.1 Case 1: Java variants

In this first setting, we have converted the three UML variants depicted in Figure 1 as Java code. The implementation resides in a unique file with the different classes, fields, and methods. Methods returning void have an empty body, while methods returning a double have a basic default implementation, the same for all variants. The overall challenge is to build an SPL out of the 3 Java variants.

(User) I have three variants of a Java implementation of a banking system. The first one is:

1In the rest, we use ChatGPT-4, ChatGPT and LLM alternately, but the three terms refer to the same thing for this paper.
Generative AI for Reengineering Variants into Software Product Lines: An Experience Report
SPLC '23, August 28-September 1, 2023, Tokyo, Japan

import java.util.List;

public class Variant1 {
    class Account {
        private String id;
        private double amount;
        private double limit;
        public void deposit(double amount) {
            //...
        }
    }
    class Client {
        private String id;
        private String name;
    }
    class Consortium {
        // Empty class as per the description.
    }
}

Listing 1: Java (first variant, corresponding to Figure 1a)

(User) The second variant is: ... The the third variant is: ... Considering the three variants, can you summarize: (1) what's common (2) what varies?

(LLM) Sure, I can summarize the common elements and the differences between the three Java variants:
Common Elements
All variants have Account, Bank, and Client classes. Each Account has a String id and a double amount. The Account class in each variant has deposit, getAmount methods. Each Bank class has methods for depositOnAccount and withdrawFromAccount. Each Client class has a String id and String name.

Differences ... Variant3: The Account class has an int currency, getCurrency, and setCurrency methods similar to Variant2. But it introduces withdrawWithoutLimit instead of withdrawWithLimit. The Bank class has a List of Account objects, a Converter, a Consortium, and a convert method. This is a mix of features from both Variant1 and Variant2. The Consortium class returns, but it’s now along with the Converter class from Variant2.

In general, the response is quite precise, but it should not be assumed as unquestionable. In particular, there are ambiguous explanations (e.g., there is no apparent link between Consortium and Converter). We continue by asking to build an annotated SPL with templating directives.
The template is correct and variant-specific elements are correctly annotated. We move forward, and challenge the LLM-based assistant to refactor the SPL (and template), this time with a feature-oriented perspective.

(User) Instead of having “variant” as a variable to configure the template, I’d like to have other variables that represent “features” of the banking system... can you propose some features?

Listing 2: Java (template-based SPL)

```java
import java.util.List;

public class Account {
    private String id;
    private double amount;
    private double limit;

    public Account(String id, double amount) {
        this.id = id;
        this.amount = amount;
    }

    public void deposit(double amount) {
        amount += this.limit;
        this.amount += amount;
    }

    public double getAmount() {
        return this.amount;
    }

    public void withdrawWithoutLimit(double amount) {
        amount -= this.limit;
        this.amount -= amount;
    }

    public void withdrawWithLimit(double amount) {
        amount -= amount;
        this.amount -= amount;
    }

    public void setAmount(double amount) {
        this.amount = amount;
    }

    public void setLimit(double limit) {
        this.limit = limit;
    }

    public void setCurrency(String currency) {
    }
}
```

(Listing 3: Improved template-based SPL with created features)

The new template allows one to configure variants in terms of 4 features (hasWithdrawalLimit, hasConsortium, supportsCurrency, isSingleAccount). The LLM-based assistant can provide a synthesis of the situation, with a product comparison matrix.

(User) Can you create a table that describes whether feature variables are supported by Variant1, Variant2, and Variant3?

(LLL) Deriving variants out of SPL: Out of the template of Listing 2 or Listing 3, we can successfully derive and retrieve the original three variants. The identification of features (see Listing 3) allows one to envision to derive new variants, beyond the three original ones. Stated differently, Table 1 could be expanded through new combinations of features and thus variants. Obviously, not all combinations of features are syntactically valid or would make sense from a domain point of view. Hence, we consider it is the responsibility of developers to build a feature model in order to properly
The challenge is to build an SPL out of the 3 UML variants. We did not use generative AI in this context. We leave as future work the problem of automating the synthesis of feature model with an assistant.

