

Large Language Models for Software Engineering: A Systematic Literature Review

XINYI HOU*, Huazhong University of Science and Technology, China

YANJIE ZHAO*, Monash University, Australia

YUE LIU, Monash University, Australia

ZHOU YANG, Singapore Management University, Singapore

KAILONG WANG, Huazhong University of Science and Technology, China

LI LI, Beihang University, China

XIAPU LUO, The Hong Kong Polytechnic University, China

DAVID LO, Singapore Management University, Singapore

JOHN GRUNDY, Monash University, Australia

HAOYU WANG[†], Huazhong University of Science and Technology, China

Large Language Models (LLMs) have significantly impacted numerous domains, including Software Engineering (SE). Many recent publications have explored LLMs applied to various SE tasks. Nevertheless, a comprehensive understanding of the application, effects, and possible limitations of LLMs on SE is still in its early stages. To bridge this gap, we conducted a systematic literature review on LLM4SE, with a particular focus on understanding how LLMs can be exploited to optimize processes and outcomes. We collect and analyze 229 research papers from 2017 to 2023 to answer four key research questions (RQs). In RQ1, we categorize different LLMs that have been employed in SE tasks, characterizing their distinctive features and uses. In RQ2, we analyze the methods used in data collection, preprocessing, and application highlighting the role of well-curated datasets for successful LLM for SE implementation. RQ3 investigates the strategies employed to optimize and evaluate the performance of LLMs in SE. Finally, RQ4 examines the specific SE tasks where LLMs have shown success to date, illustrating their practical contributions to the field. From the answers to these RQs, we discuss the current state-of-the-art and trends, identifying gaps in existing research, and flagging promising areas for future study.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Software and its engineering** → **Software development techniques**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Software Engineering, Large Language Model, Survey

*Co-first authors who contributed equally to this work.

[†]Haoyu Wang is the corresponding author (haoyuwang@hust.edu.cn).

Authors' addresses: Xinyi Hou, xinyihou@hust.edu.cn, Huazhong University of Science and Technology, Wuhan, China; Yanjie Zhao, Yanjie.Zhao@monash.edu, Monash University, Melbourne, Australia; Yue Liu, yue.liu1@monash.edu, Monash University, Melbourne, Australia; Zhou Yang, zyang@smu.edu.sg, Singapore Management University, Singapore; Kailong Wang, wangkl@hust.edu.cn, Huazhong University of Science and Technology, Wuhan, China; Li Li, lilicoding@ieee.org, Beihang University, Beijing, China; Xiapu Luo, csxluo@comp.polyu.edu.hk, The Hong Kong Polytechnic University, Hong Kong, China; David Lo, davidlo@smu.edu.sg, Singapore Management University, Singapore; John Grundy, John.Grundy@monash.edu, Monash University, Melbourne, Australia; Haoyu Wang, haoyuwang@hust.edu.cn, Huazhong University of Science and Technology, Wuhan, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

XXXX-XXXX/2023/9-ART \$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

ACM Reference Format:

Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. 2023. Large Language Models for Software Engineering: A Systematic Literature Review. 1, 1 (September 2023), 62 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

In the field of language processing, traditional **Language Models (LMs)** have been foundational elements, establishing a basis for text generation and understanding [218]. Increased computational power, advanced machine learning techniques, and access to very large-scale data have led to a significant transition into the emergence of **Large Language Models (LLMs)** [378, 395]. Equipped with expansive and diverse training data, these models have demonstrated an impressive ability to simulate human linguistic capabilities, leading to a sea of changes across multiple domains. With their capacity to learn from massive corpora and generate plausible text, LLMs are blurring the line between human and machine-produced language. They have provided researchers and engineers alike with a powerful tool to explore the complexity and richness of human communication, consequently sparking a transformational period in the field of language processing and beyond.

Software Engineering (SE) – a discipline focused on the development, implementation, and maintenance of software systems – is one of those areas reaping the benefits of the LLM revolution [201]. The utilization of LLMs in SE primarily emerges from an innovative perspective where numerous SE challenges can be effectively reframed into data, code, or text analysis tasks [330]. Using LLMs to address these SE tasks has shown a wealth of potential breakthroughs [28, 32, 158, 295, 311, 354, 355, 385]. The applicability of LLMs is particularly pronounced in tasks such as code summarization [325], which involves yielding an abstract natural language depiction of a code’s functionality, as well as the generation of well-structured code [369] and code artifacts like annotations [182]. Codex, an LLM with 12 billion parameters, has demonstrated the ability to solve 72.31% of complex Python programming challenges posed by humans [38]. GPT-4 [241], an LLM from OpenAI, has been used with a strong performance in several SE tasks, encompassing code writing, understanding, execution, and reasoning. It not only handles real-world applications and diverse coding challenges but also shows the ability to explain results in natural language and generate code from pseudocode [26].

Simultaneously, researchers have embarked on a series of research activities regarding LLM-related works, where a number of literature reviews or survey papers have been produced [31, 68, 69, 362, 395]. Table 1 summarises some of these. However, these related studies have limitations. They either focus narrowly on a single SE scope, such as the application of LLMs in software testing [328] and natural-language-to-code (NL2Code) tasks [378], or they are primarily centered on Machine Learning (ML) or Deep Learning (DL) models [330, 364], overlooking more advanced and recently emerged LLM applications, such as ChatGPT [238], which are increasingly finding applications within the SE field [196, 296, 311, 346]. Alternatively, they merely offer a preliminary exploration of the performance of LLMs in various SE tasks through empirical experiments, without conducting a systematic literature survey [61, 201, 296, 357, 373, 395]. The integration of LLMs within SE is undoubtedly a complex endeavor, requiring key considerations including the choice of the right model, comprehension of the unique features of different LLMs, devising pre-training and fine-tuning strategies, handling of data, evaluation of outcomes, and surmounting implementation challenges [378]. Despite the burgeoning interest and ongoing explorations in the field, **a detailed and systematic review of LLMs’ application in SE has been notably absent in the current literature.** This gap signifies a need for understanding the relationship between LLMs and SE. In response, our research aims to bridge this gap, providing valuable insights to the community.

Table 1. State-of-the-art surveys related to LLMs for SE.

Reference	Year	Scope of models ¹	Scope of SE tasks	SLR ²	Time frame	# Collected Papers
Zan <i>et al.</i> [378]	2023	LLM (12M+)	NL2Code	×	2020-2023	Not specified
Zhao <i>et al.</i> [395]	2023	LLM (10B+)	Beyond SE scope	×	-	Not specified
Fan <i>et al.</i> [68]	2023	LLM	Beyond SE scope	×	2017-2023	5,752
Wang <i>et al.</i> [328]	2023	LLM (117M+)	Software testing	✓	2019-2023	52
Wang <i>et al.</i> [330]	2022	ML, DL ³	General SE scope	✓	2009-2020	1,209 (ML) + 358 (DL)
Yang <i>et al.</i> [364]	2022	DL	General SE scope	✓	2015-2020	250

¹ “M” means million and “B” means billion. The numbers in parentheses indicate the parameter sizes of LLMs.

² SLR stands for Systematic Literature Review. This column denotes whether the paper follows an SLR process.

³ ML and DL refer to Machine Learning and Deep Learning, respectively.

In this paper, we conduct a systematic literature review on the utilization of LLMs in SE (LLM4SE). By mapping the current state-of-the-art, pinpointing the key strengths, weaknesses, and gaps in the existing LLM4SE literature, and proposing potential avenues for future research, our review aims to provide researchers and practitioners with a thorough guide to the convergence of LLMs and SE. We anticipate that our findings will be instrumental in guiding future inquiries and advancements in this rapidly evolving field. This work makes the following key contributions:

- We are the first to present a comprehensive systematic literature review based on 229 papers published between 2017 and 2023 that focus on the use of LLM-based solutions to address SE challenges. We conducted a detailed analysis of the selected papers based on publication trends, distribution of publication venues, etc.
- We have classified the LLMs utilized for the reported SE tasks and have provided a summary of the usage and trends of different LLM categories within the SE domain.
- We describe the reported data processing stages, encompassing data collection, categorization, preprocessing, and representation.
- We discuss optimizers used for LLM4SE tasks, including parameter and learning rate optimization, prevalent prompt engineering techniques, and commonly employed evaluation metrics.
- We describe the key applications of LLM4SE encompassing a diverse range of 55 specific SE tasks, grouped into six core SE activities – requirements engineering, software design, software development, software quality assurance, software maintenance, and software management.
- We have summarised key challenges that using LLMs encounters within the SE field and have suggested several potential research directions for LLM4SE.

Section 2 presents our research questions (RQs) and elaborates on our systematic literature review (SLR) methodology. The succeeding Sections 3 to 6 are devoted to answering each of these RQs individually. Section 7 discloses the potential threats to the validity of our study. Section 8 discusses the challenges yet to be overcome when employing LLMs to solve SE tasks and highlights promising opportunities and directions for future research. Section 9 concludes the whole paper.

2 APPROACH

This systematic literature review (SLR) follows the methodology proposed by Kitchenham *et al.* [145, 146], used in most other SE-related SLRs [170, 194, 264, 330]. Following the guidelines provided by Kitchenham *et al.*, our methodology included three main steps: planning the review (i.e., Section 2.1, 2.2), conducting the review (i.e., Section 2.3, 2.4), and analyzing the basic review results (i.e., Section 2.5).

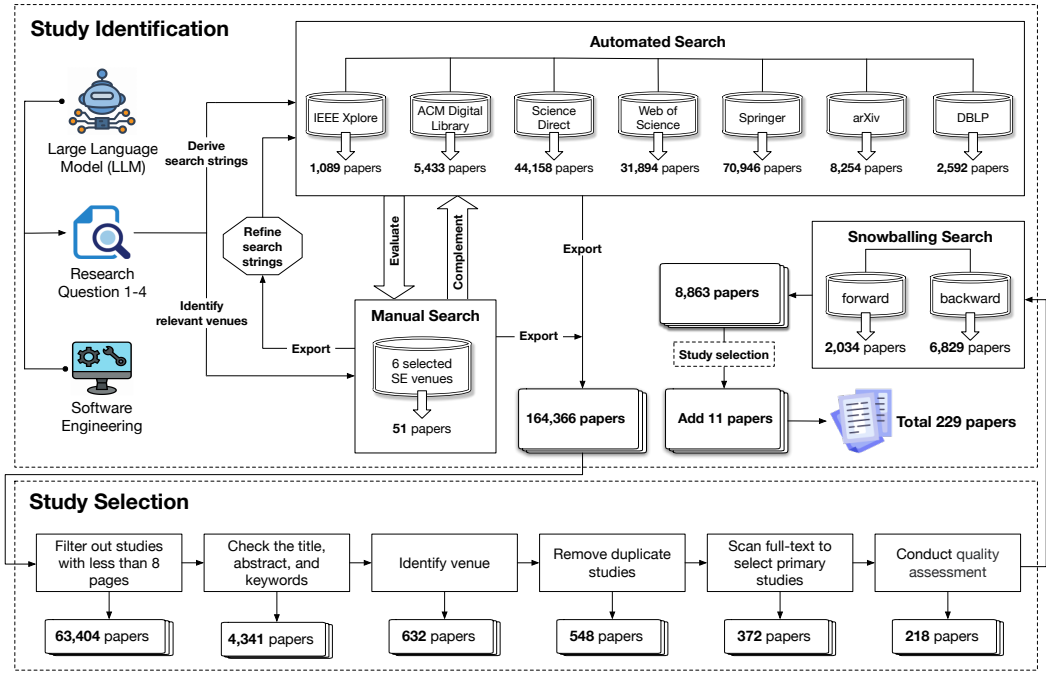


Fig. 1. Study identification and selection process.

2.1 Research Questions

To provide a comprehensive overview of the LLM4SE field, it is important to fully comprehend how these models are currently being applied in SE, the challenges they face, and their potential future research directions in SE. Thus, we aim to provide a systematic literature review of the application of LLMs to software engineering. This study thus aims to answer the following research questions:

- **RQ1: What LLMs have been employed to date to solve SE tasks?**
- **RQ2: How are SE-related datasets collected, preprocessed, and used in LLMs?**
- **RQ3: What techniques are used to optimize and evaluate LLM4SE?**
- **RQ4: What SE tasks have been effectively addressed to date using LLM4SE?**

2.2 Search Strategy

As shown in Fig.1, we employed the “Quasi-Gold Standard” (QGS) [380] approach for paper search. We conducted a manual search to identify a set of relevant studies and extracted a search string from them. This search string was then used to perform an automated search, and subsequently, a snowballing search was employed to further supplement the search results. This approach ensures both search efficiency and maximum coverage, minimizing the risk of omission. Subsequently, we employed a series of relatively strict filtering steps to obtain the most relevant studies. Specifically, we followed five steps to determine the relevance of the studies:

- (1) Select publication venues for manual search and select digital databases for automated search to ensure coverage of all the selected venues.
- (2) Establish QGS: Screen all papers for manual search and filter by inclusion/exclusion criteria (defined in Table 3).
- (3) Subjectively define the search string based on domain knowledge.

Table 2. Publication venues for manual search.

Acronym	Venues
ASE	International Conference on Automated Software Engineering
ESEC/FSE	Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering
ICSE	International Conference on Software Engineering
ISSTA	International Symposium on Software Testing and Analysis
TOSEM	Transactions on Software Engineering and Methodology
TSE	Transactions on Software Engineering

- (4) Conduct an automated search using the search string defined in Step (3).
- (5) Conduct snowballing search after performing study selection on the results of manual search and automated search.

2.2.1 Search Items. During the manual search, we selected six of the top SE conferences and journals (i.e., ICSE, ESEC/FSE, ASE, ISSTA, TOSEM, and TSE, as shown in Table 2) and searched for papers that applied LLM4SE. We systematically crawled a list comprising 4,618 published papers from the top venues. Following automated scanning via scripts, we manually verified and identified 51 papers that were relevant to our research objectives. These 51 relevant papers formed the basis for constructing the Quasi-Gold Standard (QGS). Our search string should combine two sets of keywords: one pertaining to SE tasks, and the other related to LLMs. Only if the paper contains both types of keywords there is a higher probability that it is the paper we need. The complete set of search keywords is as follows:

- *Keywords related to SE tasks:* Software Engineering, Software Development, Program*, Software Testing, Software Mainten*, SE, Software Lifecycle, Software Design*, Code representation, Code generation, Code comment generation, Code search, Code localization, Code completion, Code summarization, Method name generation, Bug detection, Bug localization, Vulnerability detection, Testing techniques, Test case generation, Program analysis, Bug classification, Defect prediction, Program repair, Code clone detection, Bug report, Software quality evaluation, SATD detection, Code smell detection, Compiled-related, Code review, Software classification, Code classification, Code change, Incident detection, Requirement extraction, Requirement traceability, Requirement validation, Effort cost prediction, Mining GitHub/Github mining, Mining SO (Stack Overflow)/SO mining, Mining app/App mining, Mining tag/Tag mining, Developer-based mining
- *Keywords related to LLMs:* LLM, Large Language Model*, Language Model*, LM, PLM, Pre-trained, Pre-training, Natural Language Processing, NLP, Machine Learning, ML, Deep Learning, DL, Artificial Intelligence, AI, Transformer, BERT, Codex, GPT, T5, Sequence Model*, Attention Model*, Transfer Learning, Neural Network*, ChatGPT, GPT-*

It is important to note that the list of keywords related to LLMs that we set up includes Machine Learning, Deep Learning, and other such terms that do not seem to be necessarily related to LLMs. The reason for this is that we want to avoid omitting papers related to our research as much as possible, so the process of performing automated searches expands our search scope.

2.2.2 Search Datasets. After determining the search string, we conducted an automated search across seven widely used databases, which are capable of covering all published or latest papers. Given that the first paper about the Transformer architecture [319], which forms the basis for LLMs, was published in 2017, we focused our search on papers published from that year onward¹. Two authors independently performed the automated search, and the search results from each database were merged and deduplicated. Specifically, we obtained 1,089 papers from IEEE Xplore,

¹The cut-off date for the paper collection process of this version is August 1st, 2023.

Table 3. Inclusion criteria and Exclusion criteria.

Inclusion criteria
1) The paper claims that an LLM is used
2) The paper claims that the study involves an SE task
3) The paper with accessible full text
Exclusion criteria
1) Short papers whose number of pages is less than 8
2) Duplicate papers or similar studies with different versions from the same authors
3) Studies belonging to books, thesis, monographs, keynotes, panels, or venues not executing a full peer-review process
4) Tool demos and editorials
5) The paper is published in a workshop or a doctoral symposium
6) The paper is a grey publication, e.g., a technical report or thesis
7) Non-English written literature
8) Literature mentioning the use of LLMs without describing the employed techniques

5,433 papers from the ACM Digital Library, 44,158 papers from ScienceDirect, 31,894 papers from Web of Science, 70,946 papers from Springer, 8,254 papers from arXiv, and 2,592 papers from DBLP.

2.3 Study Selection

2.3.1 *Study Inclusion and Exclusion Criteria.* Based on our search strategy, we initially obtained 164,366 papers that potentially relate to our research. Next, we needed to further evaluate the relevance of these papers based on inclusion and exclusion criteria, as shown in Table 3, so that the selected papers can directly address our research questions. The paper selection process, as illustrated in Fig. 1, consists of six phases.

In the first phase, we conducted automated filtering to exclude papers with less than 8 pages [19, 330] (Exclusion criteria 1), reducing the number of papers to 63,404. In the second phase, we examined the titles, abstracts, and keywords of the papers to identify those that include relevant LLM-related keywords. We then expanded the search scope to avoid missing relevant papers, including ML, DL, and other related keywords that may not directly correspond to LLM. The purpose of this phase is to narrow down the scope and filter out papers directly related to LLM (Inclusion criteria 1). Papers that are filtered out in this phase are then manually reviewed in the fifth phase. Additionally, we excluded 235 non-English written literature (Exclusion criteria 7). After the second phase, the number of papers was reduced to 4,341.

The third phase involves identifying the venues of the papers (Exclusion criteria 3). We extracted publication information such as “journal”, “URL”, “DOI”, and “series” to determine the publication sources. For papers from arXiv in 2022 and 2023, we chose to retain them, considering that this field is emerging and many works are in the process of submission. Although these papers did not undergo peer review, we have a quality assessment process to eliminate papers with low quality, ensuring the overall quality of this systematic literature review (SLR). This step resulted in 632 papers.

In the fourth phase, we merged and deduplicated the remaining papers from the seven databases and the manually searched paper list (Exclusion criteria 2), resulting in 548 papers. We then reviewed the full texts of the papers and excluded 176 papers that were grey publications or were published in workshops or doctoral symposiums (Exclusion criteria 4, 5, 6). By further assessing the quality of the papers, we identified 218 papers directly relevant to our research. This phase primarily involved excluding papers that mentioned LLMs but did not directly apply them, such as papers that only discussed LLMs in future work or focused on evaluating the performance of LLM-enabled tools [328] (Exclusion criteria 8). For SLR, survey, and review papers, we have retained them and

Table 4. Checklist of Quality Assessment Criteria (QAC) for LLM studies in SE.

ID	Quality Assessment Criteria
QAC1	Is the study relevant to SE tasks?
QAC2	Does the study utilize LLMs?
QAC3	Is the research not a secondary study, such as an SLR, review, or survey?
QAC4	Was the research published in a high-repute venue?
QAC5	Is there a clear motivation for the research?
QAC6	Does the study provide a clear description of the techniques used?
QAC7	Are the experimental setups, including experimental environments and dataset information, described in detail?
QAC8	Does the study clearly confirm the experimental findings?
QAC9	Are the key contributions and limitations of the study discussed?
QAC10	Does the study make a contribution to the academic or industrial community?

will assess their content during the quality assessment phase to determine their relevance to our research.

2.3.2 Study Quality Assessment. A well-crafted quality assessment can help to prevent biases introduced by low-quality studies and can indicate to readers where caution about conclusions should be drawn [363]. We formulated ten Quality Assessment Criteria (QAC), as shown in Table 4. These aim to assess the relevance, clarity, validity, and significance of included papers. We used a scoring system of -1, 0, 1 (irrelevant/unmet, partially relevant/met, relevant/fully met). The first three questions were designed for the remaining 382 papers in the fifth stage. If QAC1, QAC2, or QAC3 received a score of -1, there is no need to proceed with QAC4-QAC10, and the paper can be excluded directly. QAC4-QAC10 involved assessing the content of the papers using a scoring system of 0, 1, 2, 3 (poor, fair, good, excellent). Finally, we calculated the total score of QAC4-QAC10 for each paper. For published papers, the maximum score for QAC4-QAC10 should be 21 (3×7). We retained papers with a score of 16.8 (21×0.8) or above. For unpublished papers on arXiv, the score for QAC4 is always 0, and the maximum score for QAC5-QAC10 should be 18 (3×6). We retained papers with a score of 14.4 (18×0.8) or above. After this quality assessment, we obtained a final set of 218 papers.

2.4 Snowballing Search

To identify any additional possibly relevant primary studies, we conducted a snowballing search. Snowballing refers to using the reference list of a paper or the citations to the paper to identify additional papers. Snowballing could benefit from not only looking at the reference lists and citations but also complementing them with a systematic way of looking at where papers are actually referenced and where papers are cited. Using the references and the citations respectively is referred to as backward and forward snowballing.

Before conducting snowballing, a set of initial papers needs to be prepared. In this study, the initial paper list consists of the remaining 218 papers after the quality assessment. We performed forward and backward snowballing, which resulted in the collection of 2,034 and 6,829 papers, respectively. After initial deduplication, we were left with 3,350 papers. We then conducted the full study selection process on these 3,350 papers, including deduplicating them with the 218 papers from performing snowballing on the initial list. As a result, we obtained an additional 11 papers.

2.5 Data Extraction and Analysis

We finally obtained 229 relevant research papers after searching and snowballing. Fig. 2 presents an overview of the distribution of the included papers. As shown in Fig. 2 (a), 38% of papers are published in peer-reviewed venues. ICSE is the most common of these venues, with a contribution

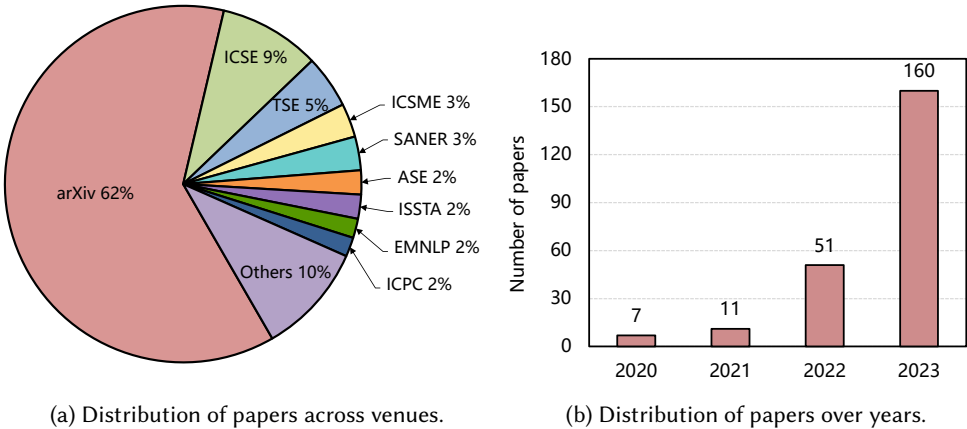


Fig. 2. Overview of the selected 229 papers' distribution.

of 9% of the total. Other venues with noteworthy contributions include TSE, ICSME, and SANER, contributing 5%, 3%, and 3% respectively. Meanwhile, the remaining 62% of papers are published on arXiv, an open-access platform that serves as a repository for scholarly articles. This finding is not surprising since much new LLM4SE research is rapidly emerging and thus many works are just completed and are likely in the peer review process. Despite the non-peer-reviewed nature of these papers, we have performed a rigorous quality assessment process on all collected papers, to ensure the quality of validity of our findings. This approach allows us to include all high-quality and relevant publications while maintaining high research standards.

