

# Large Language Models to the Rescue: Reducing the Complexity in Scientific Workflow Development Using ChatGPT

Mario Sanger<sup>a,\*</sup>, Ninon De Mecquenem<sup>a</sup>, Katarzyna Ewa Lewińska<sup>b</sup>, Vasilis Bountris<sup>a</sup>, Fabian Lehmann<sup>a</sup>, Ulf Leser<sup>a,\*</sup>,  
Thomas Kosch<sup>a,\*</sup>

<sup>a</sup>Humboldt Universitat zu Berlin, Department of Computer Science, Unter den Linden 6, Berlin, 10099, Germany

<sup>b</sup>Humboldt Universitat zu Berlin, Department of Geography, Unter den Linden 6, Berlin, 10099, Germany

---

## Abstract

Scientific workflow systems are increasingly popular for expressing and executing complex data analysis pipelines over large datasets, as they offer reproducibility, dependability, and scalability of analyses by automatic parallelization on large compute clusters. However, implementing workflows is difficult due to the involvement of many black-box tools and the deep infrastructure stack necessary for their execution. Simultaneously, user-supporting tools are rare, and the number of available examples is much lower than in classical programming languages. To address these challenges, we investigate the efficiency of Large Language Models (LLMs), specifically ChatGPT, to support users when dealing with scientific workflows. We performed three user studies in two scientific domains to evaluate ChatGPT for comprehending, adapting, and extending workflows. Our results indicate that LLMs efficiently interpret workflows but achieve lower performance for exchanging components or purposeful workflow extensions. We characterize their limitations in these challenging scenarios and suggest future research directions.

**Keywords:** Large Language Models, Scientific Workflows, User Support, ChatGPT

---

## 1. Introduction

Large-scale data analysis pipelines (also known as scientific workflows) are crucial in driving research advances for natural sciences [1]. They are pivotal in accelerating large and complex data analysis on distributed infrastructures and offer essential features, such as reproducibility and dependability [2]. In bioinformatics, for instance, scientific workflows are analyzing the terabyte-large data sets produced by modern DNA or RNA sequencing machines in a wide variety of experiments [3], thereby aiding in building a comprehensive understanding of biological processes and human diseases. Bioinformatics workflows typically include many individual computational steps, such as data pre-processing, extensive quality control, aggregation of raw sequencing data into consensus sequences, machine-learning-based tasks for classification and clustering, statistical assessments, and result visualization. Each step is carried out by a specific program, typically not written by the workflow developer but exchanged within a worldwide community of researchers [4]. Execution of a workflow on a distributed infrastructure, in principle, is taken care of by a workflow engine; however, the idiosyncrasies of the different infrastructures (e.g., file system, number and features of compute nodes, applied resource manager, and scheduler) often require workflow users to tune their scripts individually for every new system [5].

However, typical developers of workflows are researchers from heterogeneous scientific fields who possess expertise in their respective domains but often lack in-depth knowledge in

software development or distributed computing. They often encounter difficulties understanding the complex implementations of exchanged codes and the deep infrastructure stack necessary for their distributed execution. This situation challenges efficient workflow implementation, slows down or hinders data exploration and scientific innovation processes [6]. Consequently, low human productivity is a significant bottleneck in the creation, adaption, and interpretation of scientific workflows [7].

Parallel to this, there is a well-established field in human-computer interaction focusing on assisting end-user programmers and software development, as highlighted by previous work [8, 9, 10, 11], to reduce the perceived cognitive workload and improve the overall programmer performance [12]. Research in this field includes programming-by-demonstration [13, 14], visual programming [15, 16], and natural language instructions [17]. Recent work in this area particularly investigated prospects of general-purpose Large Language Models (LLMs), such as ChatGPT [18], LLaMA [19] and Bloom[20], for supporting end-user-programming [21, 22, 23] and software development in general [24, 25]. For instance, Bimbatti et al. [21] explore using ChatGPT to enhance natural language understanding within an end-user development environment, assisting non-expert users in developing programs for collaborative robots. Moreover, White et al. [24] introduce prompt design techniques for automating typical software engineering tasks, including ensuring code independence from third-party libraries and generating an API specification from a list of requirements. Surameery et al. [25] evaluate LLMs for supporting code debugging. However, results from such studies, which focus on standard programming languages, cannot easily be transferred to workflow

---

\*Corresponding authors

systems. Workflow scripts mostly call external tools with agnostic names and have little recognizable control structures or protected keywords. Publicly available examples are scarce; for instance, the community repository of the popular workflow system Nextflow [26] currently offers only 55 released workflows<sup>1</sup>. Furthermore, workflows can only be understood when the distributed system underlying their execution is considered, creating dependencies much different than usual programs. Moreover, studies investigating how LLMs support users in data science - the field in which workflows are applied extensively - do not address the unique characteristics of scientific workflows either and are limited to theoretical considerations [27, 28]. Practical studies, especially those involving real users, are badly missing.

In this work, we address these shortcomings by describing three user-studies in two different scientific fields (biomedicine and Earth observation) that evaluate the suitability of ChatGPT for comprehending, modifying, and extending scientific workflows. Specifically, we evaluate the correctness of ChatGPT regarding explainability, exchange of software components, and extension when providing real-world scientific workflow scripts. Our results show a high accuracy for comprehending and explaining scientific workflows but reduced performance for modifying and extending workflow scripts. The domain experts positively assessed the explainability in qualitative inquiries, emphasizing the time-saving capabilities of using LLMs while engineering existing workflows. Overall, our work indicates that general-purpose LLMs have the potential to improve human performance when analyzing complex scientific workflows.

## 2. Related Work

Previous research investigated how related domains, such as programming, can be augmented using interactive technologies [11, 29]. In contrast to programming, where applications use a single programming language and are often executed on a single system, scientific workflows combine multiple software artifacts on distributed stacks for advanced data processing. We ground the reader by providing a literature review about scientific workflows and introducing large language models, including their utility to facilitate the creation of software artifacts.

### 2.1. Scientific Workflows

Scientific workflows are widely used by diverse research communities, such as biomedicine [30], astronomy [31], climatology [32], and Earth observation [33] to manage the dataflow and distributed execution of complex analyses, simulations, and experiments. A scientific workflow comprises a series of interconnected computational steps, often with diverse patterns of dependencies, that define how to process and analyze data to reach a particular research objective. Scientific workflows can be regarded as directed acyclic graphs in which the nodes represent computational tasks or operations and edges model dependencies or dataflow between these tasks. An edge from one

node to another signifies that the output of the first task is used as input for the second [34]. For example, Figure 1a illustrates the computational steps, the tools, and the data flow of a bioinformatics workflow for performing differential gene expression analysis. During workflow execution, a single computation step often involves multiple processes, which are typically executed in a distributed fashion on different machines and batches of the input data, resulting in a much more complex execution graph. Consequently, scientific workflows help facilitate the reproducibility and traceability of data analyses by explicitly outlining the steps and parameters involved [35]. Furthermore, they allow for automation, scaling, and optimization of computational processes, which is especially critical in disciplines dealing with large datasets [36]. The increasing importance of scientific workflows for scientific progress has led to a growing interest in developing more user-friendly tools and methods through the research community. Scientific workflow management systems, like Apache Airflow [37], Galaxy [38], Nextflow [39], Pegasus [40], and Snakemake [41], are specifically developed to support users in designing and executing scientific workflows in various aspects. Key features of such management systems typically include workflow design and composition, (distributed) workflow execution and scheduling, provenance tracking, recovery and failure handling, and resource management [35]. Figure 1b highlights the implementation of the example workflow as well as a single computational step (see Figure 1c), i.e., reference genome alignment using the STAR toolkit in Nextflow.

Scientific workflows are often reused and adopted for complex data analysis [6]. Most of the time, users of scientific workflows are unaware of the workflow’s internal functionality and technical details. Instead, users of scientific workflows represent domain experts, such as mathematicians, physicians, or bioinformatics, who are experts in their respective domains but not necessarily in programming and interpreting scientific workflows. Existing scientific workflows were implemented and maintained by persons other than the domain user. This reduces the direct interaction with scientific workflows to a minimum, where domain experts only hand in the input data and evaluate the output data. Consequently, domain experts using a workflow often do not have the knowledge to modify, extend, or interpret the details of scientific workflows.

### 2.2. Large Language Models

Language Models, such as BERT [42], GPT-3 [43], Bloom [20] and PaLM-2 [44], build the foundation of many recent advancements in natural language processing and understanding. These models have billions of parameters and are generally pre-trained on vast sets of texts from the web and other repositories, enabling them to encode syntactic and semantic relationships in human language. As a consequence, generative language models, such as ChatGPT, LLaMA [19], and LAMDA [45], can produce machine-generated high-quality text that is indistinguishable from human writing. These generative capabilities have empowered these models to assist in diverse (creative) writing tasks and have been utilized to facilitate a wide range

<sup>1</sup><https://nf-co.re/stats> - last access 20-10-2023

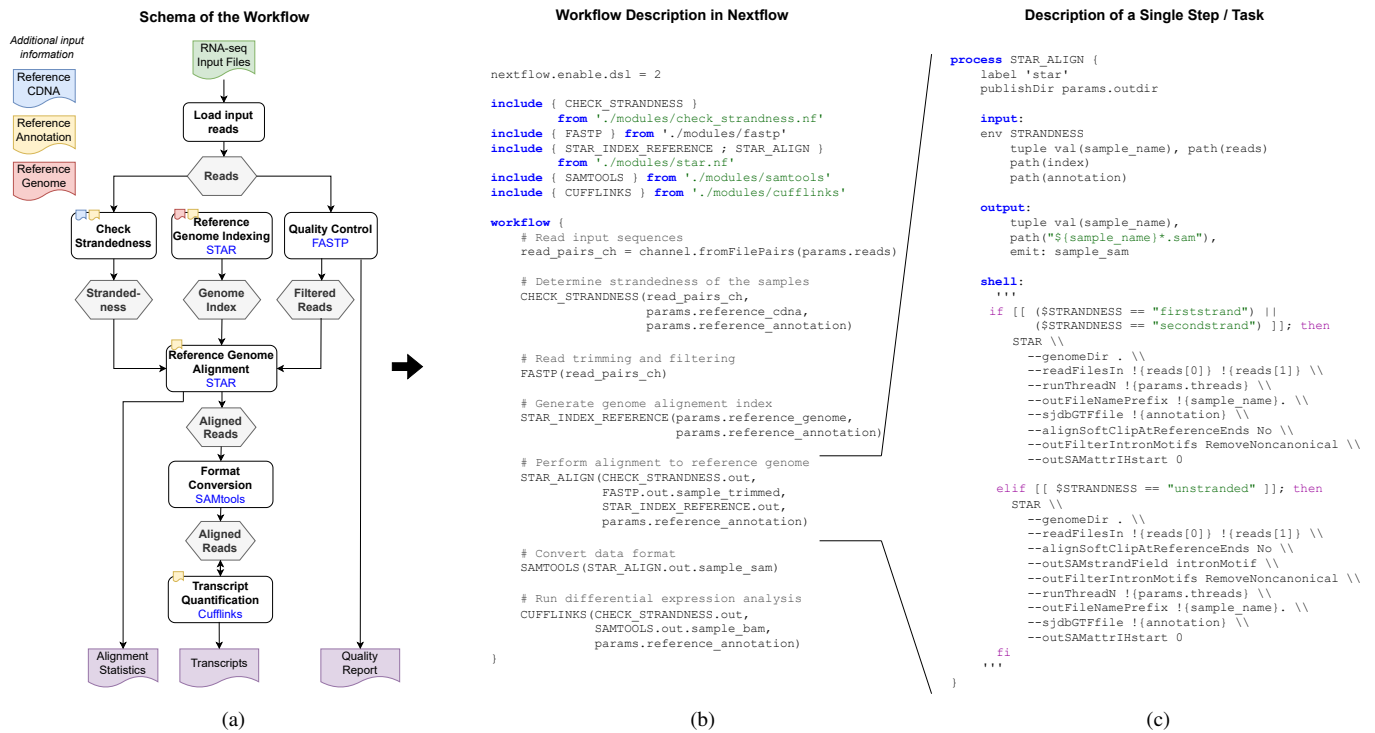


Figure 1: Example bioinformatics workflow for differential gene expression analysis created by a domain expert recruited in our study. The figure highlights the conceptual schema of the workflow (a), its implementation in Nextflow (b), and the implementation of one single step (c), i.e., reference genome alignment using the STAR tool. The workflow comprises six computational steps in total. For each step the used tool is given in blue below the task name.

of interactive language-based applications within the HCI community [46, 47, 48, 49, 50, 51]. For instance, WordCraft [46] investigate the utilization of LLMs to aid fiction writers in tasks ranging from transforming a text to resemble a “Dickensian” style to providing suggestions to combat writer’s block. Their findings indicate that writers found such text-generating models beneficial even when the generated text is not perfect. Petridis et al. [47] introduce AngleKindling, an interactive tool that employs LLMs to support journalists exploring different angles for reporting on a press release. Their study with twelve professional journalists shows that participants found the system considerably more helpful and less mentally demanding than competitor brainstorming tools. Other applications include prototyping support [48], generation of titles and synopses from keywords [51], conversational interactions with mobile user interfaces [49], and conceptual blending [50].