**Syntactical and semantics errors**: Throughout our interactions, we were exposed with two kinds of errors.

*Misleading variants*. The correct template of Listing 2 has been obtained after several iterations and fixes. Indeed, there was a subtle yet severe error during the session with the assistant. The templating code was unsound at some point and needs to be fixed. Specifically, the field limit was not included in Variant2 since the conditional was wrong. It requires to further interact as follows.

```
(User) I don’t understand why:
[% if variant == 2 %]
  private double limit;
[% endif %]
```

since the field limit is included in Variant2.

The fix was correctly provided after some interactions. This kind of error not only increases the cognitive load and effort. There is also the risk that the error goes unnoticed, thereby misleadingly authorizing unsafe variants that lack certain Java fields.

*Syntactical error affecting the whole SPL*. In Listing 3, the careful reader can notice that the method getAmount has no ending bracket. It is a real bug that causes compilation errors for all variants. The error has been fixed, and a new template has been proposed. However, it requires developers to locate the error (after the observation of compilation failures) and eventually interact again with the assistant.

```
 private double limit;
```

... (User) I’ll give you three variants of an UML model... the first one is:

```
<uml : Model xmi : version = "20110701" encoding = "UTF-8"> 
</uml : Model>
```

(LLL) Based on the provided UML model, we have a banking system with four main classes: Account, Bank, Client, and Consortium. Below is a summary of these classes and their attributes and operations:

1. Account Class
   - Attributes:
     - id (string)
     - amount (double)
     - limit (double)
   - Methods:
     - addClient

```
LLM) Comments of the assistant show a good syntactical understanding of the UML model.

(User) Here is the second UML model... here is the third UML model variant... considering the three UML variants, can you summarize: (1) what’s common (2) what varies?

(LLL) After analyzing the three UML model variants, here is a summary of what’s common and what varies among them: ...

All three variants have the following classes:
- Account - Bank - Client - Converter

These classes share some common attributes and operations: …

The main differences among the three UML model variants are as follows: Variant 1: - The Account class has an additional operation: withdraw. - The Bank class has an additional operation: addClient. - The Client class has an additional attribute: accounts.