Fig. 2 (b) shows the temporal distribution of the included papers. The number of publications has seen a rapidly growing trend since 2020. In 2020 and 2021, there are only 7 and 11 relevant papers, respectively. However, by 2022, the number of papers increases dramatically to 51. What's surprising is that, in the first half of 2023 alone, the number of published papers has already reached 160. This rapid growth trend demonstrates that there is a growing research interest in the domain of LLM4SE.

In order to visualize the main content of our collection of papers, we generated a word cloud based on the abstracts of 229 papers as shown in Fig. 3. The most frequently occurring words include “code”, “LLM”, “task”, “generation”, “performance”, and “program”, clearly indicating the main themes explored in these papers. The term “code” emphasizes the core elements of software engineering, while “LLM” denotes the use of large language models in a variety of tasks. The terms “generation” and “task” emphasize the use of the LLM for automatic code generation and other SE tasks. In addition, “performance” reflects the evaluation and assessment of the effectiveness of LLM in SE applications. The word cloud provides further visual evidence that the literature we have collected is closely related to our research topic, which is to investigate the application of LLM in SE tasks.

We then conducted data extraction during the full-text review. This extraction phase collected all relevant data that would facilitate a comprehensive and insightful response to the RQs outlined in Section 2.1. As depicted in Table 5, we extracted data including the classification of SE tasks, their corresponding activities, as well as the category, characteristics, and applicability of the LLMs. With this collected data, we systematically analyzed the relevant aspects of LLM application in the SE domain.

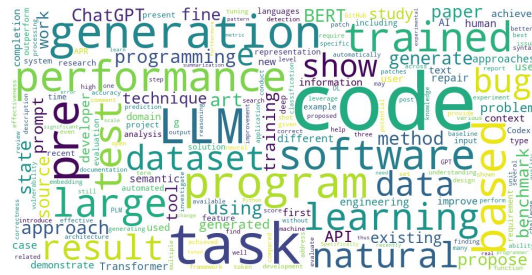


Fig. 3. Topics discussed in the collected papers.

Table 5. Extracted data items and related research questions (RQs).

RQ	Data Item
1,2,3,4	The category of SE task
1,2,3,4	The category of LLM
1,4	Characteristics and applicability of LLMs
2	The adopted data handling techniques
3	The adopted weight training algorithms and optimizer
3	The selected evaluation metrics
4	The SE activity to which the SE task belongs
4	The developed strategies and solutions

3 RQ1: WHAT LLMs HAVE BEEN EMPLOYED TO DATE TO SOLVE SE TASKS?

3.1 Large Language Models (LLMs)

Pre-trained language models (PLMs) have demonstrated impressive capabilities in solving various NLP tasks [150, 286, 342, 395]. Researchers have observed that scaling up the model sizes significantly enhances their capacity, leading to remarkable performance improvements when the parameter scale surpasses a certain threshold [111, 286, 306]. The term “Large Language Model” (LLM) was introduced to distinguish language models based on their parameter size, specifically referring to large-sized PLMs [395]. However, we note that the literature lacks a formal consensus on the minimum parameter scale for LLMs, as the model’s capacity is intertwined with both data size and total compute [328]. In this paper, we adopt the LLM scope division and taxonomy introduced by Pan *et al.*[246] and categorize the mainstream LLMs investigated in this study into three groups according to their architectures: encoder-only, encoder-decoder, and decoder-only LLMs. This taxonomy and relevant models are shown in Fig. 4.

Encoder-only LLMs. Encoder-only LLMs are a type of neural network architecture that utilizes only the encoder component of the model [52]. The encoder’s function is to process and encode the input sentence into a hidden representation, capturing the relationships between words and the overall context of the sentence. Notable instances of encoder-only LLMs include BERT [52] and its variants [74, 95, 159, 193]. As an example, BERT’s structure, based on the Transformer’s encoder architecture, has been referenced in 41 of the papers in this study. Its distinctive bidirectional attention mechanism simultaneously considers the left and right context of each word during training. In the SE domain, other prominent models like CodeBERT [74], GraphCodeBERT [95], RoBERTa [193], and ALBERT [159] have been widely employed. Specialized models such as BERTOverflow [301] and CodeRetriever [174] have been specifically developed for SE applications. These models’ innovations differ from BERT by leveraging the program structure, introducing new pre-training tasks, or engaging new modalities, thereby improving the architecture’s application to code-related tasks.

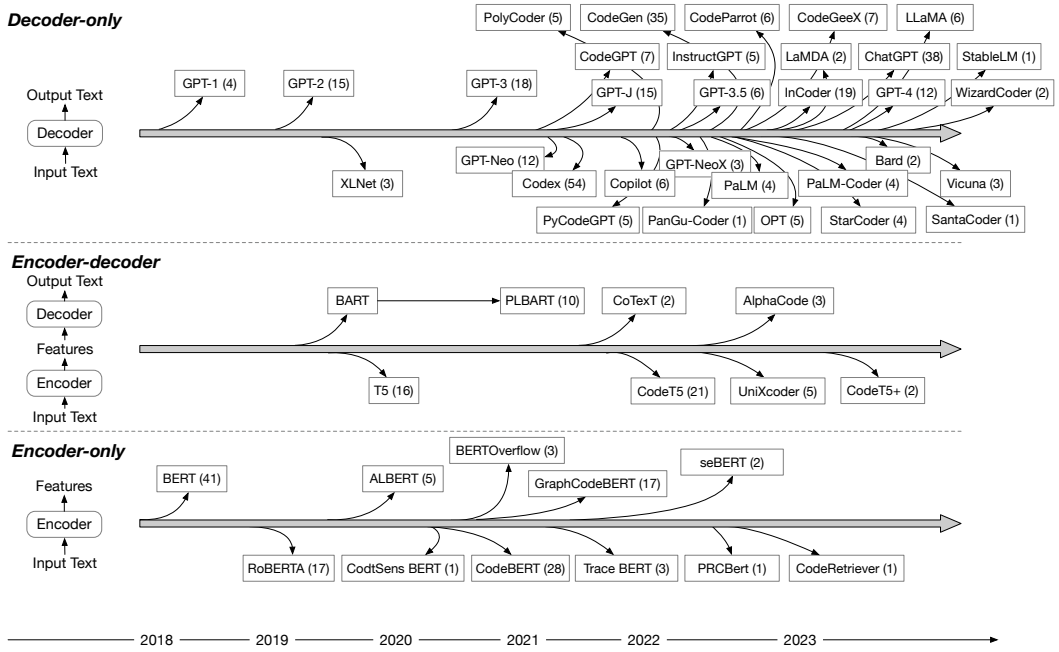


Fig. 4. Distribution of the LLMs (as well as LLM-based applications) discussed in the collected papers. The numbers in parentheses indicate the count of papers in which each LLM has been utilized.

For example, CodeBERT integrates a token prediction scheme to comprehend code by predicting subsequent tokens, enhancing its understanding of programming languages for tasks like code completion and bug detection [74]. GraphCodeBERT introduces edge-type prediction, recognizing relationships between code elements as a graph. This enables GraphCodeBERT to leverage code structure, improving its effectiveness in tasks like code summarization and program analysis [95]. These models have shown efficacy in tasks requiring a nuanced understanding of the entire sentence or code snippet. Examples include code review, bug report understanding, and named entity recognition pertaining to code entities [16, 172, 220, 258, 285, 359].

Encoder-decoder LLMs. Encoder-decoder LLMs incorporate both encoder and decoder modules [319]. The encoder ingests the input sentence and encodes it into a hidden space, effectively capturing the underlying structure and semantics. This hidden representation serves as an intermediary language, bridging the gap between diverse input and output formats. Conversely, the decoder utilizes this hidden space to generate the target output text, translating the abstract representation into concrete and contextually relevant expressions. Models such as PLBART [3], T5 [262], and CodeT5 [340] embody this architecture. Further advancements are evident in CodeT5+ [337], while AlphaCode [177] and CoText [255] showcase the architecture's adaptability to various SE tasks. The encoder-decoder design offers flexible training strategies and is proficient in handling multifaceted tasks such as summarization, translation, and question-answering. Within the field of SE, this ability has been successfully applied to tasks like code summarization [6, 91, 212]. The encoder module's capacity to understand and represent both the structure and semantics of code is pivotal, allowing the decoder to translate this comprehension into concise, human-readable summaries.

Decoder-only LLMs. Decoder-only LLMs exclusively utilize the decoder module to generate the target output text, following a distinct training paradigm that emphasizes sequential prediction [260].

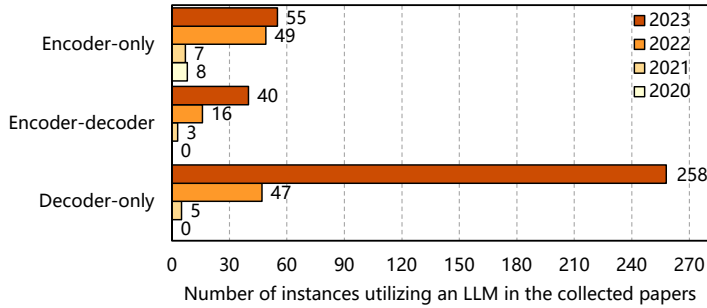


Fig. 5. Trends in the application of LLMs with different architectures in SE tasks over time.

Unlike the encoder-decoder architecture, where the encoder processes input text, the decoder-only architecture begins with an initial state and predicts subsequent tokens, gradually building the output text. This approach relies heavily on the model's ability to understand and anticipate language structure, syntax, and context. GPT-series models, such as GPT-1 [260], GPT-2 [261], GPT-3 [25], GPT-3.5 [239], GPT-4 [241], as well as their notable derivative, ChatGPT [238]², represent their major implementations. More specialized versions like CodeGPT [195], InstructGPT [242], Codex [38], Copilot [87]³, and others have been fine-tuned for specific tasks in SE. Open-source models like GPT-J [326], GPT-Neo [24], GPT-NeoX [23], LLaMA [313], and Vicuna [43] also follow this architecture. These models can generally perform downstream tasks from a few examples or simple instructions without adding prediction heads or fine-tuning, making them valuable tools in SE research. **The year 2022 marked a surge in the development of decoder-only LLMs, a trend that gained further momentum in 2023, notably with the launch of commercial products by leading Internet companies.** For example, Google launched Bard [90], Meta introduced LLaMA [313] and Llama 2 [314], Microsoft unveiled Bing Chat [214], etc. Contrary to LLMs such as GPT-4 and its derivative application, ChatGPT, released by OpenAI, which were promptly integrated into SE tasks, these new additions have not yet found widespread application within the SE field. Their potential remains largely unexplored, with opportunities for further assessment and utilization in specific tasks and challenges. The continued advancement of these models emphasizes the active exploration and innovation within decoder-only architectures.

3.2 Trend Analysis

As shown in Fig. 5, in the span from 2020 to 2023, the architecture of LLMs has witnessed notable shifts in preference and application within SE tasks. The specific choices between decoder-only, encoder-decoder, and encoder-only structures have shaped the direction of research and solutions in the SE domain [348]. This analysis explores trends in the adoption of these architectures over the years, reflecting the evolving dynamics of LLM for SE tasks.

Evolution of LLM architectures in 2021. The year 2020 saw research papers predominantly concentrating on encoder-only LLMs for SE tasks, evidenced by a total of eight papers. Decoder-only LLMs or encoder-decoder LLMs were not featured in that year's research. A marked change occurred in 2021. Out of 15 papers in 2021, five were dedicated to decoder-only LLMs, constituting 33.33% of

²ChatGPT is a conversational agent built upon the GPT architecture, with GPT-3.5 and GPT-4 being specific versions of the architecture, each representing successive advancements.

³Copilot is an application built upon LLMs tailored for coding tasks. **For convenience, all subsequent references in this paper to LLMs and their applications, such as ChatGPT and Copilot, will collectively be referred to as LLMs.**

the research. Additionally, three papers, or 20%, focused on encoder-decoder LLMs. Encoder-only LLMs witnessed a slight decline, representing 46.67% of the field with seven papers. This rapid transition can be linked to the generative capability of decoder-only LLMs. Researchers [160, 277, 296] found that these models, e.g., GPT series, requiring minimal fine-tuning, could produce not only syntactically correct but also functionally relevant code snippets. Their proficiency in grasping the context of code quickly made them a preferred choice.

Diversity of LLM architectures in 2022. 2022 experienced a significant increase in diversity, with more varied LLM architectures finding representation. Out of a total of 112 papers, 47 were centered around decoder-only LLMs, comprising 41.96% of the studies. Encoder-decoder LLMs made their presence known in 16 papers, accounting for 14.29%. Meanwhile, encoder-only LLMs led the field slightly with 49 papers, capturing 43.75% of the research interest. This diverse distribution suggests an exploration phase where researchers were actively assessing and leveraging different architectures to suit varied needs and challenges. The near-equal interest across different architectures underscores the field's richness, indicating that no single approach had become the definitive choice.

Dominance of the decoder-only architecture in 2023. 2023 signaled a strong shift towards decoder-only LLMs. An impressive 258 instances of utilizing decoder-only LLMs were recorded across 138 unique papers, reflecting that a single paper might employ multiple such models. These papers focusing on decoder-only LLMs constituted a significant 73.09% of the total research this year. In comparison, encoder-decoder LLMs were the subject of 40 papers, contributing 11.33%, while encoder-only LLMs appeared to stabilize, with 55 papers, representing 15.58% of the 2023 research landscape. This trend signifies a shift in focus and resources toward exploring and harnessing the decoder-only architecture as the primary approach in many current and future LLM4SE research and applications.

Criteria for LLM selection in SE tasks. The selection of an LLM for SE tasks should involve careful consideration rather than arbitrary choice. Key factors guiding this selection encompass the model's proficiency in understanding the context of code, its ability to generate relevant content, responsiveness to fine-tuning, and demonstrated performance on SE-specific benchmarks [168, 178, 356]. Given the stringent syntactical rules and functional requirements inherent to SE tasks, models capable of seamlessly integrating these complex aspects were typically favored.

Task-specific fine-tuning. A notable trend is the customization of LLMs for precise SE tasks [124, 172, 384]. By fine-tuning models with datasets tailored to specific functions such as bug detection or code review, researchers were able to achieve marked performance improvements [45, 152].

In conclusion, the evolution of LLMs for SE, transitioning from encoder-only to decoder-only architectures, highlights the field's vibrancy and adaptability. This shift has fundamentally altered the approach to SE tasks, reflecting the ongoing innovation within the discipline.

RQ1 - Summary

- (1) There are more than 50 different LLMs used for SE tasks in the papers we collected. Based on the underlying architecture or principles of different LLMs, we classified the summarized LLMs into three categories, i.e., encoder-only, encoder-decoder, and decoder-only LLMs.
- (2) We analyzed the trend of LLM usage for SE tasks. The most widely used LLMs are with decoder-only architectures. There are over 30 LLMs in the decoder-only category and 138 papers have researched the application of decoder-only LLMs to SE tasks.

4 RQ2: HOW ARE SE-RELATED DATASETS COLLECTED, PREPROCESSED, AND USED IN LLMS?

Data plays a crucial role in the model training phase [299]. First, data is collected to obtain diversity and richness to ensure that the model can cope with different scenarios and situations. Second, data is classified to clarify the training objectives of the model and avoid confusion and misinformation. The preprocessing of data is indispensable to clean and transform the data to improve its quality. Finally, data is formatted into a structure suitable for model processing, allowing the LLM to effectively learn the data's features and patterns. We analyze the reported processes of data collection, data classification, data preprocessing, and data representation in our selected primary studies on LLM4SE.

4.1 How are the datasets for training LLMs sourced?

Data is an indispensable and critical factor in the training of LLMs, which determines the generalization ability, effectiveness, and performance of the models [299]. Adequate, high-quality, and diverse data is critical to allow models to fully learn features and patterns, optimize parameters, and ensure reliability in validation and testing. We first investigate the methods used to obtain the dataset. By analyzing the methods of data collection, we divided the data sources into four categories: open-source datasets, collected datasets, constructed datasets, and industrial datasets. *Open-source datasets* [33, 140, 329, 379] refer to publicly accessible collections of data that are often disseminated through open-source platforms or repositories. For example, datasets like HumanEval [38], which consists of 164 manually crafted Python problems, each accompanied by its respective unit tests. The open-source nature of these datasets ensures their credibility and allows for community-driven updates, making them a reliable resource for academic research. *Collected datasets* [116, 211, 285, 311] are those that researchers compile directly from a multitude of sources, including but not limited to, major websites, forums, blogs, and social media platforms. For instance, researchers [30, 280, 344, 359] often scrape data from Stack Overflow [244] threads or GitHub [86] issue comments to create a dataset tailored to their specific research questions. *Constructed datasets* [66, 137, 149, 381] are specialized datasets that researchers create by modifying or augmenting collected datasets to better align with their specific research objectives. These modifications can be carried out through manual or semi-automatic methods and may include the generation of domain-specific test sets, annotated datasets, or synthetic data. For example, researchers often take a collected dataset of code snippets and manually annotate them with bug types to create a constructed dataset for studying automated program repair techniques [70, 131, 351]. *Industrial datasets* [7, 216, 338] are those obtained from commercial or industrial entities and often contain proprietary business data, user behavior logs, and other sensitive information. These datasets are particularly valuable for research that aims to address real-world business scenarios. However, the acquisition of such datasets is often complicated by issues related to business confidentiality and data privacy. For example, in a collaborative effort with China Merchants Bank (CMB), Wang *et al.* [338] were able to access 21 projects from CMB's repositories. Access to such data would likely require non-disclosure agreements and other legal safeguards to protect business interests. Each of these dataset types offers unique advantages and challenges, and the choice between them should be guided by the specific requirements and constraints of the research project at hand.

Fig. 6 shows the collection strategies of LLM-related datasets. As can be seen from the data in the figure, **127 studies used open-source datasets for training large models**. One of the main reasons for using open-source datasets in LLM training is their authenticity and credibility. Open-source datasets usually contain real-world data collected from various sources (such as relevant studies that have been conducted), which makes them highly reliable and representative

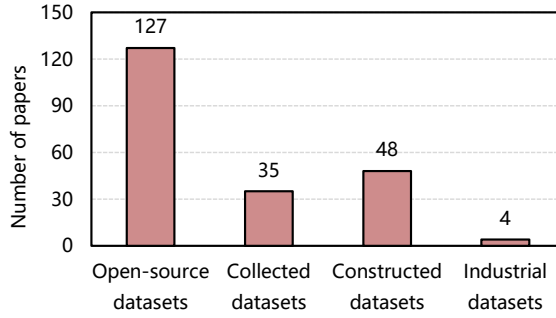


Fig. 6. The collection strategies of LLM-related datasets.

of real-world scenarios. This helps LLMs learn from real examples to better understand real-world applications and improve their performance. Second, since LLMs are a topic that has just recently emerged, a lack of suitable training sets does exist. Therefore, researchers often collect data from sites such as Stack Overflow and GitHub and build datasets to make the data more composite for SE tasks. **Out of the 229 papers we studied, we found that only four of these studies were using industrial datasets.** This suggests a potential misalignment between the properties of datasets used in academic research and those encountered in real-world industrial contexts. This divergence underscores the need for future research to investigate industrial datasets, thereby ensuring that LLMs are applicable and robust across both academic and industrial scenarios.

Note that some papers use multiple datasets that span different categories, e.g., Xu *et al.* [357] evaluated the performance of Codex, GPT-J, GPT-Neo, and other LLMs on SE tasks, and Mastropaolo *et al.* [212] investigated the use of T5 in several code-related tasks such as fixing bugs and generating code comments. For different LLMs or different SE tasks, researchers may use different training datasets. On the other hand, some papers focus on exploring how existing LLMs (e.g., ChatGPT) are used in SE tasks [346] and do not specify the dataset used for model training, as these LLMs like ChatGPT often do not require users to prepare training data themselves for general usage scenarios.

4.2 What types of SE datasets have been used in existing LLM4SE studies?

Data types play a pivotal role in shaping the architecture and selection of LLMs, as they directly influence the extraction of implicit features and subsequent model decisions[30, 84, 290, 361]. The choice of data types can significantly impact the overall performance and generalization ability of the LLMs. We examine and classify the types of SE datasets employed in LLM4SE studies. By investigating the relationship between data types, model architectures, and performance, we seek to shed light on the critical role of data types in the success of LLM4SE applications.

Data type categorization. We classified the data types of all datasets into five categories: code-based, text-based, graph-based, software repository-based, and combined data types. Table 6 describes the specific data included in the data types corresponding to the datasets we summarized from the 229 studies. We can find that **most of the studies used text-based datasets, accounting for a total of 104.** The dominance of text-based datasets in training LLMs for SE tasks highlights the models' exceptional natural language processing capabilities. These LLMs excel in understanding and processing textual data, making them an ideal choice for tasks that involve code comprehension, bug fixing, code generation, and other text-oriented SE challenges. Their ability to process and learn

Table 6. Data types of datasets involved in prior studies.