Next to their text writing capabilities, language models are further known to retain commonsense knowledge within their training data, effectively transforming them into accessible knowledge stores that can be seamlessly queried using natural language prompts. For instance, experiments with BERT [42] highlight that the model performance is competitive with traditional information extraction and open-domain question answering. Furthermore, recent studies show the potential of using ChatGPT for knowledge base construction, inspired by the fact that these language models have been pre-trained on vast internet-scale corpora that encompass diverse knowledge domains [52]. However, it is worth noting that LLMs are known to frequently

generate hallucinations, which are outputs that, while statistically plausible and seemingly believable, are factually incorrect [53, 54].

### 2.3. Using LLMs to Support Programming

The ability to generate new text and to reconstruct existing information makes LLMs highly appropriate to support users in software development, as programming often requires not only the creation of novel code segments tailored to current requirements and tasks but also depends on the application of established algorithms, software libraries, and best practices. Accordingly, a large number of papers investigate LLMs specially trained for code generation [55, 56, 57, 58] as well as different approaches leveraging these models to provide interactive programmer support [59, 60, 61, 23]. For instance, Jiang et al. [60] discuss GenLine, a natural language code synthesis tool based on a generative LLM and problem-specific prompts for creating or changing program code. The findings from a user study indicate that the approach can provide valuable support to developers. However, they also encounter several challenges, such as participants finding it difficult to form an accurate mental model of the kinds of requests that the model can reliably translate. Similarly, Vaithilingam et al. [61] conducted a user study with 24 participants evaluating their usage and experiences using the GitHub Copilot<sup>2</sup> code generation model while

<sup>2</sup><https://copilot.github.com> - last access 20-10-2023

programming. The authors find that the synthesized code often provided a helpful starting point and saved online searching efforts. However, participants encountered issues with understanding, editing, and debugging code snippets from Copilot, resulting in not necessarily improved task completion times and success rates. These findings align with the results of similar studies [62]. However, a controlled experiment in [63] records a positive effect of code generators when used in introductory programming courses for minors.

The use of generative LLMs and code generators has been scarcely explored in scientific data analysis and not yet for scientific workflows. Liu et al. [23] examine the Codex code generator [55] in the context of data analysis in spreadsheets for non-expert end-user programmers. Moreover, several studies investigate the utilization for data visualization [64, 65]. For example, the study by Maddigan et al. [64] evaluates the efficiency of ChatGPT, Codex, and GPT-3 in producing scripts to create visualizations based on natural language queries. The studies that have the most overlap with our work regarding the intention to support the design of data analysis pipelines are given by Ubani et al. [65] and Zahra et al. [66]. In the case of the former, ChatGPT is used to build a conversational, natural language-based interface between users and the scikit-learn machine learning framework [67] supporting users in several phases of a machine learning project ranging from initial task formulation to comprehensive result interpretation. For the latter, Laminar, a framework for serverless computing, is proposed, which offers possibilities for code searching, summarization, and completion. However, the framework is solely focused on Python implementations.

### 3. Methodology

This section describes the study methodology. We begin by outlining about the general study approach. Then, we explain the research process for each experiment. Based on related work and the objectives of our research, we state the following research questions:

- RQ1:** How performant is ChatGPT for comprehending and explaining scientific workflows?
- RQ2:** How suitable is ChatGPT in suggesting and applying modifications for scientific workflows?
- RQ3:** How efficient is ChatGPT in extending scientific workflows?

#### 3.1. General Study Design

To answer our research questions, we investigate the capabilities of ChatGPT, a widely-used LLM, to comprehend existing workflow descriptions (cf. Study I), to exchange tools used within a workflow (cf. Study II), and to extend a partially given workflow (cf. Study III) using three distinct user studies. We select these use cases as understanding the data flow and the analysis performed is essential for successfully applying scientific workflows. Moreover, exchanging tools and extending a partially given workflow are common use cases in adapting and

reusing existing workflows in the work context of domain scientists [6]. For each study, we specially design conversational prompts simulating the interaction between a user working with workflows and ChatGPT. For our studies, we leverage version GPT-3.5 of ChatGPT<sup>3</sup>. We decided to use GPT-3.5 since it is openly available to the public and allows other researchers to reproduce our investigations without additional incurring costs<sup>4</sup>. Additionally, we develop distinct questionnaires for evaluating the output of ChatGPT by the domain experts for each study. While conducting a study, we present a brief overview of the study’s overall goal and the developed questionnaire to the experts. Subsequently, the experts complete the questionnaire independently without the experimenters’ support. This procedure is intended as participants were not pressured by a time limit and could freely allocate their time for the study. Furthermore, we intend to avoid a Hawthorne effect, where participants can provide biased responses due to the presence of observers [68].

#### 3.2. Participants

Throughout all experiments, we recruited one expert from bioinformatics and three experts working on Earth observation workflows. Using scientific workflows is common in these two areas. In bioinformatics, scientific workflows are an important tool for enabling the automation and documentation of complex data analysis processes, ensuring reproducibility and transparency in research [39]. In Earth observation, scientific workflows streamline the complex process of acquiring, processing, and analyzing vast amounts of satellite and sensor data, enhancing the efficiency and accuracy of environmental studies [70]. Hence, scientific workflows have become commonplace in these two areas. The professions include postdocs and PhD students working at universities. All participants hold a master’s degree in their profession and several years of experience in their domain. All experts are between 25 and 40 years old (two female, two male).

#### 3.3. Scientific Workflows

In our study, we consider a total of five different workflows. We summarize the used workflows and their details in Table 1. The workflows are taken from the work context of the recruited experts, given their high degree of familiarity and expertise with them. In bioinformatics, we use two workflows that deal with the analysis of genomic data. First, the *crisprseq* workflow, sourced from the *nf-core* repository<sup>5</sup>, a hub for best-practice workflows, focuses on analyzing and evaluating gene editing experiments utilizing CRISPR-Cas9 mechanism for genome engineering. Second, the *RS-STAR* workflow, which was implemented by the recruited domain expert and performs differential gene expression analysis using RNA-seq data. The *FORCE2NXF-Rangeland* and *FORCE* are two implementations of an Earth

<sup>3</sup><https://chat.openai.com> - last access 20-10-2023

<sup>4</sup>Moreover, we provide the chat records in the Supplementary Material of this article.

<sup>5</sup><https://nf-co.re> - last access 20-10-2023

Table 1: Overview of the used workflows. We examine workflows from two scientific domains, i.e., bioinformatics and Earth observation, and two workflow systems (Nextflow and Apache Airflow). For each workflow, we report the number of (high-level) steps and used tools and in which study the workflow is used.

| Domain            | Workflow                                       | Description  | #Steps | #Tools | Study    |
|-------------------|--|--|--------|--------|----------|
| Bio-informatics   | WF1:<br>crisprseq [69]<br>(Nextflow)           | A bioinformatics data pipeline for analyzing and evaluating gene editing experiments utilizing CRISPR-Cas9 mechanism for genome engineering.<br>Repository: <a href="https://nf-co.re/crisprseq/2.0.0">https://nf-co.re/crisprseq/2.0.0</a> (created 07/2022)                          | 6      | 9      | I        |
|                   | WF2:<br>RS-Star<br>(Nextflow)                  | The general aim of this workflow is to perform differential gene expression analysis using RNA-seq data.<br>Repository: <a href="https://github.com/Nine-s/nextflow_RS1_star">https://github.com/Nine-s/nextflow_RS1_star</a> (created 11/2021)  | 5      | 5      | I,II,III |
| Earth Observation | WF3:<br>FORCE2NXF-Rangeland [33]<br>(Nextflow) | This workflow analyzes long-term vegetation dynamics in the Mediterranean using the FORCE-toolkit.<br>Repository: <a href="https://github.com/CRC-FONDA/FORCE2NXF-Rangeland">https://github.com/CRC-FONDA/FORCE2NXF-Rangeland</a> (created 11/2020)                                    | 9      | 8      | I        |
|                   | WF4:<br>Grasslands<br>(Nextflow)               | This workflow aims at understanding differences in long-term changes (1984-2022) in ground cover fractions specific to European grasslands.<br>Repository: <a href="https://github.com/kelewinska/FONDA_trends-nf">https://github.com/kelewinska/FONDA_trends-nf</a> (created 08/2023) | 6      | 3      | I,III    |
|                   | WF5:<br>FORCE<br>(Apache Airflow)              | This workflow focuses on analyzing long-term vegetation dynamics in the Mediterranean using the FORCE-toolkit.<br>Repository: <a href="https://github.com/CRC-FONDA/fonda-airflow-dags/tree/main">https://github.com/CRC-FONDA/fonda-airflow-dags/tree/main</a> (created 02/2021)      | 8      | 8      | I        |

observation workflow that is concerned with analyzing long-term vegetation dynamics in the Mediterranean using the FORCE toolkit<sup>6</sup>, which provides processing routines for satellite image archives. The former is implemented using the Nextflow scientific workflow management system<sup>7</sup> [39] and the latter by leveraging Apache Airflow<sup>8</sup> [37]. The third Earth observation workflow called *Grasslands*, builds on previous work [71, 72] aiming at understanding differences in long-term changes in ground cover fractions specific to European grasslands depending on the definition of endmembers (i.e., unique spectral signatures of a specific material or ground cover) approximating these fractions.

### 3.4. LLM Prompting

The choice and design of prompts entered into a LLM has a decisive influence on the output quality [73] and, in our case, on the suitability of ChatGPT for workflow development and implementation. Our prompts are organized first to provide the context, often including the workflow script, followed by the specific question or instruction under investigation. Suppose the workflow is divided into sub-workflows, possibly distributed over several files. In that case, we first specify the main

workflow and then all sub-workflows and task definitions in the order they occur in the main workflow. In our research’s initial stages, we experimented with various alternative prompts for each user study, incrementally modifying and enhancing them in response to the outcomes we received. For example, for Study I (i.e., workflow comprehension), we discovered that ChatGPT tends to describe properties of the workflow language or technical aspects instead of workflow characteristics. Such phenomena could be resolved by adding explicit instructions, e.g., “do not explain nextflow concepts”. We refer to Section 7.5 for a detailed discussion of prompt design challenges. We stopped this adjustment process after a few iterations as soon as no more of such artificial artifacts were generated. We have refrained from more extensive prompt engineering since workflow designers are domain experts from diverse fields who cannot be assumed to be specialists in developing and tuning prompts. However, we acknowledge that the choice of wording in our prompts influences the results [73]. We discuss the limitations of our work concerning the chosen study design and prompt strategy selection in more detail in Section 7.6.