```
</uml : Model>
```

... (User) Can you create an UML model that can represent the three variants? I’d like to have the result in PlantUML ... can you use a templating engine (like jinja2) to annotate variant-specific elements of the PlantUML model?
with the LLM-based assistant, and came to a similar conclusion to handle for transformers-like generative AI. It then becomes more problematic textual information and tokens. It then becomes more problematic explanation is that the UML models tend to exhibit much more hypothesis is that UML-like artefacts contain a different structure that hypothesis is that the generative AI technology was less mature ants, despite being on the same domain and banking system. A case of UML variants was less favorable than the case of Java variants among the three variants exhibits several errors. When asking a comparison table, we noticed similar trends: some elements are considered as common whereas it is not the case; some elements are wrongly considered as specific; some elements are simply made up (new operations, but also new classes like Manager that simply do not exist). inaccurate analysis of commonality and differences: the previous excerpt of the explanations about commonalities and differences among the three variants exhibits several errors. When asking a comparison table, we noticed similar trends: some elements are considered as common whereas it is not the case; some elements are wrongly considered as specific; some elements are simply made up (new operations, but also new classes like Manager that simply do not exist).

• inaccurate integrated model: we asked to build an integrated, 150% model without annotation, and some elements were made up or repeated, leading to an inaccurate model.
• inaccurate synthesis of template with wrong conditionals, invented elements, or unnecessary repetition of elements.
• inaccurate visualization: we also asked to produce a 150% model with colors in PlantUML to showcase specific elements. The resulting visualization was incomplete and wrong.

Obviously, the four limitations are related, i.e., a lack of understanding of what is common and varies has a ripple effect on the integrated model, possible template visualization. Overall, the case of UML variants was less favorable than the case of Java variants, despite being on the same domain and banking system. A hypothesis is that the generative AI technology was less mature to handle the UML language than the Java language. Another hypothesis is that UML-like artefacts contain a different structure that challenge the matching and merging of elements. Yet another possible explanation is that the UML models tend to exhibit much more textual information and tokens. It then becomes more problematic to handle for transformers-like generative AI.

Listed: Template of UML class diagram (excerpt)

First, the field accounts should not exist in class Client. Second, the conditional is wrong: Consortium is included in Variant 1, not in Variant 2. Same error applies for cons (see Listing 5).

Lessons and hypothesis. Compared to the case of Java, this session is less supportive of the use of generative AI. We repeated the session with different prompts and strategies for interacting with the LLM-based assistant, and came to a similar conclusion with negative experiences. We report on following limits:

• inaccurate analysis of commonality and differences: the previous excerpt of the explanations about commonalities and differences among the three variants exhibits several errors. When asking a comparison table, we noticed similar trends: some elements are considered as common whereas it is not the case; some elements are wrongly considered as specific; some elements are simply made up (new operations, but also new classes like Manager that simply do not exist).
• inaccurate integrated model: we asked to build an integrated, 150% model without annotation, and some elements were made up or repeated, leading to an inaccurate model.
• inaccurate synthesis of template with wrong conditionals, invented elements, or unnecessary repetition of elements.
• inaccurate visualization: we also asked to produce a 150% model with colors in PlantUML to showcase specific elements. The resulting visualization was incomplete and wrong.

We then continue with the overarching goal of getting a visualization of a 150% SPL that would represent all variants. We employ a step-by-step strategy, with a series of prompt/answer.

(User) Can you create an integrated model that would represent the three variants? I’d like to have the result in graphml as well... can you translate it to dot format? ... can you add color to edges that are specific to variant 1, variant 2, or variant 3? Says variant 1 is blue color, variant 2 is green color, variant 3 is red color, and if it’s common to all variants, it remains black

The final result is depicted in Figure 2, after a conversion from DOT to PDF. The 150% model is correct, and it is straightforward to engineer an annotated SPL (e.g., with a template-based approach, as previously illustrated). It is also possible to envision expanding the SPL with new variants.

Summary: The experience for Case 3 is positive, since the resulting SPL correctly represents the original variants.

5 CASE 3: GRAPHML VARIANTS

In this setting, we consider three variants in GraphML format. The case involves a different formalism (capable of representing a state transition system), language, and technological space (GraphML). It has also been used in the context of education, academic presentation, and BUT4Reuse⁴.

We start the session in a very similar fashion.

(User) I’ll give you three variants of a graphml model... Here is the first one: here is the second variant: ... here is the third variant: ...

We continue with the overarching goal of getting a visualization of a 150% SPL that would represent all variants. We employ a step-by-step strategy, with a series of prompt/answer.