Category	Data type	# Studies	Total	References
Code-based datasets	Source code	44	70	[4] [16] [27] [29] [32] [35] [46] [81] [82] [96] [112] [118] [124] [127] [135] [156] [167] [180] [185] [201] [211] [221] [229] [235] [245] [257] [275] [287] [288] [290] [293] [303] [324] [337] [344] [356] [348] [361] [371] [379] [387] [393] [399] [403]
	Bugs	4		[136] [144] [352] [354]
	Patches	4		[162] [309] [310] [390]
	Code changes	4		[84] [171] [249] [383]
	Test suites/cases	3		[70] [126] [389]
	Error code	3		[117] [250] [347]
	Vulnerable source code	2		[30] [42]
	Bug-fix pairs	2		[67] [372]
	Labeled clone pairs	1		[61]
	Buggy programs	1		[28]
	Packages	1		[283]
	Flaky test cases	1		[72]
Text-based datasets	Prompts	28	104	[29] [58] [60] [62] [85] [128] [129] [139] [140] [165] [172] [200] [282] [292] [294] [302] [304] [305] [312] [316] [320] [329] [331] [332] [345] [358] [367] [394]
	Programming problems	14		[38] [33] [107] [132] [155] [168] [176] [277] [279] [295] [311] [335] [377] [375]
	SO (i.e., Stack Overflow) posts	9		[22] [102] [103] [105] [115] [152] [263] [343] [392]
	Bug reports	9		[45] [73] [88] [105] [122] [137] [163] [197] [220]
	Programming tasks (and solutions)	7		[54] [59] [70] [143] [153] [231] [278]
	Requirements documentation	6		[66] [109] [151] [199] [217] [338]
	APIs/API documentation	5		[50] [141] [248] [359] [376]
	Q&A pairs	5		[259] [284] [321] [329] [402]
	Vulnerability descriptions	4		[10] [251] [308] [351]
	Bug reports with changesets	4		[44] [50] [105] [131]
	Methods	3		[209] [211] [374]
	Code comments	2		[257] [357]
	Project issues	2		[78] [298]
	Software specifications	1		[205]
	Dockerfiles	1		[108]
	Semantic merge conflicts	1		[382]
	Site text	1		[149]
	User intents	1		[125]
	User reviews	1		[339]
Graph-based datasets	GUI Images	1	1	[151]
Software repository-based datasets	Issues and commits	3	5	[7] [183] [392]
	Pull-requests	2		[285] [392]
Combined datasets	Source code and comments	8	18	[91] [158] [234] [257] [276] [318] [384] [388]
	Programming tasks and test suites/cases	4		[192] [236] [243] [291]
	Binary code and related annotations	1		[6]
	Failing test code and error messages	1		[355]
	Source code and Q&A pairs	1		[280]
	Source code and test suites/cases	1		[317]
	Source code, methods, and logging statements	1		[178]
	Source code, description, and code environment	1		[184]

from vast amounts of text data enables them to provide powerful insights and solutions for various

SE applications. **Text-based datasets with a large number of prompts (28) are commonly used in training LLMs for SE tasks to guide their behavior effectively.** While understanding the training data may not be essential for closed-source LLMs like ChatGPT, insights into the data handling techniques of other models remain valuable. This is particularly true as black-box models can be fine-tuned with small-sized data inputs during usage. Among the 229 surveyed papers, this understanding is reinforced by the fact that text-based datasets with a large number of prompts are the most frequently used data types for training LLMs in SE tasks. programming problems (14) are also essential as they provide diverse and challenging tasks, allowing models to generalize knowledge and skills for various SE challenges. This combination helps the models develop a robust understanding of software concepts and perform well in a wide range of tasks. There are also SO (i.e., Stack Overflow) posts (9), bug reports (9), programming tasks (and solutions) (7), etc., which are among the more numerous data types in text-based datasets.

The predominance of source code (44) as the most abundant data type in code-based datasets can be attributed to its fundamental role in SE. Source code serves as the foundation of any software project, containing the logic and instructions that define the program's behavior. Therefore, having a large volume of source code data is crucial for training LLMs to understand the intricacies of software development, enabling them to effectively generate, analyze, and comprehend code in various SE tasks. There are also common data types such as bugs (4) and patches (4) for program repair tasks. Graph-based datasets are used in some research studies for SE tasks, e.g., Kolthoff *et al.* [151] used a dataset composed of screenshots from Google Play Android applications to construct a graphical user interface (GUI) repository in their study on LLM for the rapid prototyping task. These datasets represent code using graph structures, capturing relationships and dependencies between code components.

Software repository-based datasets usually mean data collected from software version control systems (e.g., Git) and issue tracking systems (e.g., GitHub, Jira, etc.). This data includes issues and commits (3), pull requests (2), and so on. The data in software repositories can provide a wealth of information covering all aspects of the software development process, including code evolution history, records of issue fixes and feature improvements, code quality assessments, and so on. These data are valuable for studying behaviors and trends in the software development process, improving software quality and development efficiency, and evaluating the performance of software engineering techniques. Therefore, many studies have used software repository-based datasets for empirical analysis and model training.

Some studies employed combined datasets containing multiple datatypes. Among them, the most common type is "source code and comments". For instance, Tufano *et al.* [318] used a dataset comprising source code and comments to train a model and showed that a pre-trained text-to-text converter (T5) model outperforms a previous deep learning model in automating the code review task. Other combinations of data types include "binary code and related annotations", "failing test code and error messages", "source code and Q&A pairs", "source code, description, and code environment", etc.

4.3 How do data types influence the selection of data-preprocessing techniques?

For the training and application of LLMs, the raw dataset needs to be subjected to data processing to obtain a clean and suitable dataset for model training. The data processing steps [163, 206] involve operations such as data cleaning, noise removal, normalization, etc. To ensure consistency and quality of the data, different data types may require different processing methods to improve the performance and effectiveness of LLMs in SE tasks. In this section, we aim to detail the data preprocessing procedures for the two most used types of datasets, i.e., code-based datasets and text-based datasets.

Table 7. The data preprocessing procedure for code-based datasets.

Preprocessing techniques	Description	Examples	References
Data extraction	Extract relevant code blocks for specific software engineering tasks from code-based datasets, considering different granularities and requirements.	Token-level, statement-level, method-level, file-level, or project-level.	[30] [84] [136] [361]
Unqualified data deletion	Eliminate unqualified data by applying filtering rules to retain appropriate samples, ensuring the dataset's quality and relevance for various software engineering tasks.	Retain only code longer than a certain number of lines, or remove files or methods that contain a certain keyword.	[126] [167] [257] [290]
Duplicated instance deletion	Remove duplicated instances from the dataset to ensure data integrity and prevent redundancy in the training process.	Remove near-duplicate code samples using certain deduplication algorithms.	[46] [282] [357]
Data compilation	Compile the code to get correctly compilable files.	Converting java files to .class files throughout the compilation process.	[30] [210]
Uncompilable data deletion	Remove non-compilable code fragments.	Remove code fragments that fail compilation, such as those with syntax errors.	[310]
Code representation	Represented as tokens.	Tokenize source or binary code as tokens.	[357] [180]
	Represented as trees.	Parses source or binary code into AST.	[6] [124]
	Represented as graphs.	Generate source or binary code as PDG (CFG, CG).	[201] [393]
Data segmentation	Split the dataset for use in a training model, validation model, or test model.	Divide the data set according to certain rules, which can be divided into training sets, validation sets, or test sets.	[46] [344]

The data preprocessing procedure for code-based datasets. We now summarize the process of preprocessing a code-based dataset, which consists of seven steps. Table 7 describes the individual data processing steps in detail and gives examples. The first step is data extraction, which involves retrieving relevant code segments from different sources such as software repositories or version control systems [136, 361]. Depending on the requirements of the research task [212, 374], code segments can be extracted at different levels of granularity, ranging from individual methods and functions to entire source code files or even complete software projects. The next step is to remove any code segments that do not meet predefined criteria or quality standards [167, 257, 290]. This filtering process ensures that the extracted code is relevant to the specific SE task under study, thus eliminating incomplete or irrelevant code snippets. To avoid introducing bias and redundancy during model training, the third step involves removing duplicate instances [46, 357, 397]. Any duplicate code instances are identified and removed from the dataset, thus increasing the diversity and uniqueness of the data. After the data extraction and filtering steps, the fourth step, data compilation, comes into play. The extracted and filtered code segments are merged and compiled into a unified code dataset. This compilation process simplifies data storage and access and facilitates subsequent analysis and model training [30, 210]. In the fifth step, the problem of invalid or non-executable code is solved by removing data that cannot be compiled. Any code segments that cannot be compiled or executed are removed from the dataset to ensure that the remaining code instances are valid and usable during model training and evaluation. The sixth step is code representation, which consists of converting the code segments into a suitable representation that

Table 8. The data preprocessing procedure for text-based datasets.

Preprocessing techniques	Description	Examples	References
Data extraction	Extract valid text from documentation according to different software engineering tasks.	Bug report, requirement documentation, code comments, API documentation, etc.	[361] [44] [66] [45]
Initial data segmentation	Split data into different categories as required.	Split data into sentences or words.	[102] [152]
Unqualified data deletion	Delete invalid text data according to the specified rules.	Remove the source code from the bug report.	[88] [220] [257]
Text preprocessing	Preprocessing operations on text.	Remove certain symbols and words, or convert all content to lowercase.	[263] [338]
Duplicated instance deletion	Remove duplicate samples from the dataset.	Utilize the deduplication algorithm from CodeSearch-Net [119] to remove nearly duplicate methods.	[357]
Data tokenization	Use token-based text representation.	Tokenize the texts, sentences, or words into tokens.	[199]
Data segmentation	Split the dataset for use in a training model, validation model, or test model.	Divide the data set according to certain rules, which can be divided into training sets, validation sets, or test sets.	[163]

can be processed by the LLMs. This conversion can take different forms: token-based representation involves tokenizing the source or binary code into distinct tokens; tree-based representation parses the code into Abstract Syntax Trees (AST); and graph-based representation generates a Program Dependence Graph (PDG), encompassing Control Flow Graphs (CFG) and Call Graphs (CG). Finally, in the “data segmentation” step, the preprocessed dataset is partitioned into different subsets for training, validation, and testing [46, 344]. The training set is used to train the LLM, the validation set helps to tune the hyperparameters and optimize the model performance, and the testing set evaluates the model’s ability on unseen data. By strictly adhering to these seven preprocessing steps, researchers can create structured and standardized code-based datasets, thus facilitating the effective application of LLMs for a variety of SE tasks such as code completion, error detection, and code summarization.

The data preprocessing procedure for text-based datasets. As displayed in Table 8, the steps of text-based dataset preprocessing consist of a total of seven steps, but there are some differences from the code-based dataset preprocessing steps. The process begins with data extraction [44, 45, 66, 361], where relevant text is carefully extracted from SE documentation from a variety of sources, including bug reports [45], requirements documents [151], code comments [257], and API documentation [141]. This step ensures that the dataset captures diverse, task-specific textual information. After data extraction, the text is initially segmented and categorized according to the specific requirements of the research task. For example, the text can be segmented into sentences or further broken down into individual words as needed for analysis [102, 152]. To ensure the quality and relevance of the dataset, substandard data deletion is performed to eliminate any invalid or irrelevant text. For example, the dataset used by Lee *et al.* [163] was constructed from bug reports, and in the “unqualified data deletion” process the researchers filtered out bug reports with fewer than 15 words because the text was too short to contain contextual information. Next, preprocessing operations are performed on the text to standardize and clean it. Common preprocessing steps include removing certain symbols, stop words, and special characters [263, 338]. This standardized form of text facilitates the efficient processing of LLMs. To avoid introducing bias and redundancy

Table 9. The various input forms of LLMs proposed in prior studies.

Category	Input forms	# Studies	Total	References
Token-based input	Code in tokens	64	192	[4] [16] [27] [29] [30] [35] [38] [39] [42] [46] [61] [67] [72] [82] [84] [96] [118] [117] [124] [126] [127] [135] [136] [144] [156] [157] [158] [162] [167] [168] [185] [189] [210] [211] [229] [235] [245] [249] [250] [267] [276] [283] [287] [290] [288] [293] [303] [309] [310] [317] [324] [347] [352] [354] [356] [348] [371] [379] [387] [389] [390] [383] [399] [403]
	Text in tokens	99		[7] [10] [22] [29] [33] [44] [45] [51] [50] [54] [58] [59] [60] [62] [66] [73] [78] [81] [85] [88] [102] [103] [105] [107] [108] [109] [114] [115] [122] [125] [128] [129] [131] [132] [137] [139] [140] [141] [142] [149] [151] [152] [153] [155] [163] [165] [172] [176] [183] [186] [197] [199] [200] [205] [209] [212] [216] [217] [220] [243] [248] [251] [253] [259] [278] [282] [284] [285] [292] [294] [295] [298] [302] [304] [305] [308] [311] [312] [316] [320] [321] [331] [332] [335] [338] [339] [343] [344] [351] [358] [359] [367] [374] [376] [377] [382] [392] [394] [402]
	Code and text in tokens	29		[70] [91] [112] [171] [180] [178] [184] [192] [206] [231] [236] [257] [258] [263] [276] [277] [279] [280] [291] [318] [337] [355] [368] [372] [375] [381] [384] [388] [398]
Tree/Graph-based input	Code in tree structure	5	7	[6] [124] [235] [290] [399]
	Code in graph structure	2		[201] [393]
Pixel-based input	Pixel	1	1	[226]
Hybrid-based input	Hybrid input forms	1	1	[234]

in the dataset, we eliminated duplicate instances by removing any duplicate text samples [357]. This step enhances the diversity of the dataset and helps the model to generalize better to new inputs. “Data tokenization” is a key step in preparing the text for LLMs [199]. Text is labeled into smaller units, such as words or subwords, so that LLMs are easier to manage and process efficiently. Finally, the preprocessed dataset is partitioned into different subsets, usually including a training set, a validation set, and a test set.

4.4 What input formats are the datasets for LLM training converted to?

Once suitable datasets have been carefully chosen and clean data has been achieved through the preprocessing steps, the next critical aspect is the transformation of the data into appropriate formats that can effectively serve as inputs for LLMs. Table 9 shows four distinct data input types that emerged during the research: Token-based input, Tree/Graph-based input, Pixel-based input, and Hybrid-based input.

Token-based input. Token-based input [4, 6, 10, 16] involves representing code and text as sequences of tokens, which are smaller units like words or subwords. Code in tokens refers to the representation of code snippets broken down into meaningful tokens, allowing the LLMs to understand programming language syntax and semantics at a fine-grained level. Text in tokens refers to the tokenization of textual data, such as documentation, bug reports, or requirements, enabling the LLMs to process and analyze natural language descriptions effectively. Code and text in tokens combine both code and its associated textual context, allowing the model to capture the relationships between code elements and their descriptions.

Tree/Graph-based input. Tree-based input [201, 235, 393] represents code as hierarchical tree structures, capturing the syntactic relationships between code elements. Each node in the tree

represents a code element, and the edges represent the hierarchical nesting of control flow statements and other code structures. This form of input allows the LLMs to understand the code's hierarchical structure and perform tasks like code completion and bug fixing. Graph-based input represents code as a graph structure, where nodes represent code elements and edges represent the relationships between them. Unlike trees, graphs allow more flexible and complex relationships between code elements, enabling the model to capture non-linear dependencies in the code. This form of input is used in tasks like code summarization and vulnerability detection by considering the code's intricate relationships.

Pixel-based input. Pixel-based input [226] visualizes code as images, where each pixel represents a code element or token. This visual representation allows the LLMs to process and understand code through image-based learning. In this input form, LLMs learn from the visual patterns and structures in the code to perform tasks like code translation or generating code visualizations.

Hybrid-based input. Hybrid-based input [234] combines multiple modalities to provide LLMs with diverse perspectives for better code comprehension. For example, a hybrid input may combine code in tokens with visual representations of code, allowing the model to learn both from the fine-grained details in the tokenized code and from the overall visual structure of the code. This approach enhances the model's ability to understand complex code patterns and improve performance in tasks such as code comprehension and code generation.

During our investigation of LLM-based models for SE tasks, we observed distinct trends in the usage of different input forms during the training process. **Token-based input forms, namely code in tokens and text in tokens were the most prevalent, collectively constituting approximately 95.52% of the studies⁴.** Specifically, code in tokens was widely adopted in 64 studies, accounting for approximately 31.84% of the total studies, demonstrating its popularity as a primary choice for representing code snippets. This approach allowed LLMs to grasp programming language syntax and semantics effectively, making it suitable for a wide range of code-related tasks. Similarly, text in tokens was utilized in 99 studies, comprising around 49.25% of the total studies. This input form allowed LLMs to process natural language descriptions, bug reports, and documentation with greater efficiency and accuracy. The popularity of token-based input forms underscores their significance in leveraging the power of LLMs for software engineering applications.

In contrast, **tree/graph-based input forms, such as code in tree-structure, were used in only seven studies, making up approximately 3.48% of the total.** Although less prevalent, this input type emerged as a promising choice to represent the hierarchical structure and syntactic relationships within code. Its adoption indicated an ongoing exploration of tree-based representations in specialized tasks, such as code completion and bug fixing.

Pixel-based input and hybrid-based input were relatively less common, each found in one study, contributing approximately 0.5% of the total studies each. While their adoption rates were lower, these input forms presented intriguing possibilities for specific applications. Pixel-based input offered a unique visual representation of code, potentially advantageous for code translation tasks. Meanwhile, hybrid-based input, combining multiple modalities (e.g., code in tree structure and text in tokens in Niu *et al.*'s work [234]), showcased the potential for enhancing code comprehension tasks by offering diverse perspectives for the models to learn from.

In summary, the trends in input form usage reveal a strong preference for token-based input, demonstrating its versatility and effectiveness in various SE tasks. However, ongoing exploration of other input forms, such as tree/graph-based, pixel-based, and hybrid-based, suggests a dynamic and evolving landscape in the application of LLMs for SE, with potential for further innovation and

⁴This refers to studies that explicitly state input forms of LLMs, i.e., a total of 201 papers as shown in Table 9.

improvement in specialized domains. Each of these input forms caters to specific characteristics of the SE tasks being addressed, enabling LLMs to perform effectively across a wide range of code-related applications with a more comprehensive understanding of the input data.

RQ2 - Summary

- (1) We divided the datasets into four categories based on the source of data: open-source, collected, constructed, and industrial datasets. **The use of open-source datasets is the most prevalent**, constituting approximately 59.35% of the 214 papers that explicitly state the dataset.
- (2) We categorized the data types within all datasets into five groups: code-based, text-based, graph-based, software repository-based, and combined. **Text-based and code-based types are the most frequently used in applying LLMs to SE tasks**. This pattern indicates that LLMs are particularly adept at handling text and code-based data in SE tasks, leveraging their natural language processing capabilities.
- (3) We summarized the data preprocessing procedures for different data types and found several common preprocessing procedures, i.e., *data extraction*, *unqualified data deletion*, *duplicated instance deletion*, and *data segmentation*.

5 RQ3: WHAT TECHNIQUES ARE USED TO OPTIMIZE AND EVALUATE LLM4SE?

5.1 What optimizers are used to enhance model performance?

We examined the parameter and learning rate optimizers reported in our selected primary studies. As shown in Fig. 7 (a), fine-tuning emerges as the most widely used optimization algorithm in LLM studies, appearing in 87 research works [6, 51, 59, 128, 234, 308, 321, 329, 398]. This signifies the dominance of fine-tuning in adapting pre-trained models to specific tasks, resulting in enhanced performance across various natural language processing tasks [29, 33, 46]. Hyperparameter optimization is another prominent approach, found in 55 studies [235, 248, 278, 318, 324], highlighting its significance in fine-tuning the hyperparameters of language models to achieve optimal performance on specific tasks [359, 377, 393]. Of the data described above, 35 of these studies [393, 403] used both fine-tuning and hyperparameter parameter optimization algorithms. The limited occurrences of Bayesian [257] and Stochastic Gradient Descent (SGD) [226] optimization (one study each) suggest that they are less frequently employed in LLM research.

Among the learning rate optimization algorithms illustrated in Fig. 7 (b), Adam stands out with 25 occurrences in the studies [42, 45, 102, 115, 151, 152]. Adam is an adaptive optimization algorithm that combines adaptive learning rates with momentum, facilitating faster convergence and reducing the risk of getting stuck in local minima during training [215]. Similarly, AdamW appears in 21 studies [50, 72, 78, 109, 141], demonstrating its significance in improving generalization by adding weight decay regularization to Adam [372]. ZeRO [107, 308] and Adafactor [96, 367] are relatively less explored, mentioned in three and two studies respectively. NVLAMB is found in one study [321], indicating limited exploration of this specific learning rate optimization algorithm in LLM research.

5.2 What prompt engineering techniques are applied to improve the performance of LLMs in SE tasks?

Large-scale pre-trained models have demonstrated effectiveness across numerous code intelligence tasks. These models are initially pre-trained on extensive unlabeled corpora and subsequently fine-tuned on downstream tasks. However, the disparity in input formats between pre-training [3] and downstream tasks [52, 187] poses challenges in fully harnessing the knowledge embedded in pre-trained models. Furthermore, the efficacy of fine-tuning [135] strongly hinges on the volume of downstream data [94, 98, 164], a circumstance frequently characterized by data scarcity [74, 95, 340].

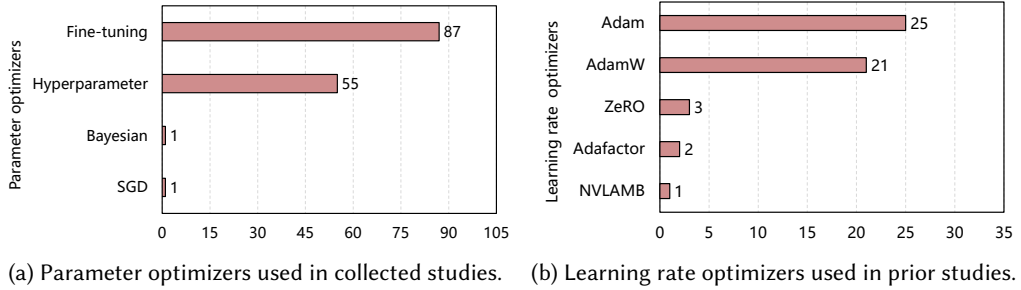


Fig. 7. Various optimizers used in the collected studies.

Recent research in the domain of Natural Language Processing (NLP) underscores that prompt engineering [73, 188], as an emerging fine-tuning [164, 175] paradigm, holds the potential to mitigate the aforementioned issues, yielding commendable outcomes across diverse NLP tasks. In the study conducted by Wang *et al.* [327], they delve into the application of prompt engineering techniques to enhance the performance of LLMs in SE tasks. The research chiefly explores two prompt types: hard prompts [94, 98, 123] and soft prompts [98, 175, 315]. Hard prompts entail manually predefined natural language instructions, while soft prompts, including plain soft prompts, replace natural language tokens in hard prompts with surrogate tokens. A variant known as prefix soft prompts adds a few surrogate tokens before the original input. In the context of prompt engineering, the task-specific knowledge imparted by inserted prompts is particularly advantageous for tasks characterized by data scarcity.

Prompt engineering, characterized by the careful design of specialized prompts, has become a fundamental technique for enhancing interactions with LLMs such as ChatGPT [60, 346] and WizardCoder [200]. These tailored prompts serve dual purposes: they direct the LLMs toward generating specific outputs and also function as an interface to access the extensive knowledge embedded in these models. This becomes particularly important in scenarios where traditional datasets, such as those derived from software repositories or standard benchmarks, are either limited or lack the granularity required for specific tasks. In the context of LLM4SE — which encompasses a range of tasks such as code generation [18, 172, 302, 309], code summarization [4, 81], program repair [117, 250, 354, 372], and test generation [293, 320, 356] — the role of prompt engineering is indispensable. In summary, recent research highlights the critical role of prompt engineering in enhancing the performance of LLMs for targeted SE tasks, thereby contributing to the evolution of automated software development methodologies.