## 4. Study I: Workflow Comprehension

In our first study, we investigate the capabilities of ChatGPT in capturing the actual purpose of a workflow. In other words, we are prompting ChatGPT to explain the purpose of a work-

<sup>6</sup><https://github.com/davidfrantz/force> - last access 20-10-2023

<sup>7</sup><https://www.nextflow.io> - last access 20-10-2023

<sup>8</sup><https://airflow.apache.org> - last access 20-10-2023

Table 2: Overview of the used prompts to investigate ChatGPT’s capabilities in capturing the content of a workflow description (Study I). *[workflow-language]* and *[workflow-text]* represent placeholders for the workflow management system, i.e., Nextflow or Apache Airflow, and the workflow description text. All prompts are executed within one conversation.

| ID   | Category             | Prompt   |
|------|----------------------|--|
| P1_1 | Overall aim          | The following text contains a scientific workflow written in <i>[workflow-language]</i> :<br><i>[workflow-text]</i><br>Explain from which research area this workflow originates and describe the general aim of this workflow. Don’t explain <i>[workflow-language]</i> concepts. |
| P1_2 | Workflow explanation | Explain all individual tasks that are implemented in this workflow. For each task explain which software programs or tools are used in this workflow to perform the task. Don’t explain <i>[workflow-language]</i> concepts.   |
| P1_3 | Workflow explanation | Explain the type of input data and the format of the input data needed for this workflow. Don’t explain the workflow itself.   |
| P1_4 | Workflow explanation | Explain the overall result of this workflow. For each individual task of the workflow report the type of data that is produced by this task.   |
| P1_5 | Research questions   | Explain up to 3 research questions for which this workflow is helpful.   |

flow. In this study, we assess ChatGPT’s quality in comprehending and explaining a workflow’s purpose in a user study involving workflow experts. Understanding the data flow and the analysis performed constitutes an important aspect of the daily work with scientific workflows. On the one hand, workflows are often precisely adapted to individual research questions, which makes it challenging even for other experts from the same domain to understand them. On the other hand, in many research institutions, an increasing number of legacy workflows whose original authors and contributors are no longer available for maintaining and refining the codes require taking over by new team members. Understanding a workflow is usually a prerequisite for adopting and applying a workflow correctly.

Thus, Study I pursues three goals: how well does ChatGPT perform on (a) identifying the domain and the overall objective of the analysis, (b) reporting the individual computation steps, used tools, their needed input data, produced output data, and (c) explaining research questions for which these analyses are helpful given the workflow description. The first two parts of the study have a reconstructive character, whereas the third is more explorative, requiring ChatGPT to reason beyond the given workflow description. We build a set of five different prompts to evaluate ChatGPT’s capabilities concerning the three dimensions. When providing the workflow definition in the prompt, delete all comments within the definition to prevent information leakage. Table 2 depicts the developed prompts.

For each workflow, all prompts are executed in one conversation, enabling ChatGPT to use the in- and output of previous prompts as context information. We ask the domain scientists to evaluate answers given by ChatGPT using a feedback questionnaire. The complete questionnaire contains nine items in total and can be found in full-length in Appendix A. The questionnaire focuses on the correctness of the prompts regarding the aim of the workflow, the explanation, and the forecast of

addressed research questions. For four of the nine items, users rate the generated explanations on a 5-point Likert scale<sup>9</sup>. In addition, three items comprise quantitative evaluations of how many computational steps are correctly detected, how many utilized software tools and programs are accurately identified, and how many valid follow-up research questions. The remaining question items concern the quality of the explanations of the workflow sequence, the description of the tools used, and the results produced. We add a comment field for each item to report issues and errors in the generated explanations if the domain expert does not apply the content.

#### 4.1. Results

The results of the expert surveys are presented in the following according to the three subcategories of the questionnaire, i.e., research area and the overall aim of the workflow, explanation of workflow details, and subsequent research questions.

##### 4.1.1. Overall Aim of the Workflow

The first two rows of Figure 2a highlight the rating distribution of the domain experts concerning ChatGPT’s explanations for the research area and the overall aim of the workflows. The experts agreed with the explanations generated, indicating a basic understanding of ChatGPT regarding the different workflows ( $\mu = 4.7$ ,  $\sigma = 0.68$ ). The lowest agreement, with a score of 4 concerning the research area and 3 for the overall aim, was recorded in the evaluation of the *WF4-Grasslands* workflow. In this case, the expert could not agree with the explanations mainly due to the wrong interpretation of an abbreviation within the workflow description, i.e., *FNF* was misinterpreted as “fraction of non-forest”, instead of *fold and fill*. This misunderstanding resulted in the workflow being explained as examining forest regions rather than grasslands.

<sup>9</sup>1: Strongly Disagree; 5 Strongly Agree.

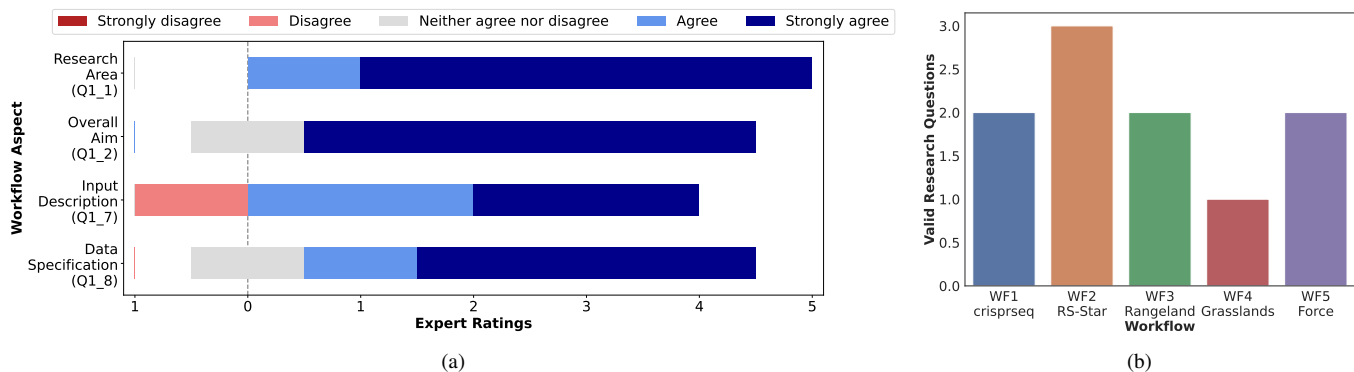


Figure 2: Results from Study I: (a) rating distribution of the domain experts for ChatGPT’s capability in identifying the research area and overall aim as well as the input and data description of a workflow. The question item identifier (see Appendix A) is given in parenthesis for each row. (b) Number of valid research questions generated by ChatGPT for the different workflows as assessed by the domain experts (question item *Q1\_9*). We prompted ChatGPT to output up to three research questions per workflow.

#### 4.1.2. Workflow Explanation

In general, the participants regard the quality of the explanations given by ChatGPT as high (see Figures 3a and 3b). All computational steps are accurately identified in three of the five workflow descriptions. Moreover, for four of these five workflows, every tool employed is correctly detected, but in *WF5-Force*, two out of eight were missing. The detailed descriptions of the tasks and tools provided by ChatGPT were also judged to be coherent by the experts. Overall, the worst performance is achieved with the output for workflow *WF4-Grasslands* for which only four out of six tasks are correctly extracted and only most of the tools are correctly described. These errors are mainly due to follow-up errors that result from the incorrect recognition of the workflow purpose.

The results of the questions items (see *Q1\_7* and *Q1\_8* in Appendix A), which assess the produced information about the format and type of input and output data of ChatGPT, can be seen in the two lower rows in Figure 2a. Similar to the previous findings, the information generated for four of the five workflows are evaluated positively ( $\mu = 4.2$ ,  $\sigma = 1.0$ ). Again, only the produced information for the workflow *WF4* was assessed as neutral or negative, i.e., input description score of 2 (disagree) and data specification of 3 (neither agree nor disagree).

#### 4.1.3. Research Questions

The last query was concerned with explaining up to three subsequent research questions to a given workflow. Figure 2b shows the result of this question item. ChatGPT achieves only moderate performance, generating just for one workflow, i.e., *WF2-RS-Star*, three valid research questions. In total, out of the 15 generated research questions 10 were correct. These figures suggest that ChatGPT offers only a reduced performance for more explorative tasks.

## 5. Study II: Workflow Modification

The second study investigates how much ChatGPT can aid domain experts in modifying and tailoring a workflow in our

second study. Researchers usually do not start from scratch when developing workflows but typically adapt or reuse parts of existing workflows from the community [6]. This strategy applies in particular to biomedicine, in which workflows are more widespread and have a longer tradition compared to other domains [74]. For example, genomic workflows will often be applied to a broad spectrum of data originating from different sources, each with its distinctive features and characteristics, making it necessary to adjust the workflow definition for more efficient data processing. Moreover, technological advancements, such as in genome sequencing technology [75], lead to new tools specially developed to leverage the capabilities of the new technologies. The continuous integration of new and alternative scientific tools into existing workflows is essential to conduct state-of-the-art research [76].

In our study, we are particularly investigating the exchange of used tools in the bioinformatics workflow *WF2-RS-Star* whose computational scheme is given in Figure 1a. We select two parts of the workflow to be modified:

1. read trimming and filtering (also called read quality control), originally performed by FASTP [77] and
2. reference genome indexing and alignment, carried out by STAR [78].

For assessing the workflow modification capabilities of ChatGPT, we build four prompts (see Table 3). The first prompt requests a list of alternative tools for a given workflow step from ChatGPT. The second and third prompts request the recommendation of two alternative tools, including an explanation of the suggestion, a comparison of the selected tools with the tool originally used in the workflow script, and their strengths and weaknesses. With the last prompt, the actual rewriting of the workflow to include the selected tool is requested. We test the inclusion of two alternative tools per computational task, i.e., the prompts *P2\_3* and *P2\_4* (see Table 3) are carried out once for each tool from the recommendation. Analogous to Study I, we use a questionnaire for evaluating ChatGPT’s output by the biomedical domain expert containing 13 items in total. For most items (9 out of 13), the generated texts and explanations

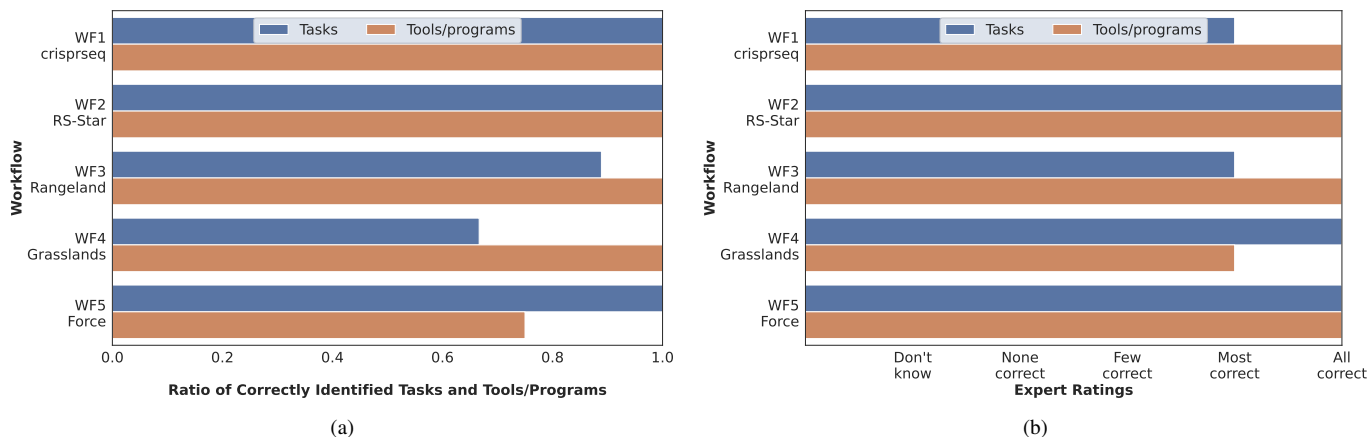


Figure 3: Result statistics highlighting the number of correctly identified tasks and tools of a workflow (a) and their explanation (b). We report separate results per investigated workflow. The results are based on the ratings of the domain experts. Results in (a) correspond to the items *Q1\_3* and *Q1\_5* and in (b) to *Q1\_4* and *Q1\_6* of the questionnaire given in Appendix A.

concerning methodical differences or pros and cons of the tools should be rated on a 5-point Likert scale. The remaining questions require numerical ratings (2 items), yes/no answers (1), and free text fields (1). Again, we add a comment field for each item to report issues and errors in the generated explanations if the domain expert does not fully apply the content. The complete questionnaire is given in Appendix B. When conducting the study, we also provide the generated workflow scripts to the domain experts and asked them to execute them on their systems. Furthermore, we request the experts to inspect and correct any non-functional scripts. For the latter, we set a time limit of 20 minutes per tool substitution.