(User) Can you create an integrated model that would represent the three variants? I’d like to have the result in graphml as well... can you translate it to dot format? ... can you add color to edges that are specific to variant 1, variant 2, or variant 3? Says variant 1 is blue color, variant 2 is green color, variant 3 is red color, and if it’s common to all variants, it remains black

The final result is depicted in Figure 2, after a conversion from DOT to PDF. The 150% model is correct, and it is straightforward to engineer an annotated SPL (e.g., with a template-based approach, as previously illustrated). It is also possible to envision expanding the SPL with new variants.

Summary: The experience for Case 3 is positive, since the resulting SPL correctly represents the original variants.

6 CASE 4: STATE CHART VARIANTS

This case involves vending machines’ model variants, encoded as state chart and using SCT format. Originally, six variants were
Figure 2: 150% model with colors (out of DOT format)

Figure 3: 150% state chart model

used. A first observation and lesson learned is that the size of the context3 (and thus prompts) precludes to consider all variants and information. Furthermore, the SCM format is composed of a state chart and a layout description. We did not consider the layout part to reduce the size of information and tokens fed to LLM. Using ChatGPT-4, we were thus forced to only consider four variants.

We followed a similar methodology as for previous cases. Due to space limitations, we report on key results and observations. Figure 3 shows a 150% model augmented with colors and depicted as PDF. ChatGPT-4 generated the DOT file after some prompts and interactions.

Limitations. There are serious errors in Figure 3. First, ChatGPT-4 takes the liberty of merging edges and transitions (for example, enterSodaCode, enterCoffeeCode, enterTeaCode). Second, some presence conditions are misleading (e.g., always transition between Product price displayed and Select payment method is not specific to a variant). Third, some transitions of the original variants are not included. When repeating our experiments with the same prompts or variations thereof, we observed similar limitations. There were sessions where the 150% model contains duplicated transitions and nodes (states). It leads to more accurate models (e.g., some transitions were this time included). However, the resulting models tend to hard to maintain; the semantics becomes ambiguous (same transition out of the same named states) and can be problematic to handle by modelling tools. Such an example is given in Figure 4, considering 3 variants instead of 4.

Summary: The experience for Case 4 is negative, since the semantics of the SCM formalism is not properly handled and important information is either lost or wrongly expressed, leading to misleading SPL.

7 CASE 5: PNG VARIANTS

In this setting, we have four images created through the copy-paste-modify reuse paradigm. This case has been considered as part of BUT4Reuse4. The images are in Portable Network Graphics (PNG) format. Each PNG file begins with an 8-byte signature and contains a series of chunks, which are binary data blocks. As a result, producing a "diff" among binaries' formats is likely to produce irrelevant information. To avoid directly operating over PNG, BUT4Reuse proposes to use an adapter based on pixels where the similarity between pixels of variants is based on checking that they share the same x y position and a color similarity formula. With current LLMs, we confirm (1) the impossibility to analyze PNG, binary files and compute a relevant diff – there are no magical solutions involved; (2) the provision of pixels location to LLM does not scale: there is simply too much information to feed in; (3) LLM is not helpful to find meaningful features: the information is too low-level and there is no semantic to infer.

Summary: The experience for Case 5 is that LLM cannot be applied "as such" for PNG, since the information is low-level and without an exploitable structure. This experience illustrates the case of binary files and the need of humans providing the desired abstraction level and means for the analysis as in traditional reengineering approaches.

https://github.com/but4reuse/but4reuse/wiki/Tutorial
8 DISCUSSION

Threats to validity: The major threat to validity is related to the prompts used in our experiments. Although we believe the used prompts conveyed our intentions, and we attempted to enhance them based on response quality, we cannot be certain that further specifications or clarifications in the prompt would have yielded improved results. We also suspect different LLMs (or LLMs with different hyperparameters like temperature) could lead to different outcomes.

Size of the input variants: Current LLMs cannot handle a large code base or numerous variants, since both can increase the number of tokens supported during inference. It is an open question how LLMs can manage long-term working memory and large context window [20, 21]. Though progress can be envisioned, this is currently a severe limitation for real-world usage.

Variants as binary files or complex structures: As demonstrated in the case of images in PNG format, LLMs were incapable of synthesizing differences and commonalities due to the non-exploitable nature of the information at the binary file level. This limitation stands in contrast to the effectiveness of BUT4Reuse Adapters in transforming input variants into usable structures. The key idea lies in operating at a different level of abstraction, specifically analyzing the commonalities and differences within the pixels themselves.

Consistency among different executions: BUT4Reuse put emphasis on obtaining deterministic results across different executions. With ChatGPT, the results vary depending on the prompt, context, and other factors, so its users should know that partially similar responses could be expected across different executions.