5.3 How are evaluation metrics utilized to assess the performance of LLM4SE tasks?

Evaluating the performance of LLM4SE is a crucial aspect of their development and deployment [137]. Benchmarking against existing datasets and using baselines are common practices to evaluate the effectiveness of LLMs [29]. However, given the diversity of SE tasks, a single evaluation metric may not suffice to capture the model's performance comprehensively. Thus, researchers often employ a range of evaluation metrics tailored to specific problem types [212, 234, 280]. We categorize the SE tasks summarized from 229 papers into four categories according to their addressed problem types, i.e., regression, classification, recommendation, and generation tasks, as displayed in Fig. 8 (b). The selection of evaluation metrics depends on the target problem types. For

Table 10. Evaluation metrics for different types of tasks.

Problem Type	Metric	# Studies	References
Regression	MAE (Mean Absolute Error)	1	[78]
Classification	F1-score	21	[7] [22] [42] [66] [72] [103] [109] [115] [141] [142] [149] [151] [162] [199] [220] [285] [287] [308] [401] [359] [392]
	Precision	20	[22] [42] [61] [66] [72] [103] [109] [115] [141] [142] [149] [151] [162] [220] [285] [287] [308] [310] [359] [392]
	Recall	18	[22] [42] [61] [66] [72] [103] [109] [115] [141] [142] [149] [151] [162] [220] [285] [287] [310] [392]
	Accuracy	12	[42] [88] [127] [135] [141] [142] [149] [162] [163] [206] [276] [310]
	ROC (Receiver Operating Characteristic)	1	[7]
	AUC (Area Under the ROC Curve)	4	[7] [310] [329] [359]
	FPR (False Positive Rate)	4	[42] [298] [308] [329]
	FNR (Falsar Negative Rate)	3	[298] [308] [329]
	MCC (Matthews Correlation Coefficient)	1	[359]
Recommendation	MRR (Mean Reciprocal Rank)	9	[44] [124] [167] [183] [197] [263] [280] [290] [343]
	Precision@k	4	[44] [102] [183] [403]
	F1-score@k	4	[102] [183] [403] [404]
	MAP/MAP@k	3	[44] [122] [183]
	Accuracy	3	[124] [167] [280]
	Recall@k	2	[102] [403]
Generation	BLEU/BLEU-4/BLEU-DC	29	[4] [6] [16] [35] [46] [54] [81] [82] [126] [155] [180] [178] [184] [209] [211] [212] [221] [234] [249] [275] [291] [318] [337] [344] [361] [367] [375] [383] [398]
	Pass@k	28	[27] [29] [33] [38] [54] [59] [60] [62] [67] [128] [129] [156] [168] [176] [192] [200] [236] [291] [307] [335] [344] [371] [376] [377] [375] [389] [388] [398]
	Accuracy/Accuracy@k	16	[73] [118] [125] [131] [137] [156] [185] [211] [212] [234] [253] [267] [284] [291] [309] [381]
	CodeBLEU	12	[16] [81] [126] [129] [184] [249] [291] [337] [344] [375] [383] [398]
	EM (Exact Match)	11	[6] [81] [96] [126] [221] [249] [291] [337] [344] [367] [393]
	ROUGE/ROUGE-L	10	[4] [6] [81] [82] [171] [178] [212] [211] [234] [375]
	METEOR	6	[4] [6] [35] [81] [82] [234]
	Precision	3	[152] [309] [339]
	Recall	3	[152] [309] [339]
	F1-score	3	[152] [309] [339]
	MRR (Mean Reciprocal Rank)	2	[35] [234]
	MFR (Mean First Ranking)	1	[316]
	MAR (Mean Average Ranking)	1	[316]
	ES (Edit Similarity)	1	[189]
	PP (Perplexity)	1	[357]

example, MAE (Mean Absolute Error) has been used for regression tasks [78]. We summarize the most frequently used evaluation metrics for each task type.

For classification tasks, the most commonly used metrics are F1-score [7, 22, 42, 66, 72, 103], Precision [22, 42, 66, 72, 103], and Recall [22, 42, 66, 72, 103, 109], with 21, 20, and 18 studies, respectively, employing these metrics. For example, in the study conducted by Khan *et al.* [141],

F1-score is utilized to evaluate the performance of an automatic bug-fixing model. Similarly, Sharma *et al.* [287] use Precision and Recall to assess the effectiveness of a transformer-based model for code summarization. These metrics are essential for evaluating the model's ability to correctly classify code snippets [72] or identify specific SE properties [42].

For recommendation tasks, MRR (Mean Reciprocal Rank) is the most frequent metric, used in 9 studies [44, 124, 167, 183, 263, 280, 290, 343]. MRR is employed to measure the effectiveness of recommendation systems for code completion, as demonstrated in the study by Ciborowska *et al.* [44]. Precision@k [44, 102, 183, 403] and F1-score@k [102, 183, 403, 404] are also utilized in recommendation tasks, with 4 studies each. These metrics are used to evaluate the precision and F1-score of the recommended code snippets or code completions.

In generation tasks, metrics like BLEU, along with its variants BLEU-4 and BLEU-DC [4, 6, 16, 35, 46], and Pass@k [27, 29, 33, 38, 54, 59] are the most commonly used, appearing in 29 and 28 studies, respectively. For instance, Wang *et al.* [337] employed BLEU to evaluate a code-to-code translation model. Pass@k is used in the research by Jiang *et al.* [129] to assess code generation models, measuring the proportion of generated code snippets that match the reference solutions. Additionally, ROUGE/ROUGE-L [4, 6, 81, 82, 171, 178, 211, 212, 234, 375], METEOR [4, 6, 35, 81, 82, 234], EM (Exact Match) [6, 81, 96, 221, 337, 344, 367, 393], and ES (Edit Similarity) [189] are used in specific studies to evaluate the quality and accuracy of generated code or natural language code descriptions.

RQ3 - Summary

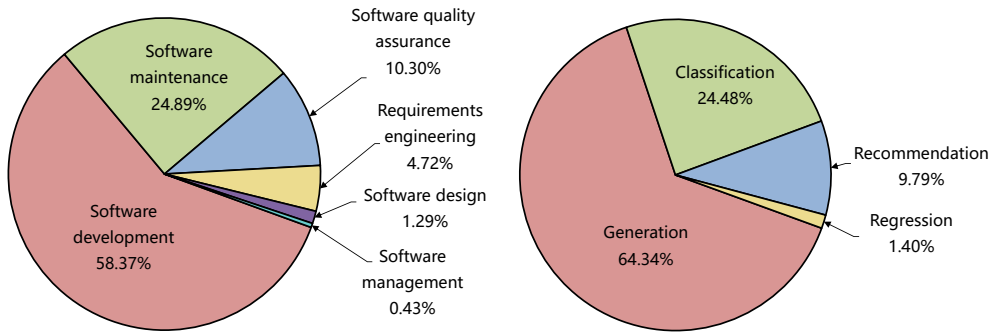
- (1) We conducted an analysis of the parameters and learning rate optimizers commonly employed in LLMs, discovering that fine-tuning and Adam stand out as the most frequently utilized techniques for parameter optimization and learning rate adjustment, respectively.
- (2) We highlighted the application and effectiveness of prompt engineering techniques in improving the performance of LLMs for SE tasks. By exploring various types of prompts, including hard and soft prompts, this emerging fine-tuning paradigm has shown to be particularly advantageous in tasks characterized by data scarcity, providing task-relevant knowledge and enhancing LLMs' versatility and efficacy across different code intelligence tasks.
- (3) We summarized the most widely used evaluation metrics according to four problem types, i.e., regression, classification, recommendation, and generation. Fifteen different evaluation metrics appeared in the generation task, while nine metrics were used for the classification task.

6 RQ4: WHAT SE TASKS HAVE BEEN EFFECTIVELY ADDRESSED TO DATE USING LLM4SE?

6.1 What are the distributions SE activities and problem types addressed to date with LLM4SE?

In this section, we provide a detailed analysis of the use of LLMs in different SE tasks. We summarise reported SE tasks [364] addressed with LLMs, following the six phases of the **Software Development Life Cycle (SDLC)** (i.e., requirements engineering, software design, software development, software quality assurance, software maintenance, and software management). Fig.8 (a) describes the distribution of LLMs in these six activities. Table 11 shows a detailed count of studies reporting specific SE tasks addressed with LLMs.

The highest number of studies is observed in the software development domain, constituting approximately 58.37% of the total research volume. This underscores the primary focus to date on utilizing LLMs to enhance coding and development processes. Software maintenance tasks account for about 24.89% of the research share, highlighting the significance of LLMs in aiding



(a) Distribution of LLM usages in SE activities. (b) Problem classification based on collected studies.

Fig. 8. Distribution of LLM utilization across different SE activities and problem types.

software updates and improvements. The software quality assurance domain holds approximately 10.3% of the research proportion, indicating a growing interest in automating testing procedures. In contrast, requirements engineering and software design activities represent approximately 4.72% and 1.29% of the research share, respectively, suggesting relatively limited exploration so far in these areas. The software management domain has the least research representation, accounting for a tiny 0.43% proportion. This distribution underscores the vital focus on development and maintenance tasks while also indicating potential avenues for further research in testing, design, and management domains.

In our collection of LLM studies for SE tasks, we've classified them based on the type of problems they address (shown in Fig. 8 (b)). The distribution reveals that **the majority of studies, about 64.34%, center around generation tasks**, showcasing the significance of LLMs in producing code or text. Following this, around 24.48% of studies fall under classification tasks, indicating the relevance of LLMs in categorizing software elements. Additionally, roughly 9.79% of studies are related to recommendation tasks, demonstrating the utility of LLMs in suggesting solutions. Lastly, a smaller portion, around 1.4%, is allocated to regression tasks, reflecting the limited exploration of LLMs for predictive modeling. **This distribution underscores the broad applicability of LLMs across different SE challenges, with a notable emphasis on code generation and classification tasks.**

6.2 How are LLMs used in requirements engineering?

This section explores the utilization of LLMs in the domain of requirements engineering. It encompasses tasks such as anaphoric ambiguity treatment, requirements classification, coreference detection, requirements elicitation, and software traceability.

Anaphoric ambiguity treatment. Ambiguity in software requirements arises when a single reader can interpret a natural language (NL) requirement in multiple ways, or different readers have varying understandings of the same requirement. Unclear and ambiguous NL software requirements can lead to suboptimal software artifacts during later development stages. Moharil *et al.* [217] and Ezzini *et al.* [66] have empirically demonstrated the significant role of LLMs such as BERT and SpanBERT in effectively addressing anaphoric ambiguity. Sridhara *et al.* [296] revealed that ChatGPT excels in addressing anaphoric ambiguity in software requirements. Through researchers' analysis of ten English requirement specifications [66] containing anaphora-related challenges,

Table 11. Distribution of SE tasks over six SE activities.

SE Activity	SE Task		Total
Requirements engineering	Anaphoric ambiguity treatment (3)	Requirements term identification (1)	11
	Requirements classification (3)	Coreference detection (1)	
	Requirement analysis and evaluation (2)	Traceability automation (1)	
Software design	GUI retrieval (1)	Software specification synthesis (1)	3
	Rapid prototyping (1)		
Software development	Code generation (62)	Agile story point estimation (1)	136
	Code completion (16)	API documentation smell detection (1)	
	Code summarization (10)	API entity and relation extraction (1)	
	Code understanding (7)	Code optimization (1)	
	Code search (5)	Code example recommendation (1)	
	Program synthesis (5)	Control flow graph generation (1)	
	API recommendation (2)	Data analysis (1)	
	API synthesis (2)	Identifier normalization (1)	
	Code comment generation (2)	Instruction generation (1)	
	Code representation (2)	Type inference (1)	
	Method name generation (2)	Others (11)	
Software quality assurance	Test generation (8)	Bug localization (1)	24
	Vulnerability detection (7)	Failure-inducing test identification (1)	
	Test automation (4)	Flaky test prediction (1)	
	Verification (2)		
Software maintenance	Program repair (23)	Duplicate bug report detection (1)	58
	Code review (6)	Decompilation (1)	
	Debugging (4)	Program merge conflicts repair (1)	
	Bug report analysis (3)	Sentiment analysis (1)	
	Code clone detection (3)	Tag recommendation (1)	
	Logging (2)	Vulnerability repair (1)	
	Bug prediction (1)	Commit classification (1)	
	Bug triage (1)	Traceability recovery (1)	
	Bug report replay (1)	Others (6)	
Software management	Effort estimation (1)		1

ChatGPT consistently demonstrated its remarkable capability to accurately identify antecedents. This empirical evidence emphasizes the valuable role ChatGPT can play in enhancing the clarity and precision of software requirements, ultimately contributing to more effective software development processes by reducing interpretational uncertainties.

Requirements classification. Originating in NL documents, requirements demand effective classification, especially for early-stage project discernment, like security-related ones [147, 166]. Automated processing hinges on identifying these requisites. Categorizing into functional (FR) or non-functional (NFR) requirements, with quality constraints, benefits automated approaches [166]. Hey *et al.*[109] employ BERT for requirement classification, where it excels in categorizing both FR and NFR requirements using a fine-tuning transfer learning technique, outstripping traditional methods. Luo *et al.*[199] introduce a BERT-based software requirement classification method, demonstrating remarkable transferability and generalization, especially in zero-shot scenarios.

Requirements term identification. Moharil *et al.* [216] propose a technique for identifying terms used in different contexts within the same domain or in interdisciplinary projects. Using BERT, which reads entire word sequences for deeper language understanding, and K-means clustering, they create and group vectors for each term in the corpora. The method has been validated on large Computer Science and multi-domain corpora comprising eight different fields.

Coreference detection. Requirements, authored by diverse stakeholders, continually evolve, leading to terminology differences and inconsistencies across domains. Entity coreference in Requirement Engineering (RE), where various expressions refer to the same real-world entity, can

cause confusion and affect comprehensibility. Wang *et al.* [338] offer a novel application of the BERT model for coreference detection.

Traceability automation. Software and system traceability refers to the ability to establish and maintain relationships between software artifacts, such as requirements, design definitions, code, and test cases, for product querying and development support [269]. Lin *et al.* [183] found that T-BERT can effectively migrate knowledge from code search to NLA-PLA (i.e., Natural Language Artifacts to Programming Language Artifacts) traceability, even with limited training instances. It outperforms existing techniques in accuracy and can be adapted to different domains without intermediate training for each project, offering a promising step toward practical, trustworthy traceability.

6.3 How are LLMs used in software design?

GUI (Graphical User Interface) retrieval. Kolthoff *et al.* [151] present the application of BERT in the task of GUI retrieval in SE. The authors fine-tune a BERT-based learning-to-rank (LTR) model for this task. GUIs, which are not standard well-structured text documents, present unique challenges for text-based ranking tasks. The BERT model is prepared by concatenating the natural language query and the GUI document text, and then this input is used to train different BERT-LTR models. The models are evaluated based on their performance in NL-based GUI ranking.

Rapid prototyping. Rapid prototyping enables developers to quickly visualize and iterate on software designs, thereby accelerating the development process and ensuring alignment with user needs. White *et al.* [346] investigate the role of LLMs in augmenting this process. The study introduces prompt design techniques, organized into patterns, providing a structured methodology to tackle prevalent challenges in LLM4SE. This research indicates that the realm of rapid prototyping stands to benefit from deeper integration with advanced machine learning techniques, thereby creating opportunities for additional research and refinement aimed at producing more intuitive and user-centric software designs.

Software specification synthesis. Software configuration is vital for system behavior, but managing configurations and specifications becomes complex with larger systems. Mandal *et al.* [205] introduce SpecSyn, a framework using an LLM for automatic software specification synthesis from natural language sources. This end-to-end approach treats the task as a sequence-to-sequence learning problem, surpassing the previous state-of-the-art tool by 21% in F1 score, and can find specifications from both single and multiple sentences.

6.4 How are LLMs used in software development?

Our analysis identifies wide-ranging applications of LLMs for software development, encompassing tasks such as code generation, code completion, and code summarization.

Code generation. Code generation has long been a task of interest: there is extensive work on program synthesis using symbolic and neural-semiotic approaches [11, 350]. Recently, LLMs trained for text generation have demonstrated the ability to complete programs [23, 25]. Since 2020, several code generation models have been trained or fine-tuned on programming language text [38, 47, 74, 77, 233, 357]. Unlike traditional program synthesis techniques, neurolinguistic models can be conditioned on natural language (e.g., code annotations) as well as generate programming language text. Researchers have experimentally demonstrated that LLMs like GPT-4 [18, 85, 128, 186], GPT-2/GPT-3/GPT-3.5 [17, 60, 139, 168, 184, 186, 225, 329, 368], BERT series [157, 379], Codex [18, 38, 54, 96, 155, 204, 371], CodeGen [54, 132, 376], InCoder [153, 186, 221, 332], Copilot [350] and CodeGeeX [398], play a key role in code generation. By pre-training on large-scale text data, these models learn rich linguistic knowledge and semantic representations that enable them to understand the meaning and structure of natural language. LLMs can automate code generation by

Table 12. The state-of-the-art applications of LLMs in code generation task.

Model	Baseline	Benchmark	Metric	Date	Reference
GPT-3.5	Codex, CodeGen, CodeGeeX, LLaMA, InCoder, PyCodeGPT, CodeParrot, GPT-2	HumanEval, MBPP, MBCPP	Pass@k	May 11, 2023	[168]
GPT-4	PaLM Coder, Codex, CodeGen-Mono, InCoder, CodeGeeX, AlphaCode	HumanEval, HumanEval-ET, MBPP, MBPP-ET	Pass@k	May 24, 2023	[60]
GPT-4	GPT-3.5, StarCoder, CodeGen, CodeGen2, Vicuna, SantaCoder, InCoder, GPT-J, GPT-Neo, PolyCoder, StableLM	HumanEval, HumanEval+, HumanEval-mini	Pass@k	Jun 12, 2023	[186]
GPT-4	GPT-3.5, WizardCoder, Instruct-StarCoder, SantaCoder, Instruct-CodeGen, CodeGeeX, InCoder, Vicuna, ChatGLM, PolyCoder	ClassEval, HumanEval	Pass@k	Aug 3, 2023	[62]

converting natural language descriptions into code [129]. These models generate program code from natural language descriptions, enhancing code-writing efficiency and accuracy. They show excellent performance in code completion, automatic code generation, and conversion of natural language annotations to code, providing software developers with powerful auxiliary tools and promoting further automation and intelligence in the code writing and development process.

Within the domain of LLMs applied to software development tasks, studies centered on code generation distinctly dominate the academic landscape. As reflected in Table 12, **the GPT series, particularly GPT-4, emerge as a key focus, with many more studies using them in the realm of code generation** [60, 62, 168, 186]. Analysing these studies, several noteworthy findings surface:

- **Programming thinking in LLMs.** Techniques that evoke “programming thinking” within LLMs, such as the TIP (i.e., Thinking in Programming) [168] methodology, have shown promising strides. By guiding LLMs to first craft a high-level code sketch before delving into detailed implementations, the synthesized code exhibits higher accuracy and robustness.
- **Class-level vs. Method-level generation.** LLMs, while adept at method-level code generation, present varied performance metrics when tasked with class-level generation [62]. This divergence underscores the evolving nature of challenges as the granularity of code synthesis shifts.
- **Expanding LLM capabilities.** The next frontier in this discipline seems to lie in harmoniously integrating LLMs with established SE tools and practices. The emergence of frameworks like EvalPlus [60] indicates a trend towards enhancing the evaluation and accuracy of LLM-generated code, possibly ushering in an era where human developers and LLMs collaboratively craft software solutions.

Code completion. Code completion is an assistive feature provided by many integrated development environments (IDEs) and code editors. Its purpose is to automatically display possible code suggestions or options as developers write code [12]. This innovation has been advanced by Language Models (LMs), evolving from n-gram and RNN models to transformer-based models like Copilot [87] and CodeGPT [134], pre-trained on extensive code datasets. Recent LLMs equipped with billions of parameters, excel in generating code snippets. These models are trained on vast amounts of natural language text, equipping them with powerful semantic understanding capabilities. In the context of code completion, LLMs such as Codex [38, 59, 181, 251], BERT series [142], Github Copilot [59, 181, 258], CodeParrot [181, 357], GPT series [235, 357], T5 [46], InCoder [77], PolyCoder [357], CodeGen [56, 58, 181, 232], and other LLMs [124, 235], can generate accurate and

intelligent code suggestions based on code context and syntax structures. They comprehend the developer's intent, predict the next possible code snippet, and provide appropriate recommendations based on the context.

With the support of LLMs, code completion achieves significant improvements in efficiency and accuracy. Developers can save time by avoiding manual input of lengthy code and reducing the risk of code errors. LLMs also learn from extensive code repositories, acquiring knowledge and best practices to offer more intelligent and precise suggestions, aiding developers in better understanding and utilizing code [46]. Additionally, these models can provide personalized code recommendations based on developers' coding styles and preferences, further enhancing the effectiveness and user experience of code completion [189].

Code summarization. Code summarization is a task that attempts to understand the code and automatically generate descriptions directly from the source code. It can also be viewed as an extended form of documentation. Successful code summarization not only facilitates the maintenance of source code [123, 228] but can also be used to improve the performance of code search using natural language queries [230, 360] and code classification [228]. LLMs play a significant role in code summarization by analyzing code structures and contexts to generate informative natural language summaries. Specifically, LLMs such as Codex [4, 16, 81], CodeBERT [35, 81, 91], and T5 [211, 212] comprehend the functionality and logic of the code, producing easily understandable human language descriptions. For example, Arakelyan *et al.* [16] rigorously evaluate the efficacy of CodeT5 and Codex across code generation and summarization tasks, shedding light on their performance under distribution shifts. It unveils practical adaptation techniques, underscoring Codex's commendable performance. Additionally, the study demonstrates that while adapted models exhibit proficiency in code generation, their generality can present trade-offs in the context of code summarization. As a result, code summarization with the support of LLMs enhances code readability, improves software documentation quality, and accelerates code comprehension and collaboration among developers. This advanced approach to code summarization demonstrates great potential for automating and streamlining various aspects of software development in modern SE practices with the employment of LLMs.

Code understanding. Code Understanding refers to the process of deeply comprehending and analyzing source code. It involves gaining insights into the logic, structure, functionality, and dependencies of the code [288], as well as understanding the programming languages, frameworks, and libraries used. LLMs can assist in code understanding by leveraging their powerful natural language processing capabilities to interpret code-related text, such as comments and documentation [135, 337]. They aid developers in grasping code functionality, identifying dependencies, and generating relevant code documentation [201, 288]. Through their ability to comprehend both code and natural language, LLMs enhance the efficiency and accuracy of code understanding, empowering developers to maintain, optimize, and integrate code effectively [135].