## 5.1. Results

The results are presented in the following according to the two subcategories of the prompts, including the exploration of alternative tools and workflow modification. We summarize the results of the two use cases (i.e., read quality control, reference genome alignment) and each of the two alternative tools when reporting the results.

### 5.1.1. Tool Exploration

When exploring possible alternative tools, ChatGPT showed a good performance, providing a fully valid list of ten alternative tools using prompt *P2\_1* (see Table 3) for both scenarios. The generated output list for both tasks can be seen in Figure 4. However, the domain expert criticized the alternative tools for reference genome alignment specified Bowtie [79] and not its successor Bowtie2 [80]. From these, ChatGPT recommended HISAT2 [81] and Bowtie as alternatives for genome alignment and Trimmomatic [82] and Cutadapt [83] for read quality control. According to the domain expert, all tools are principally valid substitutions, but Bowtie is a rather inappropriate suggestion since it is not specialized for RNAseq data. The results of explaining the tools' methodical differences and strengths and weaknesses can be seen in Figure 5. Essentially, the generated textual explanations were assessed positively, except when

describing the methodological differences between the tools. In this case, ChatGPT could offer a convincing explanation in only 50% of the cases. In the other cases, the texts were too general, and differences were not named clearly. In summary, similar to the results from Study I, ChatGPT solves these rather reconstructive tasks well, reaching an average score of 4.1 (min=3, max=5,  $\sigma=0.81$ ) when taking the results from items *Q2\_3* to *Q2\_7* (see questionnaire in Appendix B) into account. These results indicate that ChatGPT could effectively used to explore tools in a given field.

### 5.1.2. Workflow Modification

We requested ChatGPT to re-write the workflow script for each of the two recommended alternative tools in both use cases, resulting in four modified workflows in total. Table 5 summarizes the results achieved. First of all, none of the generated workflow scripts was entirely correct. Only in one of the four cases, the generated script was (at least) syntactically valid, i.e., if replacing FASTP with Trimmomatic for read quality control, and the domain expert could execute it without further adaptations. However, the script was not a semantically correct modification of the original workflow script since one particular quality control routine (i.e., PolyG trimming) was not reflected in the adopted script. This issue also occurred for Cutadapt, the other alternative tool for quality control. For both tools, ChatGPT reported in the explanation text that the tools do not support this feature; however, in reality, they do. This failure could be interpreted as a kind of LLM hallucination [53]. The second task, reference genome indexing and alignment, revealed different issues than the first. Here, the main problem was the correct linking of the two sub-parts of the task, first the index generation and then the computation of the actual alignment. For the former, each tool specifies and uses its distinct data format and defines how to store the index (e.g., saving it in one or multiple files). However, the storing strategy also affects how the output of the indexing task has to be passed on to the input of the alignment computation. In the scripts generated by ChatGPT,



### Alternative Tools for Reference Genome Alignment

#### ChatGPT Output:

Here is a list of 10 alternative tools that can be used for reference genome alignment:

1. **Bowtie**: A fast and memory-efficient aligner for short DNA sequences.
2. **BWA** (Burrows-Wheeler Aligner): A popular aligner for mapping DNA sequences against a large reference genome.
3. **HISAT2**: A fast and sensitive aligner for spliced alignment of RNA-seq reads.
4. **TopHat**: A widely used aligner for RNA-seq reads that incorporates splice junctions.
5. **Subread**: A versatile aligner capable of aligning both DNA and RNA sequences to a reference genome.
6. **Salmon**: A lightweight aligner specifically designed for accurate quantification of transcript abundance from RNA-seq data.
7. **STAR** (Spliced Transcripts Alignment to a Reference): The tool used in the provided workflow, known for its speed and ability to handle splice junctions.
8. **GSNAP**: A general-purpose aligner that allows for highly sensitive and accurate alignment of short reads against a reference genome.
9. **Novoalign**: A versatile aligner that supports both DNA and RNA sequence alignment, offering high sensitivity and accuracy.
10. **Segemehl**: A read mapper optimized for mapping short reads against large genomes with a focus on finding splice junctions accurately.

These are just a few examples of the many alignment tools available in the field of bioinformatics. The choice of tool depends on specific requirements, such as the nature of the data (DNA or RNA), read length, presence of splice junctions, and desired performance characteristics.

### Alternative Tools for Read Filtering and Trimming

#### ChatGPT Output:

Certainly! Here's a list of 10 alternative tools that can be used for read filtering and trimming in biomedical workflows:

1. **Trimomatic**: A popular tool for read trimming, adapter removal, and quality filtering.
2. **Cutadapt**: Used for adapter trimming in high-throughput sequencing data.
3. **BBduk**: Part of the BBDuk package, it offers read trimming, filtering, and quality control.
4. **SeqKit**: A fast and lightweight tool for sequence processing, including read trimming.
5. **Trimalore**: Specifically designed for trimming and quality filtering of Illumina sequencing data.
6. **Sickle**: A tool that focuses on quality-based trimming of reads from high-throughput sequencing data.
7. **AlienTrimmer**: A read trimming tool that uses a statistical approach to identify and remove adapters.
8. **Prinseq**: Provides various options for read filtering, trimming, and quality control.
9. **SolexaQA**: Offers quality assessment, read trimming, and error correction for Illumina sequencing data.
10. **AdapterRemoval**: Specialized for adapter trimming and quality filtering in paired-end sequencing data.

These tools provide different features and algorithms for read filtering and trimming, allowing researchers to choose the one that best suits their specific needs and preferences.

Figure 4: Representation of the output of ChatGPT when requested to provide a list of alternative tools for reference genome indexing and alignment (left) and read quality control (right) using prompt P2.1. All tools are assessed to be valid by the biomedical domain expert.

the actual step descriptions to invoke indexing and alignment by the tools were (generally) valid. However, the linking of these two needed to be corrected. For example, Bowtie saves its index in multiple files sharing a common file name prefix, which has to be specified as a parameter during alignment. However, in the modified script, the list of all files of a specially created directory was passed to the alignment process. For Bowtie, this problem could be easily fixed by the domain expert, but for HISAT2, it was not that trivial and hence could not be solved in the given time budget of 20 minutes.

To sum up, the study's results indicate that modifying workflow scripts poses considerable challenges for ChatGPT as it requires a detailed understanding of the tool's idiosyncrasy, the exact computations they perform, and the data formats they use.

## 6. Study III: Workflow Extension

In the third study, we investigate the capabilities of ChatGPT in extending a scientific workflow given a partial script. As discussed in the motivation for Study II (see Section 5), users often reuse parts of existing workflows from the research community and adapt them to the research question at hand by enhancing the pipeline with additional analyses and computational steps [6]. Moreover, data analysis projects are often exploratory processes, and computation pipelines are incrementally adapted and extended based on executions and findings from previous versions of the workflow, e.g., to include additional data correctness checks, add more differentiated result evaluations and provide advanced result visualizations [84]. In

our study, we simulate this incremental exploration process by taking an existing workflow and removing  $n$  steps at the end of it. We then request ChatGPT to a) enumerate the necessary steps to accomplish the original research goal and b) regenerate the next step using the tool of the original pipeline or by giving a verbal description of the task. For this study, we select one workflow from each research domain for investigation: *WF2-RS-Star* for biomedicine and *WF4-Grasslands* for earth observation. The two workflows were chosen because they offer different implementation characteristics, i.e., *WF2-RS-Star* leverages almost exclusively external tools, whereas *Grasslands* relies more strongly on specially implemented R and Python scripts. Moreover, Study I (see Section 4) already showed notable result differences of ChatGPT for both workflows. By choosing these two specific workflows, we aim to encompass a possibly broad spectrum of performance variations. We test ChatGPT's workflow extension capabilities in three scenarios: For *WF2-RS-Star*, we remove the last step *transcript quantification* as well as the last two steps *transcript quantification* and *format conversion*, forming two extension scenarios. In the case of *WF4-Grasslands*, we remove all steps at the tail of the workflow, including *autoregressive trend analysis* (see schema in Appendix D).

Table 4 illustrates the prompts developed for this purpose. This study uses slightly different prompts (see P3.2a and P3.2b) reflecting the different workflow types, i.e., tool- vs. script-based. For the latter, we include additional instructions to a) specify the programming language of the script and b) ask the domain expert for a verbal description of the computational steps to be implemented. See Appendix E for the verbal de-

Table 3: Overview of the used prompts to investigate ChatGPT’s capabilities in swapping used tools in bioinformatic workflows (Study II). Information in square brackets specifies placeholders for concrete information regarding the workflow or the tool to be replaced.

| ID   | Category              | Prompt   |
|------|-----------------------|--|
| P2.1 | Tool exploration      | The following text contains a <i>[domain]</i> workflow written in <i>[workflow-language]</i> :<br><i>[main-workflow]</i><br>The following snippets contain the source code for the step of the workflow which uses <i>[tool]</i> to perform <i>[step]</i> : <i>[step-source-code]</i> . Please provide a list of 10 alternative tools to perform <i>[tool]</i> .   |
| P2.2 | Tool exploration      | The following text contains a <i>[domain]</i> workflow written in <i>[workflow-language]</i> :<br><i>[main-workflow]</i><br>The following snippets contain the source code for the step of the workflow which uses <i>[tool]</i> to perform <i>[step]</i> : <i>[step-source-code]</i> . Alternative tools for <i>[step]</i> are: <i>[list-of-tools]</i><br>Which of the tools would you recommend as most suitable alternative for <i>[step]</i> in the given workflow. Please name the two alternatives and give an explanation why these tools are especially advisable for the given workflow.                              |
| P2.3 | Tool exploration      | <i>[original-tool]</i> and <i>[alternative-tool]</i> are two tools for <i>[step]</i> in <i>[domain]</i> workflows. First, explain the differences between the tools and the underlying approaches. Second, name strengths and weaknesses of both tools.  |
| P2.4 | Workflow modification | The following text contains a <i>[domain]</i> workflow written in <i>[workflow-language]</i> :<br><i>[main-workflow]</i><br>The following snippets contain the source code for the step of the workflow which uses the <i>[tool]</i> to perform <i>[step]</i> : <i>[step-source-code]</i><br>Please re-write the code of the workflow and the process to use <i>[alternative-tool]</i> instead of <i>[original-tool]</i> . The number of parameters of the individual process descriptions may have to be adjusted. Please explain features / options of <i>[original-tool]</i> which are not supported in <i>[new-tool]</i> . |

scription provided by the earth observation expert. The questionnaire for evaluating the generated outputs consisting of seven items can be found in Appendix C.

## 6.1. Results

The results are presented in the following according to the two subcategories of the prompts, i.e., workflow exploration and extension.

### 6.1.1. Workflow exploration

For describing further computational steps necessary to accomplish a specific research goal given a partial workflow, ChatGPT showed mixed results. The LLM provides a correct list of suitable steps in two of the three scenarios. Also, the tools and methods for implementing the steps suggested by ChatGPT were valid. However, both domain experts criticize that the specifications for the necessary steps and the proposed tools tend to be rather generic and generalized. For instance, for extending the *WF4-Grasslands* workflow the earth observation expert commented:

*Overall, the proposed workflow is very generic and does not provide a clear roadmap for the analyses. It also proposes to use very simplistic and often imperfect approaches.*

Overall, the results confirm the findings from the two previous studies that ChatGPT shows weaknesses in more exploratory tasks.

### 6.1.2. Workflow extension

Using the prompts P3.2a and P3.2b (see Table 4), we request ChatGPT to re-construct the last removed computational step in each extension scenario. Table 6 summarizes the results achieved. Like the results from Study II, ChatGPT shows considerable weaknesses in the automatic extension of workflows. None of the generated workflow scripts was executable without the intervention of the domain expert. A clear difference is revealed when comparing the two domains, biomedicine and earth observation. In the former case, the generated workflow scripts are (at least) of such a quality that the domain expert could successfully correct them within 20 minutes. In the generated scripts, mainly syntactical errors occurred (e.g., incorrect usage of variable identifiers, incomplete input definitions, or missing specification of parameters), which could be easily corrected. However, the calls to the respective programs to perform the two tasks were correct.