Genericity in terms of variants’ programming or modelling language: ChatGPT has been probably trained with almost all existing programming languages where source code was available. While the amount of information will vary depending on if it is a widely used programming or modelling language or a niche language, there is no need to adapt or provide information about the language itself for comparing variants. Despite that the reported Adapters for BUT4Reuse were relatively easy and not very time-consuming to develop (they usually rely on existing parsing libraries or there is already a generic modelling artefacts adapter that can be specialized), it is a clear advantage not having to design and provide an Adapter. When asking ChatGPT about how it calculated the common and variable parts, for instance for the Java variants, we found a similar reasoning as the one that will follow the Adapter decomposing the source code in Classes, Methods, and Fields, and then using set theory to determine if it is common or variable.

Similarity function of the Adapter: The most challenging function of a programming or modelling language Adapter is the similarity between elements. While the Adapter will allow having more control on the similarity function among elements, LLMs hide the complexity of its development and can potentially internally define more complex semantic similarity relations. In fact, one of the most relevant concerns about id-based matching is how easy it fails to refactorings in the source code. When trying some examples from a dataset of equivalent and non-equivalent program pairs (EqBench) [9] we can see that the performance of ChatGPT is not optimal, at least not from its direct response. For instance, if we provide the Java pair in pow test Eq and asks to check if they are equivalent:

(\text{LLM}) \text{ Are you sure that this alteration is not the same condition?}

\text{(User) Apologies for the confusion. Upon further review, I can see that the altered condition in the newV class is, in fact, equivalent to the original condition in the oldV class.}

This kind of experiments raise questions about the capacity of LLMs to correctly detect source code clones when responding to logical issues that can be more complicated than this example.

Extensible for visualizations: Visualization paradigms in the evolution of variants can be very diverse [46]. BUT4Reuse is extensible for integrating new visualizations, but they need to be implemented and integrated. While the currently available visualizations are specialized for the reengineering activities, with ChatGPT it is possible to ask to manipulate and show the results in different formats such as tables, graphs, and many other approaches, thus potentially focusing on any kind of perspective for the analysis task at hand.

Feature naming: BUT4Reuse integrates the VariClouds visualization for feature naming and to automatically set or suggest feature names [43]. ChatGPT is a powerful tool for summarizing source code so feature naming and creating feature descriptions through LLM could easily outperform previous approaches.

Benchmarking: The state-of-the-art in benchmarking approaches and techniques for the evolution of variants has been analyzed [58] showing significant gaps and evidence that the field could benefit from the design of new benchmarks respecting a minimum of established requirements. The existence of benchmarks are specially useful with the emergence of new techniques to quickly and unambiguously identify the strengths and limitations compared to previous approaches. Besides that, in the case of LLMs we are confronted with the limitations on the size of the information we can provide to ChatGPT while prompting, as most of the benchmarks are real systems of relatively big size.

Variants synchronization and filtering prior to SPL extraction: Fenske et al. [26] used clone detection techniques to try to synchronize variants towards simpler SPL extractions. Other more drastic solutions considered filtering outlier variants to facilitate this process [8, 39]. It will be worth exploring to which extent ChatGPT could help in that task. For instance, we repeated the Java Banking systems experiment providing again the three variants but Variant1 with the class, method, and field names in Spanish instead of English. Apparently the common and variable parts seem to take into account that they are the same even if the name was a translation, but with more difficulties and further incorrect explanations. However, it was possible to ask ChatGPT to translate Variant1 using the
naming conventions in Variant2 and Variant3 while keeping the elements of Variant1. The results were fine except for a small detail ("depositIntoAccount" should have been "depositOnAccount") that were recognized by ChatGPT as a confusion.

Complementary: Instead of thinking of ChatGPT as a replacement of BUT4Reuse or similar frameworks, the integration of both could be a more beneficial solution for the moment. For instance, the framework can create template-based prompts to get the ChatGPT results for tasks where the value has been proven.

9 CONCLUSION
We showed how LLMs can assist domain analysts and developers in migrating variants as software product lines. The inaccuracies during different steps suggest a semi-automatic approach, where it is mandatory to carefully review the generated code. We identify major barriers: LLMs are by construction stochastic and sensitive to prompt variations and context. The size of the input is also limited: It is difficult for these techniques to handle numerous variants, large code bases, or the underlying structures of binary files.

The directions of further work are multiple. This report is a call to the SPL community to start thinking about the synergies of LLMs and reengineering variants. We believe that the research direction that could have the most practical impact is the integration of LLMs with currently available tools, as well as supervision support.

Acknowledgements
We would like to acknowledge the anonymous reviewers of VariVoLution 2023 for their valuable input and feedback.

REFERENCES