Code search. Code search, or code retrieval, is the task of retrieving source code from a large code base, usually based on a user's natural language query. Despite the success of neural models in code search, such models are relatively shallow and are not capable of learning large amounts of data [280]. In recent years, some bimodal pre-training models based on the BERT neural architecture have been proposed to capture semantic links between natural and programming languages [74, 95, 274, 336], such as CodeBERT [74] and GraphCodeBERT [95]. Bimodal pre-training models learn generic representations from large amounts of data in an unsupervised manner by designing pre-training goals. Salza *et al.* [280] explored the effectiveness of LLMs such as BERT [280] and RoBERTa [35] in understanding natural language and code semantics and enhancing code search and retrieval. These studies show that pre-training tasks alone may not be sufficient for code search, which emphasizes the need for a multimodal understanding of data [290], including both natural language and code.

In addition, research has shown that the use of code generation models such as Codex [165] can enhance code retrieval by generating code snippets from natural language documents, thereby improving semantic similarity and obtaining state-of-the-art results on benchmark datasets.

Program synthesis. Program synthesis is the automated process of generating code that satisfies a given specification or set of constraints, emphasizing the derivation of functional properties of the code [40, 41, 207, 247, 297]. It differs from code generation, which primarily translates higher-level representations into target code without necessarily deriving its functionality from scratch [292, 388, 398]. Several studies have demonstrated that LLMs can be used for program synthesis tasks. LLMs have a significant impact on program synthesis due to their advanced language understanding and generation capabilities. LLMs can effectively interpret natural language descriptions, code comments, and requirements, and then generate corresponding code snippets that fulfill the given specifications. This helps developers in rapidly prototyping code and automating repetitive coding tasks [79, 155]. When applied to program synthesis, LLMs enhance productivity and reduce the burden on developers by automating the code-writing process based on high-level input [125]. Their ability to understand the nuances of both natural language and programming languages makes them valuable tools in advancing the field of SE and streamlining the development lifecycle.

API recommendation. Several methods have been proposed to automate API (Application Programming Interface) recommendations [93, 116, 190, 227], falling into two orthogonal approaches: information retrieval-based (IR-based) and neural-based. In this context, our focus is on the latter. Wei *et al.* [343] introduced CLEAR, an API recommendation method that employs the BERT sentence embedding model to represent queries, capturing continuous semantic information. Through contrast training, CLEAR enables BERT to learn precise semantic representations of queries, independent of their lexical content. Recently, Zhang *et al.* [387] developed ToolCoder, which combines API search tools with existing models to aid in code generation and API selection. This approach involves an automated data annotation method using ChatGPT, adding tool usage information to the source code data, followed by fine-tuning the code generation model. During inference, an API search tool is integrated into the generation process, allowing the model to automatically utilize the tool for suggestions when selecting APIs.

API synthesis. The automated generation of application programming interface calls, known as API synthesis, plays a crucial role in bridging human intent with machine execution. In recent studies, Wang *et al.* [331] and Patil *et al.* [248] have both explored the potential of LLMs in this realm. Utilizing models like GPT-4 and LLaMA-based architectures, these researchers showcase the prowess of LLMs in generating accurate API calls and adapting to real-time documentation changes, effectively addressing challenges like hallucination and inaccurate input arguments. The integration of LLMs in API synthesis signifies a paradigm shift, promising enhanced accuracy, adaptability, and reliability in code generation. As illuminated by these studies, the future of API synthesis may be deeply anchored in advanced machine learning, heralding new research avenues and refinements for more seamless human-machine interactions.

Code comment generation. Code comment generation, the automatic creation of comments for source code, serves to elucidate code functionality, implementation logic, and input-output details, thereby enhancing readability and maintainability [82]. As code complexity grows, manually crafting these comprehensive and accurate comments can become burdensome and prone to errors. Automation in this domain can markedly enhance the efficiency and quality of code documentation. LLMs such as Codex [82] and T5 [209] have been effectively applied to code comment generation. These models are pre-trained on vast amounts of data and possess powerful natural language processing and semantic understanding capabilities. During comment generation, LLMs analyze the structure, semantics, and context of the source code to automatically generate high-quality

comments that correspond to the code's functionality and logic. Addressing the often observed disconnect between code evolution and its accompanying documentation, Mastropaolo *et al.* [209] explore the potential of LLMs, particularly the T5 architecture, in assisting developers with code comment completion. Their empirical study juxtaposes the performance of the T5 model against an n-gram model, revealing T5's superior capabilities, though the n-gram model remains a competitive alternative. The research underscores the significance of open-source datasets for training and highlights the scant use of industrial datasets in current studies.

Code representation. Code representation learning (also known as code embedding) aims to encode the code semantics into distributed vector representations and plays a key role in recent deep-learning-based models for code intelligence. Code representation can be used to support a variety of downstream tasks, such as code completion [268], code search [92, 322], and code summarization [325, 385]. Niu *et al.* [234] propose a novel sequence-to-sequence pre-training model that utilizes structural information from source code to enhance its representation learning. The model is trained on a large corpus of source code, which enables it to capture the complex patterns and dependencies inherent in programming languages. Wan *et al.* [324] show through their research that attention is highly consistent with the syntactic structure of the code, that pre-trained code language models can preserve the syntactic structure of the code in the intermediate representations of each converter layer, and that pre-trained code models have the ability to induce a syntactic tree of the code. These revelations suggest that incorporating the syntactic structure of the code into the pre-training process results in better code representations.

Method name generation. Method names significantly affect program comprehensibility, serving as a brief summary of the source code and indicating the developer's intent [148]. The importance of method names in program comprehension is further evidenced by recent studies showing that some programmers even write down important method names to help them figure out the procedures of an application [273]. Zhu *et al.* [403] present AUMENA, a novel approach using the CodeT5 model for context-aware method naming in SE. AUMENA first learns the contextualized representation of programming and natural language, then leverages LLMs with prompt tuning to detect inconsistent method names and suggest accurate alternatives. This method avoids previous generate-then-compare consistency checking limitations, modeling the task as a two-class classification problem.

Agile story point estimation. Agile story point estimation, representing the total work needed to implement a product backlog item, is a complex task in agility. Story points are typically estimated by team consensus, using methods like plan poker and expert judgment, and considering factors like workload and complexity. However, subjective estimates may introduce uncertainty. Fu *et al.* [78] present GPT2SP, a Transformer-based approach that overcomes limitations of a previous method called Deep-SE. Unlike Deep-SE, which restricts language models to known words within a trained project, GPT2SP employs a broader context, making it transferable across projects. GPT2SP's performance is comparable to Deep-SE in within-repository evaluations and surpasses it in 62.5% of cases, with improvements ranging from 3% to 46% across various projects.

API documentation smell detection. APIs, vital for modern software development, are often accompanied by official documentation. Good documentation is key to proper API use, while poor quality can hinder adoption and negatively impact developers' productivity [1, 271, 272]. Khan *et al.* [141] identified five API documentation smells and presented a benchmark of 1,000 API documentation units containing the five smells found in the official API documentation. The authors developed classifiers to detect these odors, with BERT showing the best performance, demonstrating the potential of LLMs in automatically monitoring and warning about API documentation quality.

API entity and relation extraction. Extracting APIs and their semantic relationships from unstructured text (e.g., data from Stack Overflow) is a fundamental task in SE, but existing methods

require labor-intensive manual rule creation or data labeling. Huang *et al.* [115] present an innovative approach, AERJE, that leverages LLMs for this task. AERJE consists of a BERT-based dynamic hint generator and a T5-based joint entity-relationship extractor, which together enable efficient extraction of API entities and relationships without manual effort. The approach achieved an F1 score of 96.51% for API entity extraction and 81.2% for API relationship extraction, offering a significant advancement over traditional methods.

Code optimization. Efficiency in programming is vital, particularly in resource-limited or large-scale applications. Traditional optimizing compilers enhance efficiency through various considerations like algorithm and data structure selection [5]. Madaan *et al.* [203] explore the use of LLMs in suggesting performance-enhancing code edits. They curate a dataset of Performance-Improving Edits (PIE), showing how Codex and CodeGen can generate these edits, resulting in over 2.5x speedups for more than 25% of the C++ and Python programs, even after C++ code was compiled using the O3 optimization level.

Code example recommendation. Zhou *et al.* [400] pointed out that software developers tend to write similar code examples several times due to the need to implement similar features in different projects. Therefore, during the software development process, recommender systems can provide programmers with the most pertinent and high-quality examples written by other programmers, thus helping them to complete their tasks quickly and efficiently [53]. Open-source projects and informal documentation are the two main sources of information that developers rely on to perform programming tasks. For example, open-source projects on GitHub provide code examples and code resources for various tasks. Rahmani *et al.* [263] introduce a methodology to improve code example recommendations for Java programming language on Stack Overflow using BERT and Query-Aware Locality-Sensitive Hashing (LSH). They employ BERT to convert code into numerical vectors and then apply two LSH variants, Random Hyperplane-based, and Query-Aware, to identify Approximate Nearest Neighbors (ANN).

Control flow graph generation. Control Flow Graphs (CFGs) are a cornerstone of SE that illustrate program behavior by showing sequences of statements and their execution order conditions [8]. As a graphical representation of program behavior, CFGs are critical in many SE tasks, including code search [37, 95], code clone detection [113, 333, 341] and code classification [334, 386]. Huang *et al.* [118] presented a novel approach for generating behaviorally correct CFGs of statically typed partial code by leveraging the error-tolerant and understanding ability of LLMs. The approach involves a Chain of Thoughts (CoT) with four steps: structure hierarchy extraction, nested code block extraction, CFG generation of nested code blocks, and fusion of all nested code blocks' CFGs [161]. The CoT is broken down into an AI chain according to the single responsibility principle, along with effective prompt instructions. This results in superior node and edge coverage compared to traditional program analysis-based methods and the original CoT method.

Identifier normalization. Identifiers usually consist of multiple words, and a certain number of identifiers contain abbreviations [130]. Consequently, the lexical meaning of identifiers and the overall functionality of source code written by one developer may be challenging for other developers to comprehend. In addition, the source code cannot match the vocabulary in other software artifacts described in natural language, thus invalidating some automated algorithms. Therefore, there is a strong need to normalize identifiers with the aim of aligning the vocabulary in identifiers with the natural language vocabulary in other software artifacts. Zhang *et al.* [381] addressed this by introducing BEQAIN, an approach for identifier normalization. BEQAIN combines BERT with a Question and Answering (Q&A) system and Conditional Random Fields (CRF), treating identifier splitting as sequence labeling and abbreviation expansion as a Q&A task. It uses programming context to refine expansion results when multiple expansions are possible, aligning

identifier vocabulary with natural language and enhancing software development comprehension and automation.

Type inference. Type inference, the automated process of determining data types in programming, plays a crucial role in enhancing readability, maintainability, and reducing runtime errors [104, 256]. TypeScript, with its unique blend of optional typing, presents a nuanced challenge, especially when navigating the vast landscape of user-defined types. Addressing this complexity, Jesse *et al.* [127] introduced an approach that leverages the capabilities of a BERT-style pre-trained model. Their solution, DIVERSETYPER, adeptly infers types for user-defined classes and interfaces by uniquely correlating class and interface declarations with their respective usage contexts. Beyond merely filling the gaps of previous methodologies, DIVERSETYPER sets a new benchmark in type inference, especially for user-defined types.

6.5 How are LLMs used in software quality assurance?

Within the domain of software quality assurance, LLMs have emerged as valuable tools with diverse applications for various tasks, including vulnerability detection, test generation, bug localization, etc.

Test generation. Test generation involves automating the process of creating test cases to evaluate the correctness and functionality of software applications. It encompasses various aspects, including test case generation [389], unit test generation [283, 293, 304, 356, 374], etc. LLM application in test generation offers several advantages, including the ability to automatically generate diverse test cases, improving test coverage [283, 293] and identifying potential defects [356]. LLMs can also assist in generating test cases based on natural language descriptions, fostering better collaboration between developers and testers. Additionally, they help identify areas lacking test coverage and suggest relevant test cases, ensuring comprehensive testing and reducing the risk of undiscovered issues [389]. By enhancing test efficiency and effectiveness, LLMs contribute to producing more reliable and high-quality software products.

Vulnerability detection. The number of software vulnerabilities is rapidly increasing, as shown by the vulnerability reports from Common Vulnerabilities and Exposures (CVEs) [15] in recent years. As the number of vulnerabilities increases, there will be more possibilities for cybersecurity attacks, which can cause serious economic and social harm. Therefore, vulnerability detection is crucial to ensure the security of software systems and protect social and economic stability. Traditional static detection methods are based on static analysis and predefined matching rules, which rely on developers' expertise and make it difficult to detect unknown vulnerabilities. With the assistance of LLMs [30, 42, 308], Alqarni *et al.* [10] present an updated BERT model fine-tuned for vulnerability detection. Additionally, Tang *et al.* [303] introduced novel approaches using LLMs to enhance vulnerability detection. One of their proposed models, CSGVD, combines sequence and graph embedding for function-level vulnerability detection, outperforming other deep learning-based models on a real-world benchmark dataset. Their study also explores the application of CodeT5 for vulnerability detection, highlighting the importance of code-specific pre-training tasks.

Test automation. Automated testing methodologies offer a comprehensive array of tools and strategies designed for the evaluation of software applications' accuracy, reliability, and performance. These methodologies encompass various techniques, such as mutation testing [144] and fuzzing [50, 51]. LLMs have been used for mutation testing, introducing faults to the codebase to assess the effectiveness of test suites in identifying and detecting errors [144]. Furthermore, LLMs can aid in fuzzing, generating valid and diverse input programs that help identify vulnerabilities and bugs, particularly in challenging domains like deep learning libraries [50]. By incorporating LLMs into test techniques, software engineers benefit from improved test coverage, reduced manual effort, and enhanced bug detection [51], leading to more robust and reliable software systems.

Verification. Verification techniques, including prominent methods such as formal verification, hold a pivotal role in the domain of software quality assurance [32, 312]. These techniques validate the correctness of software systems, improving their reliability and security against potential threats. Utilizing mathematical and logical principles in the verification process facilitates thorough error detection and correction before deployment, ensuring stable and secure performance in different operational contexts. Charalambous *et al.* [32] leverage LLMs, particularly the GPT-3.5, in the realm of formal verification. Their approach combines LLMs with bounded model checking (BMC) to automatically repair software based on formal methods, showcasing the model's capability to understand intricate software structures and generate accurate repairs.

Bug localization. Bug localization refers to the process of identifying the specific source code files, functions, or lines of code that are responsible for a reported bug or software defect. Bug localization typically involves analyzing bug reports or issue descriptions provided by users or testers and correlating them with the relevant portions of the source code. This process can be challenging, especially in large and complex software projects, where codebases can contain thousands or even millions of lines of code. Traditional bug localization methods often rely on heuristics, code metrics, or stack trace analysis, which may not always provide precise results. Ciborowska *et al.* [45] investigated data augmentation techniques to enhance bug localization models. They introduce a pipeline applying token-level operations such as dictionary replacement, insertion, random swapping, and deletion, along with paragraph-level back-translation to bug reports. By employing augmented data to train BERT-based models for bug localization, they demonstrate that these techniques can substantially expand the training data and boost the models' performance.

Failure-inducing test identification. Test suites typically include two types of test cases: pass-through test cases and fault-inducing test cases [173]. In practice, there are far more pass test cases for faults than fault-inducing test cases, which hinders the effectiveness of program debugging. However, in practice, it is difficult to find fault-inducing test cases. This is because developers first need to find test inputs that trigger program faults, and the search space for such test inputs is huge [76]. Moreover, developers need to build a test oracle to automatically detect program faults, and building a test oracle is often an undecidable problem [120]. Li *et al.* [173] investigated the application of ChatGPT to the task of finding fault-inducing test cases in SE. While recognizing ChatGPT's potential, they initially observed suboptimal performance in pinpointing these cases, particularly when two versions of a program had similar syntax. The authors identified this as a weakness in ChatGPT's ability to discern subtle code differences. To enhance its performance, they devised a novel approach blending ChatGPT with difference testing. Leveraging ChatGPT's strength in inferring expected behavior from erroneous programs, they synthesized programs that amplified subtle code differences. The experimental results reveal that this approach greatly increases the probability of finding the correct fault-inducing test case.

Flaky test prediction. In many environments, it has been found that test cases can be non-deterministic, with test cases passing and failing in different executions, even for the same version of the source code. These test cases are called piecewise test cases [63, 72, 198, 405]. Fatima *et al.* [72] propose a black-box approach named Flakify that uses CodeBERT to predict flaky tests. The model is trained on a dataset of test cases labeled as flaky or non-flaky. The model's predictions can help developers focus their debugging efforts on a subset of test cases that are most likely to be flaky, thereby reducing the cost of debugging in terms of both human effort and execution time.

6.6 How are LLMs used in software maintenance?

Within the context of software maintenance, LLMs have been leveraged for bug prediction, program repair, code review, debugging, and an array of other activities.

Table 13. The state-of-the-art applications of LLMs in program repair task.

Model	Baseline	Benchmark	Metric	Date	Reference
Codex	GPT-Neo, GPT-J, GPT-NeoX, CodeT5, InCoder	QuixBugs-Python and Java, Defects4J 1.2 and 2.0, ManyBugs	Correct / plausible patches	May 20, 2023	[353]
Codex	CodeT5, CodeGen, PLBART, InCoder	Vul4J, VJBench,	Correct / plausible patches	May 29, 2023	[351]
ChatGPT	Codex, CodeGen-16B, CodeGen-6B, CodeGen-2B, CodeGen-350M	QuixBugs-Python and Java	Correct / plausible patches	Jan 30, 2023	[354]
ChatGPT	Codex, CodeBERT, SelfAPR, RewardRepair, Recoder, TBar, CURE, CoCoNuT	QuixBugs-Python and Java, Defects4J 1.2 and 2.0	Correct fixes	Apr 1, 2023	[355]

Program repair. The goal of automated program repair (APR) is to automatically identify and fix bugs or defects in software [393]. It involves leveraging automated techniques to analyze buggy code and generate correct patches to address the identified issues. LLMs, such as BERT [310, 390], CodeBERT [162], CodeT5 [249], Codex [71, 131, 351], PLBART [249, 351], T5 [211, 372] and GPT series [28, 32, 158, 295, 311, 354, 355], have shown effectiveness in generating syntactically correct and contextually relevant code. Leveraging LLMs for program repair can achieve competitive performance in generating patches for various types of bugs and defects [355]. These models can effectively capture the underlying semantics and dependencies in the code [32], leading to the production of accurate and effective patches [354, 390]. Moreover, LLMs can be fine-tuned on specific code repair datasets [211], further improving their ability to generate high-quality patches for real-world software projects. The application of LLMs in program repair not only accelerates the bug-fixing process but also enables software developers to focus on more complex tasks, leading to enhanced software reliability and maintainability.

In recent research, program repair has emerged as a prevalent application. Among the LLMs, as shown in Table 13, Codex [351, 353] and ChatGPT [354] have particularly distinguished themselves in the program repair domain. **ChatGPT edges ahead due to its inherent interactive design, enabling a continuous feedback loop that yields refined and contextually apt patches** [354, 355]. Such conversational dynamics, coupled with rigorous comparisons across diverse baselines, underscore its superior adaptability and efficiency.

Summarising several key findings from research on LLMs for program repair:

- **Interactive feedback.** Incorporating an interactive feedback loop, as observed with ChatGPT, significantly augments the accuracy of program repair [354]. This dynamic interplay between patch generation and validation fosters a deeper understanding of the software’s semantics, leading to more effective repairs.
- **Domain-specific integration.** Merging the capabilities of LLMs with domain-specific knowledge and techniques further enhances their performance. Customized prompts, project-specific fine-tuning, and leveraging SE techniques [328, 353] can dramatically elevate the efficacy of LLM-driven program repairs.
- **Comparative analysis.** Rigorous evaluation against diverse baselines reveals the versatility and adaptability of LLMs, especially ChatGPT. This wide-ranging comparison not only establishes their superiority but also underscores areas for potential improvement [355].

Code review. Code review is a critical quality assurance practice used to inspect, assess, and validate the quality and consistency of software code [285]. Code review aims to identify potential errors, vulnerabilities, and code quality issues, while also improving code maintainability, readability, and scalability. LLMs like BERT [285], ChatGPT [296], and T5 [171, 318], trained on massive code

repositories, possess the ability to understand and learn the semantics, structures, and contextual information of code [384]. In the code review process, LLMs assist reviewers in comprehensively understanding code intent and implementation details, enabling more accurate detection of potential issues and errors. Moreover, these models can generate suggestions for code improvements and optimizations, providing valuable insights and guidance to reviewers. By combining the intelligence of LLMs with the expertise of human reviewers, code review becomes more efficient and precise, further enhancing software quality and reliability.

Debugging. Debugging targets identifying, locating, and resolving software defects or errors, commonly known as bugs. The debugging process involves scrutinizing the code, tracing the execution flow, and isolating the root cause of the problem to effectively correct the error. LLMs, such as BERT and other converter-based architectures, excel at utilizing contextual information and natural language understanding. In terms of debugging, LLMs can be used to simulate the scientific debugging process, such as AutoSD proposed by Kang *et al.* [136]. This model generates hypotheses about code problems and extracts relevant values to identify potential problems. In addition, the SELF-DEBUGGING method proposed by Chen *et al.* [39] enables LLM to debug its own generated code by learning a small number of presentations and explanations, which effectively improves the accuracy and sampling efficiency of code generation. Using LLMs in debugging not only improves fixing performance by generating competitive fixes but also provides insights into and explanations of the model's decision-making process, making it an important tool for improving software quality and developer productivity.

Bug report analysis. LLMs such as Codex [137] and BERT [44] comprehensively analyze natural language text, code snippets, and contextual information within bug reports to generate precise code repair suggestions, test cases, or steps for reproducing errors. By deeply understanding the semantics and context of the issues, LLMs offer developers more intelligent solutions, expediting the error-fixing process and alleviating development burdens [88, 163]. These models excel not only in code generation but also in identifying and interpreting crucial information within error reports, aiding developers in better comprehending the underlying causes [179]. With the integration of LLMs, bug report analysis tasks are conducted more efficiently and accurately advancing optimization and enhancement of the maintenance workflow.

Code clone detection. Code clones are code samples that are identical to each other [20, 138]. These code samples can have structural or semantic equivalence [300]. Sharma *et al.* [287] investigate BERT's application in code clone detection through an exploratory study. Analyzing BERT's attention to code markers, they found that identifiers received higher attention, advocating their use in clone detection. This insight enhanced clone detection across all layers, and the implications extended beyond BERT. The researchers suggest that these findings could lead to the development of smaller models with performance akin to larger ones, thus mitigating computational accessibility issues.

Logging. Logging involves the systematic recording of events, messages, or information during the operation of a software application. It provides valuable information for understanding the behavior, performance, and potential problems of an application. Developers strategically insert logging statements throughout the code base to capture relevant data such as variable values, function calls, and error messages. These logs are an important tool for testing [34, 36], debugging [281], monitoring [100, 101], and analyzing the behavior of software operations, helping developers identify and diagnose bugs, performance bottlenecks, and other critical issues. Mastropaolo *et al.* [211] introduce LANCE, a system for automatically generating and injecting full log statements into Java code using the T5 model. Sridhara *et al.* [296] present that ChatGPT performs well in the log summarization task, generating aggregated results that are better than the current state of the art.