In contrast, the generated extension for *WF4-Grasslands* was of considerably lower quality. In this case, several syntactic and semantic errors occurred, e.g., the script uses a non-existing library function, no parallelization code is included, and not all requested computations are performed. In this state, the domain expert could not resolve the large number of problems within 20 minutes. However, when interpreting these results, one must remember that the task in this scenario is also significantly more difficult. Instead of a short task description and specification of a tool to be used, ChatGPT has to design and generate the

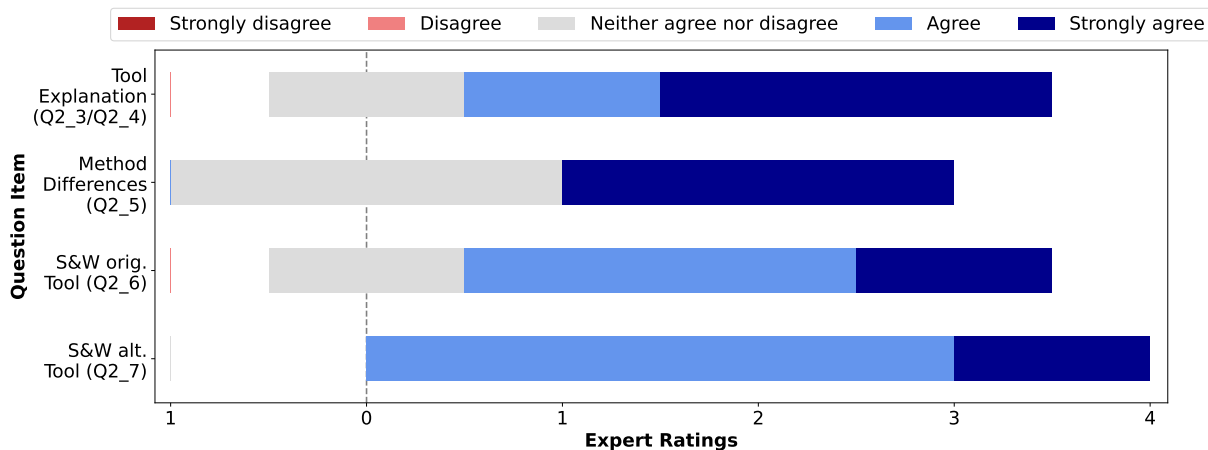


Figure 5: Overview of the rating distribution of the biomedical domain expert for ChatGPT’s capability for explaining alternative tools, methodical differences, and strengths and weaknesses (S&W) of the tools. The question item identifier (see Appendix B) is given in parenthesis for each row.

source code for a complex data analysis procedure containing multiple sub-steps.

## 7. Discussion

We conducted three studies to investigate the capabilities of using ChatGPT for comprehending, modifying, and extending scientific workflows. We discuss our methodology and the results in the following.

### 7.1. Comprehending Scientific Workflows

Study I was designed to answer **RQ1** by evaluating ChatGPT’s performance in comprehending existing workflows. The domain experts assessed that ChatGPT is good at this task while showing slight differences between the investigated research domains. In particular, the explanations for workflow *WF4-Grasslands* revealed considerable performance drops. Unlike the other workflows investigated, this one uses multiple proprietary R and Python scripts instead of leveraging external tools for assembling data processing pipelines. The lack of standardized tools makes workflow comprehension more challenging since ChatGPT has to interpret complex processing logic and has fewer possibilities to leverage static information, like the description of the general purpose of an established bioinformatics tool, seen through its training while generating the response. In addition, code quality and its readability may strongly influence the results for workflows containing proprietary scripts. For instance, one major problem while explaining *WF4-Grasslands* in Study I was the misinterpretation of the abbreviation *”fnf”* as *”fraction of non-forest”*, instead of *”fold and fill”*. Such customized and ambiguous terms challenge LLMs and reduce their applicability.

### 7.2. Modifying Scientific Workflows

We answer **RQ2** in Study II by evaluating the modification performance of ChatGPT. To this end, we requested the LLM to substitute the leveraged tools for two computational tasks, read

quality control and reference genome alignment, in the biomedical workflow *WF2-RS-Star*. The study results suggest that ChatGPT can effectively explore and explain alternative tools in the field, possibly shortening the time the experts spend searching for suitable replacements on the web. In contrast, the results also indicate that ChatGPT rather poorly supports the generation of workflow scripts for using these alternative tools. In only one scenario, i.e., substituting *FASTP* with *Trimmomatic*, the produced script could be run without syntactical errors, and in one other scenario, i.e., replacing *STAR* with *Bowtie*, the script could be fixed within 20 minutes to be syntactically and semantically valid. In the used version of ChatGPT and the selected setup, an increase in efficiency cannot be recorded or anticipated, highlighting the need for further research efforts. However, when interpreting the results, it is essential to remember that ChatGPT is a general-purpose LLM rather focusing on human language. A potential option for improvement could be testing generative models more strongly adapted to programming code, such as GitHub Copilot or Code Llama [19].

### 7.3. Extending Scientific Workflows

Finally, we investigate **RQ3** by conducting Study III. To this end, we requested ChatGPT to extend an existing (partially given) workflow to achieve specific goals. The study results confirm the findings from the two previous studies and emphasize ChatGPT’s difficulties in solving more complex and exploratory problems. In this case, explaining the necessary steps to answer the given research questions and the generation of the workflow script for the next step offered (partly) severe issues. Similar to the results of Study I, the picture is mixed regarding the different research domains, earth observation and bioinformatics. For the latter, the generated scripts form a relatively good basis for the implementation, having only (minor) syntactical issues that the expert could quickly fix. In contrast, in the case of earth observation, the script quality was considerably worse, hindering a fast correction by the expert. These results imply that efficient user support is possible for pipelines mainly leveraging external tools. However, further research is

Table 4: Overview of the used prompts to investigate ChatGPT’s capabilities in extending a given partial workflow (Study III). We distinguish two types of prompts: workflow exploration and workflow extension. For the latter, we developed two variants specially designed for tool- (P3\_2a) and script-based workflows (P3\_2b).

| ID    | Category             | Prompt   |
|-------|----------------------|--|
| P3_1  | Workflow Exploration | The following text contains a <i>[domain]</i> workflow written in Nextflow:<br><i>[workflow-description]</i><br>The workflow should be used to <i>[overall-goal]</i> . Which steps are missing in order to perform <i>[overall-goal]</i> ? Please specify only the absolutely necessary steps. For each step name up to three <i>[domain]</i> tools that can be used to perform the task.  |
| P3_2a | Workflow Extension   | The following text contains a <i>[domain]</i> workflow written in Nextflow:<br><i>[workflow-description]</i><br>Please extend to the given workflow to include one further step which <i>[step-description]</i> using <i>[tool]</i> . Please specify the new process description in a file at <i>[file-name]</i> . Please use version 2 of the Nextflow workflow language. The new process should take the output of <i>[predecessor-step]</i> as input.   |
| P3_2b | Workflow Extension   | The following text contains a <i>[domain]</i> workflow written in Nextflow:<br><i>[workflow-description]</i><br>Please extend to the given workflow to include one further task which performs <i>[step]</i> using an <i>[programming-language]</i> script. For this, please generate an <i>[programming-language]</i> script, stored in <i>[script-file-name]</i> , which performs the following computations:<br><i>[verbal-task-description]</i><br>Next to the <i>[programming-language]</i> script generate the Nextflow process description in a file named <i>[process-file-name]</i> and the updated workflow. Please use version 2 of the Nextflow workflow language. The new process should take the output of <i>[predecessor-step]</i> as input. |

necessary to investigate user-support strategies for workflows applying specially implemented analysis scripts.

#### 7.4. Scientific Workflows in LLM Training Data

LLMs are trained on large amounts of textual data from the web, including programming code and workflow scripts [42]. Therefore, it is crucial to consider whether and to what extent an LLM was already able to access the workflow scripts of our study during its pretraining. According to public information<sup>10</sup>, ChatGPT was trained on data gathered until September 2021, meaning that initial versions of two of the five tested workflows (i.e., *WF3-FORCE2NXF-Rangeland* and *WF5-Force*) could have been part of the LLM’s training routine. However, the specific training dataset used for ChatGPT is not accessible to the public, preventing a conclusive assessment. To attain a more precise estimate of the potential number of workflow scripts within the training data in general, we initiated searches for scientific workflow repositories on GitHub. We leverage the repository search engine of the website<sup>11</sup> and use the names of four widely-used workflow management systems, i.e., Apache Airflow, Nextflow, Snakemake, and Taverna as a query term. We filter all repositories with creation data less than 2021-09-01 from the query results. Of course, the results must be interpreted carefully since not every repository containing the name has to deal with scientific workflows, even if the names are very

peculiar. Detailed statistics from our search results are in Appendix F. As of 09/2021, there were between 352 and 1,900 repositories containing one of the workflow system names in their description. Moreover, the results highlight the increasing popularity of workflows since, for all systems except for Taverna, the number of repositories has almost doubled over the last two years. We also checked the number of Nextflow pipelines available in nf-core. As of September 2021, 35 pipelines were published, and 19 were under development<sup>12</sup>. Today, nf-core hosts 55 published pipelines and 33 in development. In summary, we can hypothesize from these results that ChatGPT can likely rely only on a relatively small base of workflow scripts during its training compared to classical programming code (e.g., GitHub currently hosts over 3.9 million Java and over 2.2 million Python repositories<sup>13</sup>) making user support for workflow design and implementation particularly challenging.

#### 7.5. Prompt Design Challenges

While creating prompts for the studies, we identified several challenges and issues that arose while interacting with ChatGPT.

##### 7.5.1. Representation of Workflows

For the representation of the workflow scripts, there is no straightforward option on how to include them in a prompt. The

<sup>10</sup><https://platform.openai.com/docs/models/gpt-3-5> - last access 20-10-2023

<sup>11</sup><https://github.com/search> - last access 20-10-2023

<sup>12</sup><https://nf-co.re/stats> - last access 20-10-2023

<sup>13</sup>Determined by using GitHub repository search and the search queries “*language:Java*” and “*language:Python*”. – last access 20-10-2023

Table 5: Overview of the results of the workflow modification use case in which the tools performing a specific task are replaced by alternative ones. For each task, we provide the original tool (in parenthesis) and the investigated alternative ones suggested by ChatGPT. For each combination, we highlight (✓=yes,×=no) whether the generated workflow script could be executed (Exec.), whether it is semantically valid (Val.), and whether it could be fixed within 20 minutes (Fix). For the latter, (✓) indicates cases where the script could be fixed to be executable but not entirely semantically correct. Moreover, we provide excerpts from the domain expert’s comments.

| Task   | Alt. Tool        | Exec. | Val. | Fix | Problems / Comments  |
|--|------------------|-------|------|-----|--|
| Read quality control<br>( <i>FASTP</i> [77])       | Trimmomatic [82] | ✓     | ×    | (✓) | <ul style="list-style-type: none"> <li>• Missing implementation to perform special quality control feature (i.e. PolyG trimming)</li> <li>• Invalid description of workflow differences: <ul style="list-style-type: none"> <li>- ChatGPT stated that polyG trimming isn’t available in Trimmomatic, but it actually is, however extra implementation for data adaptation is needed</li> <li>- Differences of the generated output files not described accurately</li> </ul> </li> </ul> |
|  | Cutadapt [83]    | ×     | ×    | (✓) | <ul style="list-style-type: none"> <li>• Wrong program call: syntactically incorrect specification of two parameters</li> <li>• Missing provision of valid adapter sequences</li> <li>• Invalid description of differences of the workflow script: <ul style="list-style-type: none"> <li>- ChatGPT stated that polyG trimming isn’t available in Cutadapt, but it actually is</li> <li>- Differences of the generated output files not described accurately</li> </ul> </li> </ul>      |
| Genome indexing & alignment<br>( <i>STAR</i> [78]) | HISAT2 [81]      | ×     | ×    | ×   | <ul style="list-style-type: none"> <li>• Invalid definition and linkage of in- and output between genome index generation and alignment</li> <li>• Syntactically incorrect call of the alignment process</li> <li>• Does not take parameter <i>strandedness</i> of the input data into account</li> <li>• Output files aren’t generated correctly</li> </ul>   |
|  | Bowtie [79]      | ×     | ×    | ✓   | <ul style="list-style-type: none"> <li>• Invalid definition and linkage of in- and output between genome index generation and alignment</li> <li>• Wrong output definition of genome indexing task</li> </ul>  |

workflow descriptions are often spread over several files containing sub-workflows and task descriptions. In our approach, we first specify the main workflow and then all sub-workflows and task descriptions in order of occurrence. However, there might be other, more efficient prompt solutions (with respect to the generative language model). Furthermore, the workflow scripts might exceed the maximum allowed input length of the language model, e.g., ChatGPT variants allow only for 4K to 16K words / tokens<sup>14</sup> in the input sequence. In particular, workflows heavily relying on specially implemented scripts having hundreds of code lines will face this issue.