Bug prediction. Gomes *et al.* [88] conduct a BERT and TF-IDF (Term Frequency-Inverted Document Frequency) application for long-lived bug prediction in Free/Libre Open-Source Software (FLOSS) study to compare their accuracy in predicting long-lived errors. The results show that BERT-based feature extraction consistently outperforms TF-IDF, demonstrating BERT's ability to capture the semantic context in error reports. In addition, smaller BERT architectures also show competitive results, highlighting the effectiveness of LLMs in bug prediction. This approach promises to enable more accurate error detection in FLOSS projects and improve software quality and maintenance.

Bug triage. Bug triage is pivotal for effective issue management in large projects. It entails prioritizing bugs and assigning appropriate developers for resolution. While bug triage is straightforward for smaller projects, scalability brings complexity. Finding the right developers with the needed skills becomes intricate as bugs vary in expertise requirements. Some even demand combined skills, amplifying the intricacy. Lee *et al.* [163] introduce the Light Bug Triage framework (LBT-P). This innovative approach employs BERT to extract semantic information from bug reports. To surmount challenges with LLMs in bug triage, the researchers employ techniques like model compression, knowledge preservation fine-tuning, and a new loss function.

Bug report replay. Bug reports are crucial for software maintenance, allowing users to inform developers of problems encountered while using the software. Therefore, researchers have invested significant resources in automating error playback to speed up the software maintenance process. The success of current automated approaches depends heavily on the characteristics and quality of error reports, as they are limited by manually created schemas and predefined vocabularies. Inspired by the success of the LLMs in natural language understanding, Feng *et al.* [73] propose AdbGPT, which utilizes natural language understanding and logical reasoning capabilities of the LLM to extract Steps to Reproduce (S2R) entities from bug reports and guide the bug replay process based on the current graphical user interface (GUI) state. The researchers describe how cue engineering, a small amount of learning, and thought chain reasoning can be utilized to leverage the knowledge of the LLM for automated error replay. This approach is significantly lightweight compared to traditional approaches, which utilize a single LLM to address both phases of S2R entity extraction and guided replay through novel hint engineering.

Duplicate bug report detection. In large software projects, multiple users may encounter and report the same or similar bugs independently, resulting in a proliferation of duplicate bug reports [122]. Duplicate bug report detection involves analyzing the textual content of bug reports and comparing them to find similarities and redundancies. LLM models, such as BERT [122], ChatGPT [296], and other transformer-based architectures, are well-suited for natural language understanding and contextual representation. When applied to this task, LLMs can effectively capture the semantic similarities between bug reports, even in cases with slight variations in language or phrasing. The utilization of LLMs in this context not only enhances efficiency in managing bug reports but also contributes to improving the overall software development and maintenance workflow, reducing redundancy, and ensuring prompt bug resolution [391].

Decompilation. Decompilation is crucial in many security and SE tasks. For example, decompilation is often the first step in malware analysis [222], where human analysts examine malware code to understand its behavior. It is also important for binary vulnerability analysis (where analysts want to identify critical vulnerabilities in executables) [55, 224], software supply chain analysis [106, 237], and code reuse (where legacy executables may need to be ported or hardened) [57, 208, 254]. Decompilation tools, such as IDA and Ghidra, have been useful in security threat analysis [223] proves its importance. Xu *et al.* [358] propose a new technique for recovering symbolic names during decompilation that leverages the synergy between LLMs (especially ChatGPT) and program analysis. The method employs an iterative algorithm to propagate ChatGPT query results based on program semantics. This propagation in turn provides better context for ChatGPT. The results

show that 75% of the recovered names are perceived as good by the users and that the technique outperforms the state-of-the-art by 16.5% and 20.23% in terms of precision and recall, respectively. **Program merge conflicts repair.** Program merge conflicts repair addresses the challenges faced when integrating individual code changes, which can lead to textual or semantic inconsistencies. Zhang *et al.* [382] explored the potential of using k-shot learning with LLMs like GPT-3 to automate this repair process. While these models showed promise in resolving semantic conflicts for Microsoft Edge, they didn't fully replace the benefits of domain-specific languages for certain synthesis patterns.

Sentiment analysis. Sentiment analysis involves determining emotions in text data related to software products, such as user feedback or comments [97, 121, 133]. The goal of sentiment analysis is to automatically classify the sentiment of the text as positive, negative, or neutral, providing valuable insights into how users perceive and react to software applications. Zhang *et al.* [392] conducted a study comparing pre-trained Transformer models like BERT, RoBERTa, XLNet, and ALBERT with existing SA4SE tools across six datasets. The results show that the Transformer models outperformed previous tools by 6.5% to 35.6% in macro/micro-averaged F1-scores, albeit with a trade-off in runtime efficiency. However, this accuracy boost comes with some runtime costs, indicating that while Transformer models are less efficient than existing SA4SE approaches, their runtime cost is not prohibitively high.

Tag recommendation. Improper tagging in software Q&A sites can lead to redundancy and other issues such as tag explosion. He *et al.* [103] introduced PTM4Tag, a framework utilizing PLMs with a triplet architecture to recommend tags for posts. By separately modeling the title, description, and code snippets of posts, PTM4Tag was compared using five popular PLMs, including BERT, CodeBERT, etc. The SE-specialized CodeBERT showed the best performance, notably surpassing CNN-based methods. An ablation study revealed that while the title was crucial in tag prediction, using all post components achieved the optimal result.

Vulnerability repair. Vulnerability repair is the process of identifying and fixing security holes or weaknesses in software applications. Pearce *et al.* [251] investigate how to use LLMs for software zero-point vulnerability remediation. The authors explore the challenges faced in designing hints to induce LLMs to generate fixed versions of insecure code. It shows that while the approach is promising, with LLMs capable of fixing 100% of synthetic and hand-created scenarios, a qualitative assessment of the model's performance on a corpus of historical real-life examples reveals challenges in generating functionally correct code. It is concluded that despite the potential for future targeted LLM applications in this area, challenges remain. For a complete end-to-end system, the full system needs to be evaluated in conjunction with error localization and an improved testbed.

Traceability recovery. Traceability recovery focuses on re-establishing lost or unclear connections between related software artifacts, thereby facilitating coherent software evolution and maintenance [83]. While traditional methods have offered some solutions, the integration of LLMs has recently emerged as a promising avenue for enhancing the accuracy and efficiency of this task. Zhu *et al.* [404] present TRACEFUN, a traceability link recovery framework enhanced with unlabeled data, serves as a testament to this potential, leveraging LLMs to bridge the gap between labeled and unlabeled data, thereby refining traceability link predictions.

6.7 How are LLMs used in software management?

Research papers describing the utilization of LLMs in software management are still limited.

Effort estimation. Effort estimation refers to the process of predicting the amount of time, resources, and manpower required to complete a software development project. Alhamed *et al.* [7] conduct an evaluation of the application of BERT in the task of effort estimation for software maintenance. Their study underscores BERT's potential to offer valuable insights and aid in the

decision-making process while also highlighting the associated challenges and need for further investigation.

RQ4 - Summary

- (1) We categorized SE tasks into six activities: requirements engineering, software design, software development, software quality assurance, software maintenance, and software management. Subsequently, we summarized the specific applications of LLMs in these SE activities.
- (2) We identified a total of 55 SE tasks and found that LLMs are most widely used in software development, with 136 papers mentioning 21 SE tasks. The least applied area, software management, was mentioned in only one study.
- (3) **Code generation and program repair are the most prevalent tasks for employing LLMs in software development and maintenance activities.** We analyze the top-performing LLMs repeatedly validated in these tasks and summarize novel findings.

7 THREATS TO VALIDITY

Paper search omission. One key limitation is the possibility of omitting relevant papers during the search process. When gathering papers related to LLM4SE tasks from various publishers, it is possible to miss some papers due to incomplete summarization of keywords for software engineering tasks or LLMs. To address this concern, we adopted a comprehensive approach, combining manual search, automated search, and snowballing techniques, to minimize the risk of missing relevant papers. For the manual search, we diligently searched for LLM papers related to SE tasks in six top-tier SE venues and extracted authoritative and comprehensive SE tasks and LLM keywords from these sources. With these numbered keyword search strings in place, we conducted automated searches on seven widely used publisher platforms. Additionally, to further augment our search results, we employed both forward and backward snowballing.

Study selection bias. Another limitation is the potential study selection bias. We established inclusion and exclusion criteria to perform the initial selection of papers, followed by manual verification based on quality assessment criteria (QAC). This process involves a combination of automated and manual procedures. The automated selection process may result in mislabeling of papers due to incomplete or ambiguous information in their corresponding BibTeX records. To mitigate this issue, any papers that cannot be confidently excluded are temporarily retained for manual verification. However, the manual verification stage could be influenced by the subjective judgment biases of the researchers, affecting the accuracy of the quality assessment of papers. To address these concerns, we invited two experienced reviewers in the fields of SE and LLM research to conduct a secondary review of the study selection results. This step aims to enhance the accuracy of our paper selection and minimize the likelihood of omission or misclassification. By implementing these measures, we strive to ensure that the selected papers are accurate and comprehensive, minimizing the impact of study selection bias and enhancing the reliability of our systematic literature review. We additionally provide a replication package⁵ for others to view.

8 CHALLENGES AND OPPORTUNITIES

8.1 Challenges

8.1.1 Challenges in LLM Applicability.

Model size and deployment. The size of LLMs has seen a marked increase over time, moving from GPT-1's 117M parameters to GPT-2's 1.5B, and further to GPT-3's 175B parameters [362]. The

⁵https://docs.google.com/spreadsheets/d/1iomMvoDL2znNDQ_J4aGnqb3BhZpEMlfz

billions and even trillions [219] of parameters pose significant storage, memory, and computational challenges, which can hinder LLMs in resource-limited and real-time scenarios, especially when developers lack access to powerful GPUs or TPUs. CodeBERT [74], a pre-trained model proposed in 2019, has a total of 125M parameters, resulting in a large model size of 476 MB. Recently proposed models like Codex [38] and CodeGen [232], have over 100 billion parameters and over 100 GB in size. The large sizes also require more computational resources. As pointed out by Hugging Face team [21], training a 176B model (i.e., BLOOM [282]) on 1.5 TB datasets consumes an estimated 1,082,880 GPU hours. Similarly, the training of the GPT-NeoX-20B model [23] on the Pile dataset [80], encompassing over 825 GiB of raw text data, requires the deployment of eight NVIDIA A100-SXM4-40GB GPUs. Each of these GPUs comes with a price tag of over 6,000 dollars [14], and the training extends to 1,830 hours or approximately 76 days. Moreover, even training a relatively smaller model like the PolyCoder (2.7B) [357], employing eight NVIDIA RTX 8000 GPUs on a single machine, demands a commitment of around 6 weeks. These examples illustrate the significant computational costs associated with training LLMs. These also have significant energy costs with predictions of massively increased energy usage by LLM-based platforms [270]. Fortunately, there are preliminary studies on reducing code models' size and improving their efficiency. Shi *et al.* [289] use a genetic algorithm to compress CodeBERT into only 3 MB and reduce its response latency by more than 70%. Overall, the challenge of increasing model sizes and efficient deployment requires further attention from the communities.

Data dependency. In Section 4, we provide a detailed analysis of the datasets used in 229 studies and the data preprocessing process, finding that LLMs rely heavily on a large number of different datasets for training and fine-tuning, posing the data dependency challenge. The quality, diversity, and quantity of data directly affect the performance and generalizability of the models. Given their size, LLMs often require large amounts of data to capture nuances, but obtaining such data can be challenging. Relying on limited or biased datasets may cause the model to inherit these biases, resulting in biased or inaccurate predictions. In addition, the domain-specific data required for fine-tuning can be a bottleneck. Due to the relatively short period of time since the emergence of LLM, such large-scale datasets are still relatively rare, especially in the SE domain. Another issue is the risk of benchmark data contamination, where training and test data overlaps could lead to inflated performance metrics [397]. For instance, Brown *et al.* [25] discovered a code bug that prevented them from fully removing all overlapping data. They were unable to afford retraining and resorted to using “cleaned” variants of the benchmarks to mitigate the issue. Moreover, there are grave concerns around the inclusion of Personally Identifiable Information (PII) in pre-training corpora. Instances of PII, such as phone numbers and email addresses, have led to privacy leaks during the prompting process [64, 154].

Ambiguity in code generation. Ambiguity in code generation poses a significant challenge for LLMs in SE tasks. When code intent is unclear (e.g., multiple valid solutions exist), LLMs may struggle to produce accurate and contextually appropriate code. This can lead to syntactically correct but functionally incorrect code, impacting the reliability and effectiveness of LLM-based code generation. Addressing this issue requires exploring techniques to incorporate additional context, domain-specific knowledge, or multi-model ensembles to improve LLMs' ability to handle ambiguity and generate precise code, ensuring their successful integration into real-world software development processes.

8.1.2 Challenges in LLM Generalizability. The generalizability of LLMs refers to the ability of these models to consistently and accurately perform tasks in different tasks, datasets, or domains outside their training environment. While LLMs are trained on massive amounts of data, ensuring extensive knowledge capture, their performance is sometimes problematic when confronted with specific or

idiosyncratic tasks outside the scope of their training. This challenge is particularly evident in the SE domain, where we present the application of LLMs to 55 SE tasks in Section 6. We observed that the context and semantics of code or documents vary greatly across projects, languages, or domains. Ensuring that the LLM generalizes well requires careful fine-tuning, validation on different datasets, and continuous feedback loops. Without these measures, models run the risk of over-adapting their training data, thus limiting their usefulness in a variety of real-world applications. Recent studies have shown that the LLMs cannot generalize their good performance to inputs after semantic-preserving transformations. For example, Yang *et al.* [365] show that the performance of CodeBERT on different tasks decreases significantly after substituting the variables' names in the input.

8.1.3 Challenges in LLM Evaluation. We summarized key evaluation metrics used in different types of SE tasks according to four task types: regression, classification, recommendation, and generation (Section 6). We found that when applying LLMs in the software engineering domain, the methodology for evaluating the performance of the models is usually based on a set of predefined metrics. Unfortunately, these metrics (e.g., Accuracy, Recall, or F1-score), while useful in some cases, may not fully capture all the effects and impacts of a model in a given SE task. For example, a model may perform well in terms of accuracy but may fail in processing specific types of inputs or in some specific situations. In addition, these metrics may not capture certain qualitative aspects of the model, such as its interpretability, robustness, or sensitivity to specific types of errors. Some of the most recent studies on LLM4SE tasks [112, 294, 358, 373, 388], in which researchers customized some evaluation metrics to assess the performance of models, also further illustrate the limitations of some of the widely used evaluation metrics in the field of LLM.

8.1.4 Challenges in LLM Interpretability, Trustworthiness, and Ethical Usage. Interpretability and trustworthiness are crucial aspects in the adoption of LLMs for SE tasks. The challenge lies in understanding the decision-making process of these models, as their black-box nature often makes it difficult to explain why or how a particular code snippet or recommendation is generated. Recent studies [169, 323, 366] also show that LLM of code trained on low-quality datasets can have vulnerabilities (e.g., generating insecure code). The lack of interpretability and trustworthiness can lead to uncertainty and hesitation among developers, who may be hesitant to rely on LLM-generated code without a clear understanding of how it was derived. Establishing trust in LLMs requires efforts to develop techniques and tools that provide insights into the model's internal workings and enable developers to comprehend the reasoning behind the generated outputs. Enhancing interpretability and trustworthiness can ultimately promote the widespread adoption of LLMs in SE, leading to more efficient and effective development practices. Many LLMs are not open and it is unclear what data they have been trained on, both quality and representativeness but also ownership of the source training data. This brings into question ownership of the derivative data, e.g., generated designs, code, or test cases. There is also potential for various adversarial attacks e.g. deliberately seeding LLMs with code vulnerabilities so that automatically generated code snippets have subtle but vulnerable aspects.

8.2 Opportunities

8.2.1 Optimization of LLM4SE.

The advent of code-specialized LLMs in SE. The recent emergence of code-specialized LLMs, such as GitHub Copilot [87], Amazon's CodeWhisperer [13], OpenAI Code Interpreter [240] integrated into ChatGPT, and Code Llama [213] from Meta's Llama family, signals a transformative phase in LLM4SE. These specialized LLMs, fine-tuned on code-specific datasets, are not merely incremental improvements but paradigm shifts in code understanding, generation, and efficiency.

They offer new avenues for automated coding, personalized developer assistance, enhanced code review, and quality assurance, among other tasks, setting the stage for groundbreaking advancements in the SE domain.

Influence and applications of ChatGPT. ChatGPT's popularity in recent academic research, as evidenced by its large presence in our 229 analyzed papers, emphasizes its escalating influence and acceptance within academia. Researchers' preference for ChatGPT over other LLMs and LLM-based applications since its release can be attributed to its computational efficiency, adaptability to various tasks, and potential cost-effectiveness [160, 168, 354]. Its applications extend beyond mere code efficiency and debugging, fostering a collaborative era in development. This paradigm shift signifies a broader move towards integrating advanced natural language understanding into conventional coding practices [160, 201, 277]. By thoughtfully analyzing these dynamics and trends, we can foresee the potential pathways for LLMs and LLM applications like ChatGPT in shaping more robust, efficient, and collaborative software development procedures. Such insights stand as a promising indication of the future revolutionary impact of LLMs on SE.

Performance enhancement from task-specific model training. The choice between leveraging commercially available pre-trained models like GPT-4 and building upon open-source frameworks such as LLaMA [313], Llama 2 [314], and Alpaca [9] (fine-tuned from LLaMA 7B on 52K instruction-following demonstrations) provides a nuanced set of options for individual or organizational customization in specialized tasks. The distinction between these two approaches lies in the degree of control and customization. Pre-trained models like GPT-4 are generally not designed for large-scale retraining due to their proprietary nature, but they allow quick task-specific adaptations with limited data, thereby minimizing computational overhead. On the other hand, frameworks like LLaMA offer an open-source foundation for more extensive customization. While they come pre-trained, organizations often modify the source code and retrain these models on their own large-scale datasets to meet specialized requirements [110, 370]. This process is computationally intensive, leading to greater resource allocation and cost, but affords the advantage of creating highly domain-specific models. Hence, the primary trade-off is between the ease of use and quick deployment offered by models like GPT-4, and the deep customization capabilities but higher computational demands associated with open-source frameworks like LLaMA.

Collaborative LLMs. From our review it is evident that LLMs have made significant strides in addressing various SE challenges. However, as the complexity of SE tasks continues to grow, there's an emerging need for more sophisticated and tailored solutions. One promising direction is the concept of Collaborative LLMs. This approach involves integrating multiple LLMs [60, 396] or combining LLMs with specialized machine-learning models [66, 381] to enhance their efficacy for SE tasks. By harnessing the collective strengths of different models, we believe that the SE community can achieve more precise and efficient outcomes, from code completion to bug detection.

8.2.2 Expanding LLM's NLP Capabilities in More SE Phases.

Integration of new input forms. In our analysis we observed that the predominant input forms were code-based datasets and text-based datasets. However, there was a noticeable scarcity of graph-based datasets [151] (Section 4). Leveraging new input forms of natural language, such as spoken language, diagrams, and multimodal inputs, presents an opportunity to enhance the LLMs' ability to understand and process diverse user requirements. Integrating spoken language could improve interactions between developers and models, enabling more natural and context-rich communication. Diagrams can facilitate visual representations of code and requirements, offering a complementary perspective for code generation. Furthermore, multimodal inputs that combine text, audio, and visual cues could offer a more comprehensive context understanding, leading to more accurate and contextually appropriate code generation. Additionally, exploring graph-based

datasets could be crucial for addressing complex code scenarios, as graphs capture the structural relationships and dependencies in code, allowing LLMs to better comprehend code interactions and dependencies.

Widening LLM applications across SE phases. We observed a pronounced emphasis on the application of LLMs in software development and maintenance. These areas have undoubtedly benefited from the capabilities of LLMs, leading to enhanced code completion [124, 181, 189], bug detection [45, 73, 136], and other related tasks. The current application of LLMs in requirements engineering, software design, and software management remains relatively sparse. This presents a significant opportunity: by expanding the use of LLMs to these under-explored areas, we can potentially improve how requirements are elicited, how software designs are conceptualized, and how projects are managed.

8.2.3 Enhancing LLM's Performance in Existing SE Tasks.

Tackling domain-specific challenges. Many SE domains, including safety-critical systems and specific industries, suffer from a scarcity of open-source datasets, hindering the application of LLMs in these specialized areas. Future research can focus on creating domain-specific datasets and fine-tuning LLMs to cater to the unique challenges and intricacies of these fields [22, 298]. Collaboration with domain experts and practitioners is vital to curate relevant data, and fine-tuning LLMs on this data can enhance their effectiveness and ensure better alignment with the specific requirements of each domain, paving the way for LLMs to address real-world challenges [26] in diverse software engineering domains [173].

Establishing a comprehensive evaluation framework for LLM4SE. The necessity for a universal, yet adaptable, evaluation framework for LLM4SE is pressing for both academic and industrial sectors. In academia, such a framework enables streamlined assessments of LLM performance, efficacy, and limitations, serving as a benchmark to verify the models' practical readiness. On the industrial side, collaborations with real-world development teams using this framework yield empirical insights into LLMs' utility, including their impacts on productivity, code quality, and team collaboration, while also revealing challenges like model biases, misinterpretation of code semantics, and context-specific limitations. Establishing this framework is critical for standardizing assessments and facilitating responsible LLM adoption in both academic research and practical applications [22, 89].

8.3 Roadmap

We provide a roadmap for future development in leveraging Large Language Models for Software Engineering (LLM4SE), with an additional high-level perspective that acknowledges the reciprocal relationship and emerging exploration of Software Engineering for Large Language Models (SE4LLM).

Automated coding, development and personalized developer assistance. The pursuit of automation in coding encompasses the auto-generation of code snippets, bug fixes, system optimization, and the creation of intelligent, personalized assistance for developers that is context-aware and adaptable to individual needs. LLM's generative capabilities can be leveraged to help developers better understand requirements and generate syntactically and semantically correct code, thereby accelerating development cycles and improving software quality. Leveraging LLM's natural language processing to develop context-aware tools allows for interaction with developers in a more intuitive and responsive manner. Additionally, fine-tuning LLMs for specific coding tasks and developer assistance can further enhance their accuracy and efficiency, customizing the automation process to suit the unique demands of different projects and individuals.

Advancing testing and analysis. The inclusion of LLMs in software testing methods opens up avenues for enhanced test case generation, bug classification, and defect prediction, thereby improving the precision and efficiency of the software testing process. For instance, LLMs show potential to be fine-tuned to a project's specific requirements to generate customized test cases, which elevates the likelihood of early detection of subtle bugs or security vulnerabilities. Furthermore, the integration of LLMs with traditional SE techniques, including both static and dynamic program analysis presents a compelling direction for more rigorous code analysis. The potential for utilizing LLMs in formal analysis methodologies, including formal verification, is another area that merits investigation [32]. These advancements not only facilitate the early discovery of complex errors but also lead to reduced development costs and quicker time-to-market, ultimately contributing to the robustness and reliability of the software products.

Integrating programming knowledge into LLMs. One critical future direction lies in the integration of specialized code representation methods and programming domain knowledge into LLM4SE [202, 324]. This integration aims to enhance the capability of LLMs to generate code that is not only functionally accurate but also secure and compliant with programming standards. Leveraging advanced techniques in code embedding, syntax tree parsing, and semantic analysis could significantly refine the generation capabilities of LLMs. Moreover, embedding domain-specific rules and best practices into these models would enable them to auto-generate code that adheres to industry or language-specific guidelines for security and style.

Enhanced code review and quality assurance. The transformation of the code review process can be supported by employing LLMs to analyze code context, perform intelligent comparisons, and offer insights that go beyond traditional automated review systems. The application of fine-tuned LLMs for code review can allow for more precise error detection and tailored feedback, offering a more nuanced understanding of code quality and potential improvements.

Extracting insights from data mining. LLMs can play a critical role in mining insights from platforms like GitHub, StackOverflow, and app stores. Through the application in tasks such as requirement extraction, traceability, validation, and various types of mining (tag, app, developer-based), LLMs can provide valuable insights that inform development strategies and decision-making. By automating and enhancing these mining tasks, LLMs contribute to a deeper understanding of user needs, emerging trends, and the efficiency of development practices.

Empowering predictive analytics and decision support. Leveraging LLMs for effort cost prediction, software classification, code classification, incident detection, and software quality evaluation may support better data-driven insights and predictive analytics. This empowers organizations to make informed decisions throughout the development lifecycle. LLMs' ability to model and analyze vast amounts of data enables more accurate forecasts of project timelines, resource needs, and potential risks.

LLMs in software security. The growing impact of LLM4SE offers both unparalleled opportunities and challenges in the domain of software security. On the one hand, LLMs offer promising solutions for automated security audits, compliance verifications, and vulnerability detection. These models can potentially be leveraged for automated code reviews to ensure compliance with industry standards and legal regulations, while also identifying potential security vulnerabilities [2, 49, 73, 75, 99, 252]. For instance, Ferrag *et al.* [75] showcased the efficacy of LLMs in cyber reasoning tasks related to software security. On the other hand, the usage of LLMs introduces novel security concerns. Their complexity makes them susceptible to attacks, demanding novel strategies to fortify the models themselves [48, 65, 191, 265, 266, 349]. As an example, Wu *et al.* [349] delve into methods to secure LLMs against jailbreak attacks. An intriguing direction for future research lies in enabling LLMs to automatically identify and rectify their own vulnerabilities. Specifically, the focus could be on equipping LLMs to generate self-applied patches to their underlying code, thereby enhancing

their inherent security, as opposed to merely implementing application-layer restrictions. Given this landscape, future research should adopt a balanced approach, aiming to exploit LLMs for automating and enhancing existing software security protocols while concurrently developing techniques to secure the LLMs themselves. This dual focus is crucial for fully realizing the potential of LLMs in enhancing the security and compliance assurance of software systems.

Software Engineering for Large Language Models (SE4LLM). As the capabilities and complexities of LLMs continue to expand, there arises a reciprocal need for specialized SE practices tailored for the development, optimization, and maintenance of these models. SE4LLM encompasses a range of challenges and opportunities, including the design of scalable and maintainable architectures, the creation of efficient training algorithms, the development of rigorous testing frameworks for model robustness and fairness, and the implementation of ethical guidelines and compliance mechanisms. The convergence of SE with LLMs not only facilitates the growth of more sophisticated and adaptable models but also opens up new avenues for interdisciplinary research and innovation, bringing together the expertise of both the AI and SE communities. This aligns with a broader vision where SE practices become an integral part of the lifecycle of LLMs, ensuring their robustness, efficiency, and ethical alignment with societal values.

9 CONCLUSION

LLMs are bringing significant changes to the field of SE. The potential of these models to handle complex tasks can fundamentally reshape many SE practices and tools. In this systematic literature review, we analyzed the emerging utilization of LLMs for software engineering, encompassing papers published since the inception of the first LLM (BERT). We examined the diverse LLMs that have been employed in SE tasks and explored their distinct features and applications (RQ1). We then investigated the processes involved in data collection, preprocessing, and usage, emphasizing the significant role well-curated datasets play in the successful application of LLMs to solve SE tasks (RQ2). Following this, we investigated the various strategies utilized to optimize and assess the performance of LLMs for SE tasks (RQ3). Lastly, we reviewed the wide range of SE tasks where LLMs have been applied to date, shedding light on the practical contributions LLMs have made (RQ4). We summarised some key existing challenges of LLM4SE and provided a research roadmap, outlining promising future research directions.

REFERENCES

- [1] Emad Aghajani, Csaba Nagy, Mario Linares-Vásquez, Laura Moreno, Gabriele Bavota, Michele Lanza, and David C Shepherd. 2020. Software documentation: the practitioners' perspective. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 590–601.
- [2] Baleegh Ahmad, Shailja Thakur, Benjamin Tan, Ramesh Karri, and Hammond Pearce. 2023. Fixing Hardware Security Bugs with Large Language Models. *arXiv preprint arXiv:2302.01215* (2023).
- [3] Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Unified pre-training for program understanding and generation. *arXiv preprint arXiv:2103.06333* (2021).
- [4] Toufique Ahmed, Kunal Suresh Pai, Premkumar Devanbu, and Earl T Barr. 2023. Improving Few-Shot Prompts with Relevant Static Analysis Products. *arXiv preprint arXiv:2304.06815* (2023).
- [5] Alfred V Aho, Ravi Sethi, Jeffrey D Ullman, et al. 2007. *Compilers: principles, techniques, and tools*. Vol. 2. Addison-wesley Reading.
- [6] Ali Al-Kaswan, Toufique Ahmed, Maliheh Izadi, Anand Ashok Sawant, Premkumar Devanbu, and Arie van Deursen. 2023. Extending Source Code Pre-Trained Language Models to Summarise Decompiled Binaries. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 260–271.
- [7] Mohammed Alhamed and Tim Storer. 2022. Evaluation of Context-Aware Language Models and Experts for Effort Estimation of Software Maintenance Issues. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 129–138.
- [8] Frances E Allen. 1970. Control flow analysis. *ACM Sigplan Notices* 5, 7 (1970), 1–19.

- [9] Stanford Alpaca. 2023. Stanford Alpaca: An Instruction-following LLaMA Model. https://github.com/tatsu-lab/stanford_alpaca.
- [10] Mansour Alqarni and Akramul Azim. 2022. Low level source code vulnerability detection using advanced bert language model. In *Proceedings of the Canadian Conference on Artificial Intelligence-Https://caiac. pubpub. org/pub/gdhhb8oq4 (may 27 2022)*.
- [11] Rajeev Alur, Rastislav Bodik, Garvit Juniwal, Milo MK Martin, Mukund Raghothaman, Sanjit A Seshia, Rishabh Singh, Armando Solar-Lezama, Emina Torlak, and Abhishek Udupa. 2013. *Syntax-guided synthesis*. IEEE.
- [12] Sven Amann, Sebastian Proksch, Sarah Nadi, and Mira Mezini. 2016. A study of visual studio usage in practice. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, Vol. 1. IEEE, 124–134.
- [13] Amazon. 2023. Amazon CodeWhisperer. <https://aws.amazon.com/cn/codewhisperer/>.
- [14] Amazon. 2023. NVIDIA Tesla A100 Ampere 40 GB Graphics Card - PCIe 4.0 - Dual Slot. <https://www.amazon.com/NVIDIA-Tesla-A100-Ampere-Graphics/dp/B0BGZJ27SL>.
- [15] M Anon. 2022. National vulnerability database. <https://www.nist.gov/programs-projects/national-vulnerability-database-nvd>.
- [16] Shushan Arakelyan, Rocktim Jyoti Das, Yi Mao, and Xiang Ren. 2023. Exploring Distributional Shifts in Large Language Models for Code Analysis. *arXiv preprint arXiv:2303.09128* (2023).
- [17] Amos Azaria, Rina Azoulay, and Shulamit Reches. 2023. ChatGPT is a Remarkable Tool-For Experts. *arXiv preprint arXiv:2306.03102* (2023).
- [18] Patrick Bareiß, Beatriz Souza, Marcelo d’Amorim, and Michael Pradel. 2022. Code generation tools (almost) for free? a study of few-shot, pre-trained language models on code. *arXiv preprint arXiv:2206.01335* (2022).
- [19] Rabih Bashroush, Muhammad Garba, Rick Rabiser, Iris Groher, and Goetz Botterweck. 2017. Case tool support for variability management in software product lines. *ACM Computing Surveys (CSUR)* 50, 1 (2017), 1–45.
- [20] Ira D Baxter, Andrew Yahin, Leonardo Moura, Marcelo Sant’Anna, and Lorraine Bier. 1998. Clone detection using abstract syntax trees. In *Proceedings. International Conference on Software Maintenance (Cat. No. 98CB36272)*. IEEE, 368–377.
- [21] Stas Bekman. 2022. The Technology Behind BLOOM Training. <https://huggingface.co/blog/bloom-megatron-deepspeed>.
- [22] Eeshita Biswas, Mehmet Efruz Karabulut, Lori Pollock, and K Vijay-Shanker. 2020. Achieving reliable sentiment analysis in the software engineering domain using bert. In *2020 IEEE International conference on software maintenance and evolution (ICSME)*. IEEE, 162–173.
- [23] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745* (2022).
- [24] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*. <https://doi.org/10.5281/zenodo.5297715>
- [25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [26] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [27] Nghi DQ Bui, Hung Le, Yue Wang, Junnan Li, Akhilesh Deepak Gotmare, and Steven CH Hoi. 2023. CodeTF: One-stop Transformer Library for State-of-the-art Code LLM. *arXiv preprint arXiv:2306.00029* (2023).
- [28] Jialun Cao, Meiziniu Li, Ming Wen, and Shing-chi Cheung. 2023. A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. *arXiv preprint arXiv:2304.08191* (2023).
- [29] Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. 2023. MultiPL-E: a scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering* (2023).
- [30] Aaron Chan, Anant Kharkar, Roshanak Zilouchian Moghaddam, Yevhen Mohylevskyy, Alec Helyar, Eslam Kamal, Mohamed Elkamhawy, and Neel Sundaresan. 2023. Transformer-based Vulnerability Detection in Code at EditTime: Zero-shot, Few-shot, or Fine-tuning? *arXiv preprint arXiv:2306.01754* (2023).
- [31] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).
- [32] Yiannis Charalambous, Norbert Tihanyi, Ridhi Jain, Youcheng Sun, Mohamed Amine Ferrag, and Lucas C Cordeiro. 2023. A New Era in Software Security: Towards Self-Healing Software via Large Language Models and Formal Verification. *arXiv preprint arXiv:2305.14752* (2023).

- [33] Angelica Chen, J  r  my Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, and Ethan Perez. 2023. Improving code generation by training with natural language feedback. *arXiv preprint arXiv:2303.16749* (2023).
- [34] Boyuan Chen, Jian Song, Peng Xu, Xing Hu, and Zhen Ming Jiang. 2018. An automated approach to estimating code coverage measures via execution logs. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 305–316.
- [35] Fuxiang Chen, Fatemeh H Fard, David Lo, and Timofey Bryksin. 2022. On the transferability of pre-trained language models for low-resource programming languages. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. 401–412.
- [36] Jinfu Chen, Weiye Shang, Ahmed E Hassan, Yong Wang, and Jiangbin Lin. 2019. An experience report of generating load tests using log-recovered workloads at varying granularities of user behaviour. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 669–681.
- [37] Long Chen, Wei Ye, and Shikun Zhang. 2019. Capturing source code semantics via tree-based convolution over API-enhanced AST. In *Proceedings of the 16th ACM International Conference on Computing Frontiers*. 174–182.
- [38] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [39] Xinyun Chen, Maxwell Lin, Nathanael Sch  rli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128* (2023).
- [40] Xinyun Chen, Chang Liu, and Dawn Song. 2017. Towards synthesizing complex programs from input-output examples. *arXiv preprint arXiv:1706.01284* (2017).
- [41] Xinyun Chen, Dawn Song, and Yuandong Tian. 2021. Latent execution for neural program synthesis beyond domain-specific languages. *Advances in Neural Information Processing Systems* 34 (2021), 22196–22208.
- [42] Yizheng Chen, Zhoujie Ding, Xinyun Chen, and David Wagner. 2023. DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection. *arXiv preprint arXiv:2304.00409* (2023).
- [43] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023).
- [44] Agnieszka Ciborowska and Kostadin Damevski. 2022. Fast changeset-based bug localization with BERT. In *Proceedings of the 44th International Conference on Software Engineering*. 946–957.
- [45] Agnieszka Ciborowska and Kostadin Damevski. 2023. Too Few Bug Reports? Exploring Data Augmentation for Improved Changeset-based Bug Localization. *arXiv preprint arXiv:2305.16430* (2023).
- [46] Matteo Ciniselli, Nathan Cooper, Luca Pascarella, Antonio Mastropaolo, Emad Aghajani, Denys Poshyvanyk, Massimiliano Di Penta, and Gabriele Bavota. 2021. An empirical study on the usage of transformer models for code completion. *IEEE Transactions on Software Engineering* 48, 12 (2021), 4818–4837.
- [47] Colin B Clement, Dawn Drain, Jonathan Timcheck, Alexey Svyatkovskiy, and Neel Sundaresan. 2020. PyMT5: multi-mode translation of natural language and Python code with transformers. *arXiv preprint arXiv:2010.03150* (2020).
- [48] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated Jailbreak Across Multiple Large Language Model Chatbots. *arXiv preprint arXiv:2307.08715* (2023).
- [49] Gelei Deng, Yi Liu, V  ctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. 2023. PentestGPT: An LLM-empowered Automatic Penetration Testing Tool. *arXiv preprint arXiv:2308.06782* (2023).
- [50] Yinlin Deng, Chunqiu Steven Xia, Haoran Peng, Chenyuan Yang, and Lingming Zhang. 2023. Large Language Models are Zero-Shot Fuzzers: Fuzzing Deep-Learning Libraries via Large Language Models. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2023)*.
- [51] Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. 2023. Large language models are edge-case fuzzers: Testing deep learning libraries via fuzzgpt. *arXiv preprint arXiv:2304.02014* (2023).
- [52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [53] Juri Di Rocco, Davide Di Ruscio, Claudio Di Sipio, Phuong T Nguyen, and Riccardo Rub  i. 2021. Development of recommendation systems for software engineering: the CROSSMINER experience. *Empirical Software Engineering* 26, 4 (2021), 69.
- [54] Victor Dibia, Adam Fourney, Gagan Bansal, Forough Poursabzi-Sangdeh, Han Liu, and Saleema Amershi. 2022. Aligning Offline Metrics and Human Judgments of Value of AI-Pair Programmers. *arXiv preprint arXiv:2210.16494*

(2022).

- [55] Sushant Dinesh, Nathan Burow, Dongyan Xu, and Mathias Payer. 2020. Retrowrite: Statically instrumenting cots binaries for fuzzing and sanitization. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1497–1511.
- [56] Hantian Ding, Varun Kumar, Yuchen Tian, Zijian Wang, Rob Kwiatkowski, Xiaopeng Li, Murali Krishna Ramanathan, Baishakhi Ray, Parminder Bhatia, Sudipta Sengupta, et al. 2023. A static evaluation of code completion by large language models. *arXiv preprint arXiv:2306.03203* (2023).
- [57] Steven HH Ding, Benjamin CM Fung, and Philippe Charland. 2019. Asm2vec: Boosting static representation robustness for binary clone search against code obfuscation and compiler optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 472–489.
- [58] Tuan Dinh, Jinman Zhao, Samson Tan, Renato Negrinho, Leonard Lausen, Sheng Zha, and George Karypis. 2023. Large Language Models of Code Fail at Completing Code with Potential Bugs. *arXiv preprint arXiv:2306.03438* (2023).
- [59] Jean-Baptiste Döderlein, Mathieu Acher, Djamel Eddine Khelladi, and Benoit Combemale. 2022. Piloting Copilot and Codex: Hot Temperature, Cold Prompts, or Black Magic? *arXiv preprint arXiv:2210.14699* (2022).
- [60] Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023. Self-collaboration Code Generation via ChatGPT. *arXiv preprint arXiv:2304.07590* (2023).
- [61] Shihan Dou, Junjie Shan, Haoxiang Jia, Wenhao Deng, Zhiheng Xi, Wei He, Yueming Wu, Tao Gui, Yang Liu, and Xuanjing Huang. 2023. Towards Understanding the Capability of Large Language Models on Code Clone Detection: A Survey. *arXiv preprint arXiv:2308.01191* (2023).
- [62] Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. 2023. ClassEval: A Manually-Crafted Benchmark for Evaluating LLMs on Class-level Code Generation. *arXiv preprint arXiv:2308.01861* (2023).
- [63] Moritz Eck, Fabio Palomba, Marco Castelluccio, and Alberto Bacchelli. 2019. Understanding flaky tests: The developer’s perspective. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 830–840.
- [64] El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyen Hoang, Rafael Pinot, Sébastien Rouault, and John Stephan. 2023. On the Impossible Safety of Large AI Models. *arXiv:2209.15259* [cs.LG]
- [65] Andre Elizondo. 2023. LangKit: Making Large Language Models Safe and Responsible. <https://whylabs.ai/blog/posts/langkit-making-large-language-models-safe-and-responsible>.
- [66] Saad Ezzini, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. 2022. Automated handling of anaphoric ambiguity in requirements: a multi-solution study. In *Proceedings of the 44th International Conference on Software Engineering*. 187–199.
- [67] Sarah Fakhoury, Saikat Chakraborty, Madan Musuvathi, and Shuvendu K Lahiri. 2023. Towards Generating Functionally Correct Code Edits from Natural Language Issue Descriptions. *arXiv preprint arXiv:2304.03816* (2023).
- [68] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2023. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020* (2023).
- [69] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).
- [70] Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated repair of programs from large language models. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1469–1481.
- [71] Zhiyu Fan, Xiang Gao, Abhik Roychoudhury, and Shin Hwei Tan. 2022. Automated Repair of Programs from Large Language Models. *arXiv preprint arXiv:2205.10583* (2022).
- [72] Sakina Fatima, Taher A Ghaleb, and Lionel Briand. 2022. Flakify: A black-box, language model-based predictor for flaky tests. *IEEE Transactions on Software Engineering* (2022).
- [73] Sidong Feng and Chunyang Chen. 2023. Prompting Is All Your Need: Automated Android Bug Replay with Large Language Models. *arXiv preprint arXiv:2306.01987* (2023).
- [74] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155* (2020).
- [75] Mohamed Amine Ferrag, Ammar Battah, Norbert Tihanyi, Merouane Debbah, Thierry Lestable, and Lucas C Cordeiro. 2023. SecureFalcon: The Next Cyber Reasoning System for Cyber Security. *arXiv preprint arXiv:2307.06616* (2023).
- [76] Gordon Fraser, Matt Staats, Phil McMinn, Andrea Arcuri, and Frank Padberg. 2015. Does automated unit test generation really help software testers? a controlled empirical study. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 24, 4 (2015), 1–49.
- [77] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen-tau Yih, Luke Zettlemoyer, and Mike Lewis. 2022. InCoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999* (2022).

- [78] Michael Fu and Chakkrit Tantithamthavorn. 2022. GPT2SP: A transformer-based agile story point estimation approach. *IEEE Transactions on Software Engineering* 49, 2 (2022), 611–625.
- [79] Apurva Gandhi, Thong Q Nguyen, Huitian Jiao, Robert Steen, and Ameya Bhatawdekar. 2023. Natural Language Commanding via Program Synthesis. *arXiv preprint arXiv:2306.03460* (2023).
- [80] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027* (2020).
- [81] Shuzheng Gao, Xin-Cheng Wen, Cuiyun Gao, Wenxuan Wang, and Michael R Lyu. 2023. Constructing Effective In-Context Demonstration for Code Intelligence Tasks: An Empirical Study. *arXiv preprint arXiv:2304.07575* (2023).
- [82] Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large Language Models are Few-Shot Summarizers: Multi-Intent Comment Generation via In-Context Learning. (2024).
- [83] Malcom Gethers, Rocco Oliveto, Denys Poshyvanyk, and Andrea De Lucia. 2011. On integrating orthogonal information retrieval methods to improve traceability recovery. In *2011 27th IEEE International Conference on Software Maintenance (ICSM)*. IEEE, 133–142.
- [84] Lobna Ghadhab, Ilyes Jenhani, Mohamed Wiem Mkaouer, and Montassar Ben Messaoud. 2021. Augmenting commit classification by using fine-grained source code changes and a pre-trained deep neural language model. *Information and Software Technology* 135 (2021), 106566.
- [85] Henry Gilbert, Michael Sandborn, Douglas C Schmidt, Jesse Spencer-Smith, and Jules White. 2023. Semantic Compression With Large Language Models. *arXiv preprint arXiv:2304.12512* (2023).
- [86] Github. 2023. Github. <https://github.com/>.
- [87] GitHub. 2023. Github copilot. <https://copilot.github.com>.
- [88] Luiz Gomes, Ricardo da Silva Torres, and Mario Lúcio Côrtes. 2023. BERT-and TF-IDF-based feature extraction for long-lived bug prediction in FLOSS: a comparative study. *Information and Software Technology* 160 (2023), 107217.
- [89] Lina Gong, Jingxuan Zhang, Mingqiang Wei, Haoxiang Zhang, and Zhiqiu Huang. 2023. What is the intended usage context of this model? An exploratory study of pre-trained models on various model repositories. *ACM Transactions on Software Engineering and Methodology* 32, 3 (2023), 1–57.
- [90] Google. 2023. Bard. <https://bard.google.com/>.
- [91] Jian Gu, Pasquale Salza, and Harald C Gall. 2022. Assemble foundation models for automatic code summarization. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 935–946.
- [92] Xiaodong Gu, Hongyu Zhang, and Sunghun Kim. 2018. Deep code search. In *Proceedings of the 40th International Conference on Software Engineering*. 933–944.
- [93] Xiaodong Gu, Hongyu Zhang, Dongmei Zhang, and Sunghun Kim. 2016. Deep API learning. In *Proceedings of the 2016 24th ACM SIGSOFT international symposium on foundations of software engineering*. 631–642.
- [94] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332* (2021).
- [95] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *arXiv preprint arXiv:2009.08366* (2020).
- [96] Priyanshu Gupta, Avishree Khare, Yasharth Bajpai, Saikat Chakraborty, Sumit Gulwani, Aditya Kanade, Arjun Radhakrishna, Gustavo Soares, and Ashish Tiwari. 2023. GrACE: Generation using Associated Code Edits. *arXiv preprint arXiv:2305.14129* (2023).
- [97] Emitza Guzman, David Azócar, and Yang Li. 2014. Sentiment analysis of commit comments in GitHub: an empirical study. In *Proceedings of the 11th working conference on mining software repositories*. 352–355.
- [98] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. Ptr: Prompt tuning with rules for text classification. *AI Open* 3 (2022), 182–192.
- [99] Andreas Happe and Jürgen Cito. 2023. Getting pwn'd by AI: Penetration Testing with Large Language Models. *arXiv preprint arXiv:2308.00121* (2023).
- [100] Julian Harty, Haonan Zhang, Lili Wei, Luca Pascarella, Mauricio Aniche, and Weiyi Shang. 2021. Logging practices with mobile analytics: An empirical study on firebase. In *2021 IEEE/ACM 8th International Conference on Mobile Software Engineering and Systems (MobileSoft)*. IEEE, 56–60.
- [101] Wilhelm Hasselbring and André van Hoorn. 2020. Kieker: A monitoring framework for software engineering research. *Software Impacts* 5 (2020), 100019.
- [102] Junda He, Zhou Xin, Bowen Xu, Ting Zhang, Kisub Kim, Zhou Yang, Ferdian Thung, Ivana Irsan, and David Lo. 2023. Representation Learning for Stack Overflow Posts: How Far are We? *arXiv preprint arXiv:2303.06853* (2023).
- [103] Junda He, Bowen Xu, Zhou Yang, DongGyun Han, Chengran Yang, and David Lo. 2022. PTM4Tag: sharpening tag recommendation of stack overflow posts with pre-trained models. In *Proceedings of the 30th IEEE/ACM International*