### 7.5.2. Loss of Focus

Some of the prompts are very long due to the specification of the entire workflow script, which challenges ChatGPT to maintain focus. Adding additional instructions to the prompt helped to avoid or reduce this phenomenon, e.g. for the explanation use case (Study I) we added to the prompt “*Don’t explain nextflow concepts*” (see P1\_1 and P1\_2 in Table 2) and “*Don’t explain the workflow itself*” (P1\_3) to prevent ChatGPT to generate outputs describing features of the workflow man-

agement system or the complete workflow when requesting input data specification.

### 7.5.3. Technological Details

In some cases, adaptation to technological details of the specified workflows were necessary. For example, the Nextflow system offers two language versions for describing processing pipelines. The Nextflow workflows in our study all used the new version of the language. However, when extending workflows in Study III, we had to specify the desired version (see P3\_2 in Table 4) to get the correct output. This observation is surprising since the partially given workflow is already in the respective version. Interestingly, this was only necessary for the workflow extension but not for their modification (P2\_4 in Table 3) in which the phenomena did not occur.

In summary, the efficient and effective formation of prompts offers a wide range of possible solutions. In our study, we identified initial clues and difficulties, but further research is needed to detect further potential for improving the interaction between domain experts and ChatGPT and generative LLMs in general.

## 7.6. Limitations and Future Work

In the following, we highlight the limitations of this work that merit further research.

<sup>14</sup><https://platform.openai.com/docs/models/gpt-3-5> - last access 20-10-2023

Table 6: Overview of the results of the workflow extension use case in which we provide ChatGPT a partial workflow and request the LLM to extend it by one further computational step. For each investigated use case, we highlight (✓=yes, ×=no) whether the generated workflow script could be executed (Exec.), whether it is semantically valid (Val.), and whether it could be fixed within 20 minutes (Fix). For the latter, (✓) indicates cases where the script could be fixed to be executable but not entirely semantically correct. Moreover, we provide excerpts from the domain expert’s comments.

| Workflow          | Task/Tool   | Exec. | Val. | Fix | Problems / Comments   |
|-------------------|---|-------|------|-----|---|
| WF2<br>RS-Star    | Transcript<br>quantification/<br><i>Cufflinks</i><br>[85] | ×     | ×    | ✓   | <ul style="list-style-type: none"> <li>• Syntax errors: process definition for CUFFLINKS declares one input channel but two were specified</li> <li>• Input tuple does not match input set cardinality declared by process definition</li> <li>• Wrong variable name: sorted.bam (wrong) instead of sample.bam</li> </ul>   |
|                   | Format<br>conversion/<br><i>SAMtools</i><br>[86]          | ×     | ×    | ✓   | <ul style="list-style-type: none"> <li>• Syntactical errors: Usage of wrong variable name (sample.sam)</li> <li>• Incorrect syntax for connecting the new task to the previous one SAMTOOLS(STAR_ALIGN.sample.sam) (wrong) vs. SAMTOOLS(STAR_ALIGN.out.sample.sam) (correct)</li> </ul>   |
| WF4<br>Grasslands | AR<br>analysis/<br><i>R script</i>                        | ×     | ×    | ×   | <ul style="list-style-type: none"> <li>• Input to the R script is a path to a directory, not a TIFF file</li> <li>• Incorrect use of remotePARTS library: There is no function called <i>autoTrend</i> in this package</li> <li>• Calculation needs to be parallelized (as specified in the request)</li> <li>• Computation should be implemented for four types of inputs, namely: GV, NPV, SOIL, and SHADE.</li> <li>• Desired outputs from the AR model needs to be retrieved and written out (missing)</li> <li>• Script declares a Conda environment (Python), not R environment.</li> </ul> |

### 7.6.1. Study Design

In each of our three studies, we created and provided the prompts for testing ChatGPT’s capabilities concerning the different use cases and the domain experts only evaluated the outputs of ChatGPT, leading to a rather indirect interaction between the domain scientist and the LLM. An alternative design for the study would be to have the experts interact directly with ChatGPT by developing and refining the prompts independently. In addition to assessing the capabilities of ChatGPT, this would have the advantage of gaining initial insights into interaction forms and patterns of the different experts with ChatGPT. Moreover, this would allow for improved customization of the prompts to the particular research domain and the idiosyncratic properties and characteristics of each workflow. Extended optimization of the prompting strategy by the domain scientist could lead to better results but reduce potential time savings in solving the actual task. Our study design was motivated by the fact that the experts had strongly limited time budgets for the study. For example, even for evaluating ChatGPT’s outputs in Study I, the experts already needed up to three hours to accurately check the generated explanations. A study design that envisages direct interaction involves high efforts in terms of introduction and explanation to ChatGPT and prompting strategies for the domain scientists, thus limiting the scope of research questions that can be investigated. In addition, the selected study design has the advantage of using the same prompts for the different domains, which contributes to better comparability of the results and eliminates the influence of differences for

individual prompt differences.

In our study, we focused solely on ChatGPT as generative language model. However, there are many other general-purpose models available (e.g., PaLM-2 [44], BARD<sup>15</sup> or Llama-2 [19]) as well as models more specially designed for programming tasks (e.g., GitHub Copilot, Code Llama [87] or OpenAI Codex<sup>16</sup>) publicly available and worth investigating. Our studies only highlight the results of ChatGPT in the version used (GPT-3.5) but do not claim generalizability for other LLMs. Finally, recent research showed that placebo effects can undermine the validity of study results when user expectations are altered through the presence of an AI [88, 89]. In future work and in the case of using LLMs, placebo conditions must be included to avoid findings that are not a result of increased user expectations towards the capabilities of ChatGPT.

### 7.6.2. Prompting Strategy

Next to other models, the prompts used in our studies also constitute a limiting factor. We cannot exclude the possibility that other prompts, using a different structure or wordings, may achieve better results for the investigated use cases. In addition, it should be emphasized that the generated texts are subject to stochastic processes, which can lead to deviations even when reusing the same prompts.

<sup>15</sup><https://bard.google.com/> - last access 20-10-2023

<sup>16</sup><https://openai.com/blog/openai-codex> - last access 20-10-2023

### 7.6.3. Limited Number of Domain Experts

In the context of our studies, only four domain experts evaluated the outputs of ChatGPT. In some cases, generated explanations were assessed by one person only (e.g., Study II). This low number of experts limits the validity and generalizability of the results and offers the risk of subjective bias. However, recruitment for such studies is difficult because the number of potential participants is small and they often have strongly limited time budgets, making study design challenging. Please note that for experts in the field, even “just” familiarizing themselves with an unfamiliar workflow is a challenging and time-consuming endeavor.

### 7.6.4. Investigated domains and selected workflows

Our study explores real-world workflows from the two domains, bioinformatics and earth observation. Of course, these only represent part of the full range of workflows in the natural sciences. It constitutes an exciting follow-up research question: how suitable ChatGPT and other generative LLMs are in other research contexts, such as climate research [32] and astronomy [31], and whether it is possible to identify categories or groups of domains which are particularly well (or poorly) supported. Furthermore, we only examined two workflow systems, Nextflow and Apache Airflow, leaving other alternatives, such as Snakemake, Taverna, and Pegasus, for future work.

### 7.6.5. Explored Use Cases

This work focused on comprehending, modifying, and extending workflows with ChatGPT. These use cases represent only a partial scope of user support opportunities and are worth considering and evaluating other use cases. For instance, migrating workflows implemented in legacy workflow management systems to more recent ones, e.g., transforming Taverna [74] scripts to Snakemake or Nextflow, or adapting them to different infrastructure stacks poses an interesting research question. Moreover, user support in workflow debugging, error identification, or optimization, as done in classical programming [61], would be a valuable contribution to research scientists.

## 8. Conclusion

The significance of large-scale data analysis workflows in advancing research in the natural sciences is growing steadily. Developers of such workflows, primarily researchers from diverse scientific fields, are challenged with the increasing complexity and scale of their analyses, which involve (next to their domain knowledge) working with different frameworks, tools, programming languages, and infrastructure stacks. Although a few tools for creating and maintaining workflows are available, improving user efficiency remains an open research area. In this work, we contribute to this situation by evaluating the suitability of ChatGPT for comprehending, modifying, and extending scientific workflows. In three user studies with four researchers from different scientific domains, we evaluated the correctness of ChatGPT regarding explainability, exchange of software components, and extension when providing real-world

scientific workflow descriptions. Our results show a high accuracy for comprehending and explaining scientific workflows while achieving a reduced performance for modifying and extending workflow descriptions. These findings clearly illustrate the need for further research in this area.

## Acknowledgments

This work is supported by *German Research Foundation (DFG)*, CRC 1404: “FONDA: Foundations of Workflows for Large-Scale Scientific Data Analysis” (Project-ID 414984028).

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Author contributions

**Mario Sänger:** Conceptualization, Methodology, Investigation, Visualization, Writing - Original Draft **Ninon De Mequenem:** Resources, Data Curation, Writing - Review & Editing **Katarzyna Ewa Lewińska:** Resources, Data Curation, Writing - Review & Editing **Vasilis Bountris:** Resources, Data Curation, Writing - Review & Editing **Fabian Lehmann:** Resources, Data Curation, Writing - Review & Editing **Ulf Leser:** Conceptualization, Writing - Review & Editing, Project administration, Funding acquisition **Thomas Kosch:** Conceptualization, Methodology, Writing - Original Draft , Writing - Review & Editing, Validation, Project administration

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly and ChatGPT for linguistic revision. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## References

- [1] S. B. Davidson, J. Freire, Provenance and scientific workflows: Challenges and opportunities, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, Association for Computing Machinery, New York, NY, USA, 2008, p. 1345–1350.
- [2] S. Cohen-Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsén, P. Larmande, Y. L. Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal, C. Blanchet, Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities, *Future Generation Computer Systems* 75 (2017) 284–298.
- [3] L. Wratten, A. Wilm, J. Göke, Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers, *Nature methods* 18 (10) (2021).

- [4] J. Ison, H. Ienasescu, P. Chmura, E. Rydzka, H. Ménager, M. Kalaš, V. Schwämmle, B. Grüning, N. Beard, R. Lopez, S. Duvaud, H. Stockinger, B. Persson, R. Vařeková, T. Raček, J. Vondrášek, H. Peterson, A. Salumets, I. Jonassen, R. Hooff, T. Nyrönen, A. Valencia, S. Capella, J. Gelpí, F. Zambelli, B. Savakis, B. Leskošek, K. Rapacki, C. Blanchet, R. Jimenez, A. Oliveira, G. Vriend, O. Collin, J. van Helden, P. Løngreen, S. Brunak, The bio.tools registry of software tools and data resources for the life sciences, *Genome Biology (Online Edition)* 20 (1) (2019).
- [5] R. Ferreira da Silva, H. Casanova, K. Chard, I. Altintas, R. M. Badia, B. Balis, T. Coleman, F. Coppens, F. Di Natale, B. Enders, T. Fahringer, R. Filgueira, G. Fursin, D. Garijo, C. Goble, D. Howell, S. Jha, D. S. Katz, D. Laney, U. Leser, M. Malawski, K. Mehta, L. Pottier, J. Ozik, J. L. Peterson, L. Ramakrishnan, S. Soiland-Reyes, D. Thain, M. Wolf, A community roadmap for scientific workflows research and development, in: *Workshop on Workflows in Support of Large-Scale Science (WORKS)*, 2021.
- [6] S. Cohen-Boulakia, U. Leser, Search, adapt, and reuse: the future of scientific workflows, *ACM SIGMOD Record* 40 (2) (2011) 6–16.
- [7] E. Deelman, T. Peterka, I. Altintas, C. D. Carothers, K. K. van Dam, K. Moreland, M. Parashar, L. Ramakrishnan, M. Taufer, J. Vetter, The future of scientific workflows, *The International Journal of High Performance Computing Applications* 32 (1) (2018) 159–175.
- [8] B. A. Myers, M. B. Rosson, Survey on user interface programming, in: *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1992, pp. 195–202.
- [9] B. R. Barricelli, F. Cassano, D. Fogli, A. Piccinno, End-user development, end-user programming and end-user software engineering: A systematic mapping study, *Journal of Systems and Software* 149 (2019) 101–137.
- [10] S. Lau, S. S. Srinivasa Ragavan, K. Milne, T. Barik, A. Sarkar, Tweakit: Supporting end-user programmers who transmogrify code, in: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–12.
- [11] N. Peitek, A. Bergum, M. Rekrut, J. Mucke, M. Nadig, C. Parnin, J. Siegmund, S. Apel, Correlates of programmer efficacy and their link to experience: A combined eeg and eye-tracking study, in: *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022*, Association for Computing Machinery, New York, NY, USA, 2022, p. 120–131.
- [12] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt, P. W. Woźniak, A Survey on Measuring Cognitive Workload in Human-Computer Interaction, *ACM Computing Surveys* (jan 2023).
- [13] T. J.-J. Li, A. Azaria, B. A. Myers, Sugilite: creating multimodal smartphone automation by demonstration, in: *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 6038–6049.
- [14] A. R. Sereshkeh, G. Leung, K. Perumal, C. Phillips, M. Zhang, A. Fazly, I. Mohamed, Vasta: a vision and language-assisted smartphone task automation system, in: *Proceedings of the 25th international conference on intelligent user interfaces*, 2020, pp. 22–32.
- [15] S. G. Tamilselvam, N. Panwar, S. Khare, R. Aralikatte, A. Sankaran, S. Mani, A visual programming paradigm for abstract deep learning model development, in: *Proceedings of the 10th Indian Conference on Human-Computer Interaction*, 2019, pp. 1–11.
- [16] E. Coronado, D. Deuff, P. Carreno-Medrano, L. Tian, D. Kulić, S. Sumartojo, F. Mastrogiovanni, G. Venture, Towards a modular and distributed end-user development framework for human-robot interaction, *IEEE Access* 9 (2021) 12675–12692.
- [17] T. J.-J. Li, End user programming of intelligent agents using demonstrations and natural language instructions, in: *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, 2019, pp. 143–144.
- [18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* 35 (2022) 27730–27744.
- [19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, *arXiv preprint arXiv:2302.13971* (2023).
- [20] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, *arXiv preprint arXiv:2211.05100* (2022).
- [21] G. Bimbatti, D. Fogli, L. Gargioni, Can chatgpt support end-user development of robot programs? (2023).
- [22] D. Sobania, M. Briesch, C. Hanna, J. Petke, An analysis of the automatic bug fixing performance of chatgpt, *arXiv preprint arXiv:2301.08653* (2023).
- [23] M. X. Liu, A. Sarkar, C. Negreanu, B. Zorn, J. Williams, N. Toronto, A. D. Gordon, “what it wants me to say”: Bridging the abstraction gap between end-user programmers and code-generating large language models, in: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–31.
- [24] J. White, S. Hays, Q. Fu, J. Spencer-Smith, D. C. Schmidt, Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design, *arXiv preprint arXiv:2303.07839* (2023).
- [25] N. M. S. Surameery, M. Y. Shakor, Use chat gpt to solve programming bugs, *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290* 3 (01) (2023) 17–22.
- [26] P. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. Garcia, P. Di Tommaso, S. Nahnsen, The nf-core framework for community-curated bioinformatics pipelines., *Nature biotechnology* 38 (3) (2020).
- [27] B. Almarie, P. E. Teixeira, K. Pacheco-Barrios, C. Rossetti, F. Fregni, The use of large language models in science: Opportunities and challenges, *Principles and Practice of Clinical Research* 9 (1) (2023) 1–4.
- [28] H. Hassani, E. S. Silva, The role of chatpt in data science: how ai-assisted conversational interfaces are revolutionizing the field, *Big data and cognitive computing* 7 (2) (2023) 62.
- [29] C. Liang, J. Karolus, T. Kosch, A. Schmidt, On the suitability of real-time assessment of programming proficiency using gaze properties, in: *Proceedings of the 7th ACM International Symposium on Pervasive Displays, PerDis '18*, Association for Computing Machinery, New York, NY, USA, 2018.
- [30] L. Wratten, A. Wilm, J. Göke, Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers, *Nature methods* 18 (10) (2021) 1161–1168.
- [31] F. Ahmad, W. Ahmad, An efficient astronomical image processing technique using advance dynamic workflow scheduler in cloud environment, *International Journal of Information Technology* 14 (6) (2022) 2779–2791.
- [32] J. M. Kunkel, L. R. Pedro, Potential of i/o aware workflows in climate and weather, *Supercomputing Frontiers and Innovations* 7 (2) (2020) 35–53.
- [33] F. Lehmann, D. Frantz, S. Becker, U. Leser, P. Hostert, Force on nextflow: Scalable analysis of earth observation data on commodity clusters., in: *CIKM Workshops*, 2021.
- [34] J. Yu, R. Buyya, A taxonomy of scientific workflow systems for grid computing, *ACM Sigmod Record* 34 (3) (2005) 44–49.
- [35] C. S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, J. I. V. Hemert, Scientific workflows: moving across paradigms, *ACM Computing Surveys (CSUR)* 49 (4) (2016) 1–39.
- [36] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, J. Myers, Examining the challenges of scientific workflows, *Computer* 40 (12) (2007) 24–32.
- [37] B. P. Harenslak, J. de Ruiter, Data Pipelines with Apache Airflow, Simon and Schuster, 2021.
- [38] J. Goecks, A. Nekrutenko, J. Taylor, G. T. team@ galaxyproject. org, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, *Genome biology* 11 (2010) 1–13.
- [39] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows, *Nature biotechnology* 35 (4) (2017) 316–319.
- [40] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. F. Da Silva, M. Livny, et al., Pegasus, a workflow management system for science automation, *Future Generation Computer Systems* 46 (2015) 17–35.
- [41] J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine, *Bioinformatics* 28 (19) (2012) 2520–2522.
- [42] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of*



- NAACL-HLT, 2019, pp. 4171–4186.
- [43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [44] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al., Palm 2 technical report, arXiv preprint arXiv:2305.10403 (2023).
- [45] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al., Lamda: Language models for dialog applications, arXiv preprint arXiv:2201.08239 (2022).
- [46] A. Yuan, A. Coenen, E. Reif, D. Ippolito, Wordcraft: story writing with large language models, in: 27th International Conference on Intelligent User Interfaces, 2022, pp. 841–852.
- [47] S. Petridis, N. Diakopoulos, K. Crowston, M. Hansen, K. Henderson, S. Jastrzebski, J. V. Nickerson, L. B. Chilton, Anglekindling: Supporting journalistic angle ideation with large language models, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–16.
- [48] E. Jiang, K. Olson, E. Toh, A. Molina, A. Donsbach, M. Terry, C. J. Cai, Promptmaker: Prompt-based prototyping with large language models, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–8.
- [49] B. Wang, G. Li, Y. Li, Enabling conversational interaction with mobile ui using large language models, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.
- [50] S. Wang, S. Petridis, T. Kwon, X. Ma, L. B. Chilton, Popblends: Strategies for conceptual blending with large language models, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–19.
- [51] H. Osone, J.-L. Lu, Y. Ochiai, Buncho: ai supported story co-creation via unsupervised multitask learning to increase writers’ creativity in japanese, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–10.
- [52] C. Wang, X. Liu, D. Song, Language models are open knowledge graphs, arXiv preprint arXiv:2010.11967 (2020).
- [53] P. Manakul, A. Liusie, M. J. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, arXiv preprint arXiv:2303.08896 (2023).
- [54] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, et al., Check your facts and try again: Improving large language models with external knowledge and automated feedback, arXiv preprint arXiv:2302.12813 (2023).
- [55] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374 (2021).
- [56] C. Clement, D. Drain, J. Timcheck, A. Svyatkovskiy, N. Sundaresan, Pymt5: multi-mode translation of natural language and python code with transformers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9052–9065.
- [57] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, et al., Competition-level code generation with alphacode, *Science* 378 (6624) (2022) 1092–1097.
- [58] H. Le, Y. Wang, A. D. Gotmare, S. Savarese, S. C. H. Hoi, Coderl: Mastering code generation through pretrained models and deep reinforcement learning, *Advances in Neural Information Processing Systems* 35 (2022) 21314–21328.
- [59] N. Jain, S. Vaidyanath, A. Iyer, N. Natarajan, S. Parthasarathy, S. Rajamani, R. Sharma, Jigsaw: Large language models meet program synthesis, in: Proceedings of the 44th International Conference on Software Engineering, 2022, pp. 1219–1231.
- [60] E. Jiang, E. Toh, A. Molina, K. Olson, C. Kayacik, A. Donsbach, C. J. Cai, M. Terry, Discovering the syntax and strategies of natural language programming with generative language models, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–19.
- [61] P. Vaithilingam, T. Zhang, E. L. Glassman, Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models, in: Chi conference on human factors in computing systems extended abstracts, 2022, pp. 1–7.
- [62] A. M. Dakhel, V. Majdinasab, A. Nikanjam, F. Khomh, M. C. Desmarais, Z. M. J. Jiang, Github copilot ai pair programmer: Asset or liability?, *Journal of Systems and Software* 203 (2023) 111734.
- [63] M. Kazemitabaar, J. Chow, C. K. T. Ma, B. J. Ericson, D. Weintrop, T. Grossman, Studying the effect of ai code generators on supporting novice learners in introductory programming, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–23.
- [64] P. Maddigan, T. Susnjak, Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models, *IEEE Access* (2023).
- [65] M. M. Hassan, A. Knipper, S. K. K. Santu, Chatgpt as your personal data scientist, arXiv preprint arXiv:2305.13657 (2023).
- [66] Z. Zahra, Z. Li, R. Filgueira, Laminar: A new serverless stream-based framework with semantic code search and code completion, arXiv preprint arXiv:2309.00584 (2023).
- [67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the *Journal of machine Learning research* 12 (2011) 2825–2830.
- [68] P. Sedgwick, N. Greenwood, Understanding the hawthorne effect, *Bmj* 351 (2015).
- [69] M. Sanvicente-García, A. García-Valiente, S. Jouide, J. Jaraba-Wallace, E. Bautista, M. Escobosa, A. Sánchez-Mejías, M. Güell, Crispr-analytics (crispr-a): A platform for precise analytics and simulations for gene editing, *PLoS Computational Biology* 19 (5) (2023) e1011137.
- [70] M. Sudmanns, D. Tiede, S. Lang, H. Bergstedt, G. Trost, H. Augustin, A. Baraldi, T. Blaschke, Big earth data: disruptive changes in earth observation data management and analysis?, *International Journal of Digital Earth* 13 (7) (2020) 832–850.
- [71] K. E. Lewińska, J. Buchner, B. Bleyhl, P. Hostert, H. Yin, T. Kuemmerle, V. C. Radeloff, Changes in the grasslands of the caucasus based on cumulative endmember fractions from the full 1987–2019 landsat record, *Science of Remote Sensing* 4 (2021) 100035.
- [72] K. E. Lewińska, P. Hostert, J. Buchner, B. Bleyhl, V. C. Radeloff, Short-term vegetation loss versus decadal degradation of grasslands in the caucasus based on cumulative endmember fractions, *Remote Sensing of Environment* 248 (2020) 111969.
- [73] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. El-nashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, arXiv preprint arXiv:2302.11382 (2023).
- [74] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, et al., Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics* 20 (17) (2004) 3045–3054.
- [75] E. L. Van Dijk, H. Auger, Y. Jaszczyszyn, C. Thermes, Ten years of next-generation sequencing technology, *Trends in genetics* 30 (9) (2014) 418–426.
- [76] T. Hu, N. Chitnis, D. Monos, A. Dinh, Next-generation sequencing technologies: An overview, *Human Immunology* 82 (11) (2021) 801–811.
- [77] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one fastq preprocessor, *Bioinformatics* 34 (17) (2018) i884–i890.
- [78] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, Star: ultrafast universal rna-seq aligner, *Bioinformatics* 29 (1) (2013) 15–21.
- [79] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short dna sequences to the human genome, *Genome biology* 10 (3) (2009) 1–10.
- [80] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie 2, *Nature methods* 9 (4) (2012) 357–359.
- [81] D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with hisat2 and hisat-genotype, *Nature biotechnology* 37 (8) (2019) 907–915.
- [82] A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, *Bioinformatics* 30 (15) (2014) 2114–2120.
- [83] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet. journal* 17 (1) (2011) 10–12.
- [84] C. Zeyen, L. Malburg, R. Bergmann, Adaptation of scientific workflows by means of process-oriented case-based reasoning, in: Case-Based Reasoning Research and Development: 27th International Conference, IC-CBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings

- 27, Springer, 2019, pp. 388–403.
- [85] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. Van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nature biotechnology* 28 (5) (2010) 511–515.
- [86] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools, *GigaScience* 10 (2), giab008 (02 2021). [arXiv:https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf](https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf), [doi:10.1093/gigascience/giab008](https://doi.org/10.1093/gigascience/giab008).  
URL <https://doi.org/10.1093/gigascience/giab008>
- [87] B. Rozière, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, et al., Code llama: Open foundation models for code, *arXiv preprint arXiv:2308.12950* (2023).
- [88] A. M. Kloft, R. Welsch, T. Kosch, S. Villa, "AI enhances our performance, I have no doubt this one will do the same": The Placebo Effect Is Robust to Negative Descriptions of AI, 2023. [arXiv:2309.16606](https://arxiv.org/abs/2309.16606).
- [89] T. Kosch, R. Welsch, L. Chuang, A. Schmidt, The Placebo Effect of Artificial Intelligence in Human-Computer Interaction, *ACM Transaction Computer-Human Interaction* 29 (6) (jan 2023).

## Appendix A. Questionnaire Study I

Table A.1: Feedback form for the first user study which investigates the capabilities of ChatGPT to capture the content of a workflow description. For each item, we added a comment field to report issues and errors in the generated explanations if the domain expert does not fully apply the content.

| Prompt | ID   | Question  | Answer options   |
|--------|------|---|--|
| P1_1   | Q1_1 | The generated explanation matches the research area of the workflow                               | 5-point Likert scale   |
|        | Q1_2 | The generated explanation matches the overall aim of the workflow                                 | 5-point Likert scale   |
| P1_2   | Q1_3 | How many of the tasks of the workflow (see task overview below) are contained in the description? | Numerical  |
|        | Q1_4 | The explanation describes the tasks of the workflow correctly                                     | (1) Don't know<br>(2) None of the tasks is correct<br>(3) Few tasks are correct<br>(4) Most tasks are correct<br>(5) All tasks are correct   |
|        | Q1_5 | How many of the software programs / tools used in the workflow are mentioned in the explanation?  | Numerical  |
|        | Q1_6 | The explanation of the software programs / tools used is correct                                  | (1) Don't know<br>(2) None of the software programs / tools is correct<br>(3) Few software programs / tools are correct<br>(4) Most software programs / tools are correct<br>(5) All software programs / tools are correct |
| P1_3   | Q1_7 | The explanation matches the input data specification of the workflow                              | 5-point Likert scale   |
| P1_4   | Q1_8 | The generated explanation matches the overall result of the workflow                              | 5-point Likert scale   |
| P1_5   | Q1_9 | How many of the generated research questions are valid?   | Numerical  |

## Appendix B. Questionnaire Study II

Table B.1: Feedback form for the second user study which investigates the capabilities of ChatGPT in exchanging the used tools in a scientific workflow. For each item we added a comment field to report issues and errors in the generated explanations if the domain expert doesn't fully apply with the content.

| Prompt | ID    | Question   | Answer options       |
|--------|-------|--|----------------------|
| P2_1   | Q2_1  | How many of the 10 alternative tools are valid?  | Numerical            |
| P2_2   | Q2_2  | The selected tools are reasonable alternatives for the task?   | 5-point Likert scale |
|        | Q2_3  | The generated explanation for the first tool highlights the suitability of the tool for the task well  | 5-point Likert scale |
|        | Q2_4  | The generated explanation for second tool highlights the suitability of the tool for the task well     | 5-point Likert scale |
| P2_3   | Q2_5  | The generated explanation helps to understand the methodical differences                               | 5-point Likert scale |
|        | Q2_6  | The generated explanation highlights strengthens and weaknesses of <i>[original-tool]</i> correctly    | 5-point Likert scale |
|        | Q2_7  | The generated explanation highlights strengthens and weaknesses of <i>[alternative-tool]</i> correctly | 5-point Likert scale |
| P2_4   | Q2_8  | The adaption of the workflow is correct  | 5-point Likert scale |
|        | Q2_9  | The adaption of the task <i>[task-name]</i> is correct   | 5-point Likert scale |
|        | Q2_10 | The explanation of the not-supported features by <i>[alternative-tool]</i> is correct                  | 5-point Likert scale |
|        | Q2_11 | Can the workflow be executed without errors?   | Yes/No               |
|        | Q2_12 | How long did it take to correct the workflow?  | Numerical            |
|        | Q2_13 | What had to be adapted to make the workflow executable?  | Free-text            |

### Appendix C. Questionnaire Study III

Table C.1: Feedback form for the third user study which investigates the capabilities of ChatGPT to extend a given (partial) workflow script. For each item we added a comment field to report issues and errors in the generated explanations if the domain expert doesn't fully apply with the content.

| Prompt  | ID   | Question   | Answer options       |
|---------|------|--|----------------------|
| P3_1    | Q3_1 | Is the list of necessary steps complete?   | Yes/No               |
|         | Q3_2 | How many of the proposed steps are correct? Please give your answer in the form of "x out of y steps".                 | Free text            |
|         | Q3_3 | How many of the proposed tools are correct? Please give your answer in the form of "x out of y tools".                 | Free text            |
| P3_2a/b | Q3_4 | The extension of the workflow is correct   | 5-point Likert scale |
|         | Q3_5 | Can the workflow be executed without errors?   | Yes/No               |
|         | Q3_6 | How long did it take to correct the workflow? Enter "not finished" when you set yourself a time limit that has run up. | Free text            |
|         | Q3_7 | What had to be adapted to make the workflow executable?  | Free text            |

## Appendix D. Schema WF4-Grasslands

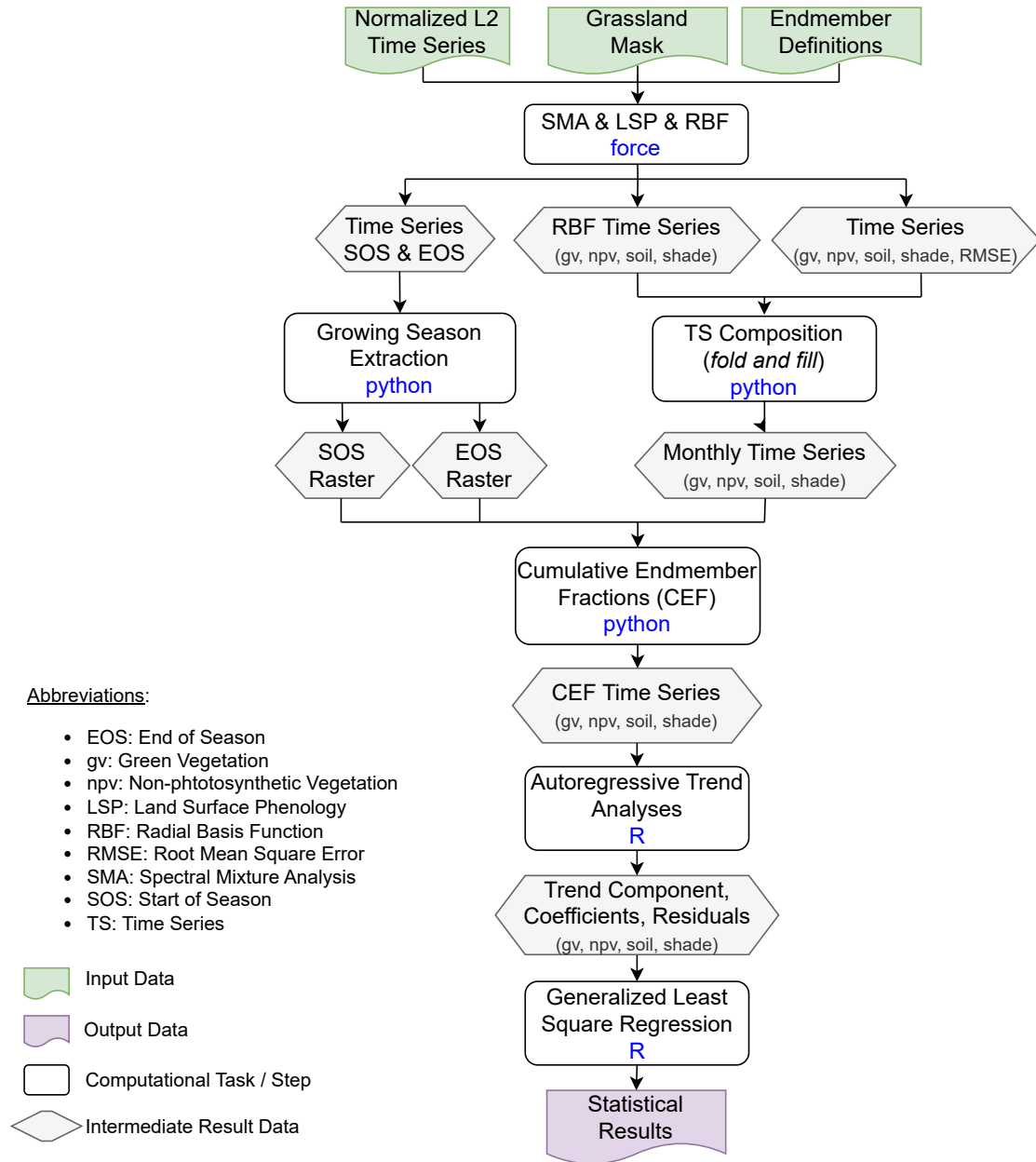


Figure D.1: Overview of the earth observation workflow *WF4-Grasslands* developed by one of the domain experts. The workflow aims at understanding differences in long-term changes (1984-2022) in ground cover fractions specific to European grasslands depending on the definition of endmembers (i.e., unique spectral signatures of a specific material or ground cover) approximating these fractions. The figure highlights the conceptual schema and data flow of the workflow.

## Appendix E. Verbal Task Description

In Study III, we use the following task description, provided by the earth observation expert, for extending *WF4-Grasslands* with an autoregressive trend analysis:

- step 1: it reads the data from a multi-band TIFF file into a raster brick as defined with the raster package. The input file needs to be selected based on a location specified by a string parameter passed into the script
- step 2: it converts all 0s in the data brick to 'NA'
- step 3: it fits per-pixel time trend with Auto Regressive trend function from the remotePARTS R package. Each raster in a data brick represents one time step in a time series over which the trend is fitted. The result of the trend fitting should be a multi-band raster in a TIFF format that comprises bands with information on trend's slope, intercept, p-value, temporal auto-correlation, and residuals for each time step.
- The steps 1-3 should be applied repetitively and independently to four raster datasets available for each location, namely: GV, NPV, SOIL, and SHADE.
- The script should be implemented using parallel cluster processing functionality in the raster package and provide control over the number of nodes, and RAM available to the process.

## Appendix F. GitHub Search Results

Table F.1: Statistics of the search results for four different scientific workflow systems using the GitHub search engine. For each system we use the system name as search term and restrict the result repositories to be created before the date give by the column (group).

| Search Term    | < 2021-09-01   |                 | < 2023-09-01   |           |
|----------------|----------------|-----------------|----------------|-----------|
|                | # Repositories | # Pull requests | # Repositories | # Commits |
| apache airflow | 1,200          | 27,000          | 2,600          | 85,000    |
| nextflow       | 1,700          | 5,000           | 3,200          | 14,000    |
| snakemake      | 1,900          | 5,000           | 3,400          | 18,000    |
| taverna        | 352            | 244             | 472            | 645       |