- Conference on Program Comprehension*. 1–11.
- [104] Vincent J Hellendoorn, Christian Bird, Earl T Barr, and Miltiadis Allamanis. 2018. Deep learning type inference. In *Proceedings of the 2018 26th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering*. 152–162.
 - [105] Robert Kraig Helmececi, Mucahit Cevik, and Savas Yildirim. 2023. Few-shot learning for sentence pair classification and its applications in software engineering. *arXiv preprint arXiv:2306.08058* (2023).
 - [106] Armijn Hemel, Karl Trygve Kalleberg, Rob Vermaas, and Eelco Dolstra. 2011. Finding software license violations through binary code clone detection. In *Proceedings of the 8th Working Conference on Mining Software Repositories*. 63–72.
 - [107] Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, et al. 2021. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938* (2021).
 - [108] Jordan Henkel, Denini Silva, Leopoldo Teixeira, Marcelo d’Amorim, and Thomas Reps. 2021. Shipwright: A human-in-the-loop system for dockerfile repair. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 1148–1160.
 - [109] Tobias Hey, Jan Keim, Anne Kozirolek, and Walter F Tichy. 2020. Norbert: Transfer learning for requirements classification. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 169–179.
 - [110] hiyouga. 2023. LLaMA Efficient Tuning. <https://github.com/hiyouga/LLaMA-Efficient-Tuning>.
 - [111] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).
 - [112] Jie Hu, Qian Zhang, and Heng Yin. 2023. Augmenting Greybox Fuzzing with Generative AI. *arXiv preprint arXiv:2306.06782* (2023).
 - [113] Xing Hu, Ge Li, Xin Xia, David Lo, and Zhi Jin. 2018. Deep code comment generation. In *Proceedings of the 26th conference on program comprehension*. 200–210.
 - [114] Di Huang, Ziyuan Nan, Xing Hu, Pengwei Jin, Shaohui Peng, Yuanbo Wen, Rui Zhang, Zidong Du, Qi Guo, Yewen Pu, et al. 2023. ANPL: Compiling Natural Programs with Interactive Decomposition. *arXiv preprint arXiv:2305.18498* (2023).
 - [115] Qing Huang, Yanbang Sun, Zhenchang Xing, Min Yu, Xiwei Xu, and Qinghua Lu. 2023. API Entity and Relation Joint Extraction from Text via Dynamic Prompt-tuned Language Model. *arXiv preprint arXiv:2301.03987* (2023).
 - [116] Qiao Huang, Xin Xia, Zhenchang Xing, David Lo, and Xinyu Wang. 2018. API method recommendation without worrying about the task-API knowledge gap. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 293–304.
 - [117] Qing Huang, Jiahui Zhu, Zhenchang Xing, Huan Jin, Changjing Wang, and Xiwei Xu. 2023. A Chain of AI-based Solutions for Resolving FQNs and Fixing Syntax Errors in Partial Code. *arXiv preprint arXiv:2306.11981* (2023).
 - [118] Qing Huang, Zhou Zou, Zhenchang Xing, Zhenkang Zuo, Xiwei Xu, and Qinghua Lu. 2023. AI Chain on Large Language Model for Unsupervised Control Flow Graph Generation for Statically-Typed Partial Code. *arXiv preprint arXiv:2306.00757* (2023).
 - [119] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436* (2019).
 - [120] Ali Reza Ibrahimzade, Yigit Varli, Dilara Tekinoglu, and Reyhaneh Jabbarvand. 2022. Perfect is the enemy of test oracle. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 70–81.
 - [121] Md Rakibul Islam and Minhaz F Zibran. 2017. Leveraging automated sentiment analysis in software engineering. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. IEEE, 203–214.
 - [122] Haruna Isotani, Hironori Washizaki, Yoshiaki Fukazawa, Tsutomu Nomoto, Saori Ouji, and Shinobu Saito. 2021. Duplicate bug report detection by using sentence embedding and fine-tuning. In *2021 IEEE international conference on software maintenance and evolution (ICSME)*. IEEE, 535–544.
 - [123] Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, and Luke Zettlemoyer. 2016. Summarizing source code using a neural attention model. In *54th Annual Meeting of the Association for Computational Linguistics 2016*. Association for Computational Linguistics, 2073–2083.
 - [124] Maliheh Izadi, Roberta Gismondi, and Georgios Gousios. 2022. Codefill: Multi-token code completion by jointly learning from structure and naming sequences. In *Proceedings of the 44th International Conference on Software Engineering*. 401–412.
 - [125] Naman Jain, Skanda Vaidyanath, Arun Iyer, Nagarajan Natarajan, Suresh Parthasarathy, Sriram Rajamani, and Rahul Sharma. 2022. Jigsaw: Large language models meet program synthesis. In *Proceedings of the 44th International Conference on Software Engineering*. 1219–1231.

- [126] Prithwish Jana, Piyush Jha, Haoyang Ju, Gautham Kishore, Aryan Mahajan, and Vijay Ganesh. 2023. Attention, Compilation, and Solver-based Symbolic Analysis are All You Need. *arXiv preprint arXiv:2306.06755* (2023).
- [127] Kevin Jesse, Premkumar T Devanbu, and Anand Sawant. 2022. Learning to predict user-defined types. *IEEE Transactions on Software Engineering* 49, 4 (2022), 1508–1522.
- [128] Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. SelfEvolve: A Code Evolution Framework via Large Language Models. *arXiv preprint arXiv:2306.02907* (2023).
- [129] Xue Jiang, Yihong Dong, Lecheng Wang, Qiwei Shang, and Ge Li. 2023. Self-planning code generation with large language model. *arXiv preprint arXiv:2303.06689* (2023).
- [130] Yanjie Jiang, Hui Liu, Jiahao Jin, and Lu Zhang. 2020. Automated expansion of abbreviations based on semantic relation and transfer expansion. *IEEE Transactions on Software Engineering* 48, 2 (2020), 519–537.
- [131] Matthew Jin, Syed Shahriar, Michele Tufano, Xin Shi, Shuai Lu, Neel Sundaresan, and Alexey Svyatkovskiy. 2023. Inferfix: End-to-end program repair with llms. *arXiv preprint arXiv:2303.07263* (2023).
- [132] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35 (2022), 11785–11799.
- [133] Robbert Jongeling, Subhajit Datta, and Alexander Serebrenik. 2015. Choosing your weapons: On sentiment analysis tools for software engineering research. In *2015 IEEE international conference on software maintenance and evolution (ICSME)*. IEEE, 531–535.
- [134] Judini. 2023. The future of software development powered by AI. <https://codegpt.co/>.
- [135] Aditya Kanade, Petros Maniatis, Gogul Balakrishnan, and Kensen Shi. 2020. Learning and evaluating contextual embedding of source code. In *International conference on machine learning*. PMLR, 5110–5121.
- [136] Sungmin Kang, Bei Chen, Shin Yoo, and Jian-Guang Lou. 2023. Explainable Automated Debugging via Large Language Model-driven Scientific Debugging. *arXiv preprint arXiv:2304.02195* (2023).
- [137] Sungmin Kang, Juyeon Yoon, and Shin Yoo. 2022. Large language models are few-shot testers: Exploring llm-based general bug reproduction. *arXiv preprint arXiv:2209.11515* (2022).
- [138] Rafael-Michael Karampatsis and Charles Sutton. 2020. Scelmo: Source code embeddings from language models. *arXiv preprint arXiv:2004.13214* (2020).
- [139] Li Ke, Hong Sheng, Fu Cai, Zhang Yunhe, and Liu Ming. 2023. Discriminating Human-authored from ChatGPT-Generated Code Via Discernable Feature Analysis. *arXiv:2306.14397* [cs.SE]
- [140] Adam Khakhar, Stephen Mell, and Osbert Bastani. 2023. PAC Prediction Sets for Large Language Models of Code. *arXiv preprint arXiv:2302.08703* (2023).
- [141] Junaed Younus Khan, Md Tawkat Islam Khondaker, Gias Uddin, and Anindya Iqbal. 2021. Automatic detection of five api documentation smells: Practitioners’ perspectives. In *2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 318–329.
- [142] Junaed Younus Khan and Gias Uddin. 2022. Automatic detection and analysis of technical debts in peer-review documentation of r packages. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 765–776.
- [143] Mohammad Abdullah Matin Khan, M Saiful Bari, Xuan Long Do, Weishi Wang, Md Rizwan Parvez, and Shafiq Joty. 2023. xCodeEval: A Large Scale Multilingual Multitask Benchmark for Code Understanding, Generation, Translation and Retrieval. *arXiv preprint arXiv:2303.03004* (2023).
- [144] Ahmed Khanfir, Renzo Degiovanni, Mike Papadakis, and Yves Le Traon. 2023. Efficient Mutation Testing via Pre-Trained Language Models. *arXiv preprint arXiv:2301.03543* (2023).
- [145] Barbara Kitchenham, Stuart Charters, et al. 2007. Guidelines for performing systematic literature reviews in software engineering.
- [146] Barbara Kitchenham, Lech Madeyski, and David Budgen. 2022. SEGRESS: Software engineering guidelines for reporting secondary studies. *IEEE Transactions on Software Engineering* 49, 3 (2022), 1273–1298.
- [147] Eric Knauss, Siv Houmb, Kurt Schneider, Shareeful Islam, and Jan Jürjens. 2011. Supporting requirements engineers in recognising security issues. In *Requirements Engineering: Foundation for Software Quality: 17th International Working Conference, REFSQ 2011, Essen, Germany, March 28-30, 2011. Proceedings* 17. Springer, 4–18.
- [148] Amy J Ko, Brad A Myers, Michael J Coblenz, and Htet Htet Aung. 2006. An exploratory study of how developers seek, relate, and collect relevant information during software maintenance tasks. *IEEE Transactions on software engineering* 32, 12 (2006), 971–987.
- [149] Takashi Koide, Naoki Fukushima, Hiroki Nakano, and Daiki Chiba. 2023. Detecting Phishing Sites Using ChatGPT. *arXiv preprint arXiv:2306.05816* (2023).
- [150] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [151] Kristian Kolthoff, Christian Bartelt, and Simone Paolo Ponzetto. 2023. Data-driven prototyping via natural-language-based GUI retrieval. *Automated Software Engineering* 30, 1 (2023), 13.

- [152] Bonan Kou, Muhao Chen, and Tianyi Zhang. 2023. Automated Summarization of Stack Overflow Posts. *arXiv preprint arXiv:2305.16680* (2023).
- [153] Bonan Kou, Shengmai Chen, Zhijie Wang, Lei Ma, and Tianyi Zhang. 2023. Is Model Attention Aligned with Human Attention? An Empirical Study on Large Language Models for Code Generation. *arXiv preprint arXiv:2306.01220* (2023).
- [154] Amit Kulkarni. 2021. GitHub Copilot AI Is Leaking Functional API Keys. <https://analyticsdrift.com/github-copilot-ai-is-leaking-functional-api-keys/>.
- [155] Kirby Kuznia, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Less is more: Summary of long instructions is better for program synthesis. *arXiv preprint arXiv:2203.08597* (2022).
- [156] Shuvendu K Lahiri, Aaditya Naik, Georgios Sakkas, Piali Choudhury, Curtis von Veh, Madanlal Musuvathi, Jeevana Priya Inala, Chenglong Wang, and Jianfeng Gao. 2022. Interactive code generation via test-driven user-intent formalization. *arXiv preprint arXiv:2208.05950* (2022).
- [157] Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. 2023. DS-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*. PMLR, 18319–18345.
- [158] Márk Lajkó, Viktor Csuvi, and László Vidács. 2022. Towards JavaScript program repair with generative pre-trained transformer (GPT-2). In *Proceedings of the Third International Workshop on Automated Program Repair*. 61–68.
- [159] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [160] Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. *arXiv preprint arXiv:2305.18486* (2023).
- [161] Thanh Le-Cong, Hong Jin Kang, Truong Giang Nguyen, Stefanus Agus Haryono, David Lo, Xuan-Bach D Le, and Quyet Thang Huynh. 2022. Autopruner: transformer-based call graph pruning. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 520–532.
- [162] Thanh Le-Cong, Duc-Minh Luong, Xuan Bach D Le, David Lo, Nhat-Hoa Tran, Bui Quang-Huy, and Quyet-Thang Huynh. 2023. Invalidator: Automated patch correctness assessment via semantic and syntactic reasoning. *IEEE Transactions on Software Engineering* (2023).
- [163] Jaehyung Lee, Kisun Han, and Hwanjo Yu. 2022. A Light Bug Triage Framework for Applying Large Pre-trained Language Model. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–11.
- [164] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).
- [165] Dong Li, Yelong Shen, Ruoming Jin, Yi Mao, Kuan Wang, and Weizhu Chen. 2022. Generation-Augmented Query Expansion For Code Retrieval. *arXiv preprint arXiv:2212.10692* (2022).
- [166] Feng-Lin Li, Jennifer Horkoff, John Mylopoulos, Renata SS Guizzardi, Giancarlo Guizzardi, Alexander Borgida, and Lin Liu. 2014. Non-functional requirements as qualities, with a spice of ontology. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE, 293–302.
- [167] Jingxuan Li, Rui Huang, Wei Li, Kai Yao, and Weiguo Tan. 2021. Toward less hidden cost of code completion with acceptance and ranking models. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 195–205.
- [168] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023. Enabling Programming Thinking in Large Language Models Toward Code Generation. *arXiv preprint arXiv:2305.06599* (2023).
- [169] Jia Li, Zhuo Li, Huangzhao Zhang, Ge Li, Zhi Jin, Xing Hu, and Xin Xia. 2022. Poison Attack and Defense on Deep Source Code Processing Models. <https://doi.org/10.48550/ARXIV.2210.17029>
- [170] Li Li, Tegawendé F Bissyandé, Mike Papadakis, Siegfried Rasthofer, Alexandre Bartel, Damien Octeau, Jacques Klein, and Le Traon. 2017. Static analysis of android apps: A systematic literature review. *Information and Software Technology* 88 (2017), 67–95.
- [171] Lingwei Li, Li Yang, Huaxi Jiang, Jun Yan, Tiejian Luo, Zihan Hua, Geng Liang, and Chun Zuo. 2022. AUGER: automatically generating review comments with pre-training models. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1009–1021.
- [172] Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large Code Generation Models are Better Few-Shot Information Extractors. *arXiv preprint arXiv:2305.05711* (2023).
- [173] Tsz-On Li, Wenxi Zong, Yibo Wang, Haoye Tian, Ying Wang, and Shing-Chi Cheung. 2023. Finding Failure-Inducing Test Cases with ChatGPT. *arXiv preprint arXiv:2304.11686* (2023).
- [174] Xiaonan Li, Yeyun Gong, Yelong Shen, Xipeng Qiu, Hang Zhang, Bolun Yao, Weizhen Qi, Daxin Jiang, Weizhu Chen, and Nan Duan. 2022. CodeRetriever: A Large Scale Contrastive Pre-Training Method for Code Search. In *Proceedings*

- of the 2022 Conference on Empirical Methods in Natural Language Processing. 2898–2910.
- [175] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).
 - [176] Xin-Ye Li, Jiang-Tian Xue, Zheng Xie, and Ming Li. 2023. Think Outside the Code: Brainstorming Boosts Large Language Models in Code Generation. *arXiv preprint arXiv:2305.10679* (2023).
 - [177] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science* 378, 6624 (2022), 1092–1097.
 - [178] Yichen Li, Yintong Huo, Zhihan Jiang, Renyi Zhong, Pinjia He, Yuxin Su, and Michael R Lyu. 2023. Exploring the Effectiveness of LLMs in Automated Logging Generation: An Empirical Study. *arXiv preprint arXiv:2307.05950* (2023).
 - [179] Yao Li, Tao Zhang, Xiapu Luo, Haipeng Cai, Sen Fang, and Dawei Yuan. 2022. Do Pre-trained Language Models Indeed Understand Software Engineering Tasks? *arXiv preprint arXiv:2211.10623* (2022).
 - [180] Zongjie Li, Chaozheng Wang, Zhibo Liu, Haoxuan Wang, Dong Chen, Shuai Wang, and Cuiyun Gao. 2023. Cctest: Testing and repairing code completion systems. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 1238–1250.
 - [181] Zongjie Li, Chaozheng Wang, Zhibo Liu, Haoxuan Wang, Shuai Wang, and Cuiyun Gao. 2022. CCTEST: Testing and Repairing Code Completion Systems. *arXiv preprint arXiv:2208.08289* (2022).
 - [182] Yuding Liang and Kenny Zhu. 2018. Automatic generation of text descriptive comments for code blocks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
 - [183] Jinfeng Lin, Yalin Liu, Qingkai Zeng, Meng Jiang, and Jane Cleland-Huang. 2021. Traceability transformed: Generating more accurate links with pre-trained bert models. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 324–335.
 - [184] Chao Liu, Xuanlin Bao, Hongyu Zhang, Neng Zhang, Haibo Hu, Xiaohong Zhang, and Meng Yan. 2023. Improving ChatGPT Prompt for Code Generation. *arXiv preprint arXiv:2305.08360* (2023).
 - [185] Fang Liu, Ge Li, Yunfei Zhao, and Zhi Jin. 2020. Multi-task learning based pre-trained language model for code completion. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 473–485.
 - [186] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210* (2023).
 - [187] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101* (2016).
 - [188] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
 - [189] Tianyang Liu, Canwen Xu, and Julian McAuley. 2023. RepoBench: Benchmarking Repository-Level Code Auto-Completion Systems. *arXiv preprint arXiv:2306.03091* (2023).
 - [190] Xiaoyu Liu, LiGuo Huang, and Vincent Ng. 2018. Effective API recommendation without historical software repositories. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 282–292.
 - [191] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv:2306.05499* (2023).
 - [192] Yue Liu, Thanh Le-Cong, Ratnadira Widyasari, Chakkrit Tantithamthavorn, Li Li, Xuan-Bach D Le, and David Lo. 2023. Refining ChatGPT-Generated Code: Characterizing and Mitigating Code Quality Issues. *arXiv preprint arXiv:2307.12596* (2023).
 - [193] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
 - [194] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. 2022. Deep learning for android malware defenses: a systematic literature review. *Comput. Surveys* 55, 8 (2022), 1–36.
 - [195] Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. 2021. Codexglue: A machine learning benchmark dataset for code understanding and generation. *arXiv preprint arXiv:2102.04664* (2021).
 - [196] James H Lubowitz. 2023. ChatGPT, an artificial intelligence chatbot, is impacting medical literature. *Arthroscopy* 39, 5 (2023), 1121–1122.
 - [197] Dipeeka Luitel, Shabnam Hassani, and Mehrdad Sabetzadeh. 2023. Improving Requirements Completeness: Automated Assistance through Large Language Models. *arXiv preprint arXiv:2308.03784* (2023).
 - [198] Qingzhou Luo, Farah Hariri, Lamyaa Eloussi, and Darko Marinov. 2014. An empirical analysis of flaky tests. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*. 643–653.

- [199] Xianchang Luo, Yinxing Xue, Zhenchang Xing, and Jiamou Sun. 2022. PRCBERT: Prompt Learning for Requirement Classification using BERT-based Pretrained Language Models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–13.
- [200] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. WizardCoder: Empowering Code Large Language Models with Evol-Instruct. *arXiv preprint arXiv:2306.08568* (2023).
- [201] Wei Ma, Shangqing Liu, Wenhan Wang, Qiang Hu, Ye Liu, Cen Zhang, Liming Nie, and Yang Liu. 2023. The Scope of ChatGPT in Software Engineering: A Thorough Investigation. *arXiv preprint arXiv:2305.12138* (2023).
- [202] Wei Ma, Mengjie Zhao, Xiaofei Xie, Qiang Hu, Shangqing Liu, Jie Zhang, Wenhan Wang, and Yang Liu. 2023. Are Code Pre-trained Models Powerful to Learn Code Syntax and Semantics? *arXiv:2212.10017* [cs.SE]
- [203] Aman Madaan, Alexander Shypula, Uri Alon, Milad Hashemi, Parthasarathy Ranganathan, Yiming Yang, Graham Neubig, and Amir Yazdanbakhsh. 2023. Learning performance-improving code edits. *arXiv preprint arXiv:2302.07867* (2023).
- [204] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128* (2022).
- [205] Shantanu Mandal, Adhrik Chethan, Vahid Janfaza, SM Mahmud, Todd A Anderson, Javier Turek, Jesmin Jahan Tithi, and Abdullah Muzahid. 2023. Large Language Models Based Automatic Synthesis of Software Specifications. *arXiv preprint arXiv:2304.09181* (2023).
- [206] Dung Nguyen Manh, Nam Le Hai, Anh TV Dau, Anh Minh Nguyen, Khanh Nghiem, Jin Guo, and Nghi DQ Bui. 2023. The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation. *arXiv preprint arXiv:2305.06156* (2023).
- [207] Zohar Manna and Richard Waldinger. 1980. A deductive approach to program synthesis. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 2, 1 (1980), 90–121.
- [208] Andrea Marcelli, Mariano Graziano, Xabier Ugarte-Pedrero, Yanick Fratantonio, Mohamad Mansouri, and Davide Balzarotti. 2022. How machine learning is solving the binary function similarity problem. In *31st USENIX Security Symposium (USENIX Security 22)*. 2099–2116.
- [209] Antonio Mastropaolo, Emad Aghajani, Luca Pascarella, and Gabriele Bavota. 2021. An empirical study on code comment completion. In *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 159–170.
- [210] Antonio Mastropaolo, Nathan Cooper, David Nader Palacio, Simone Scalabrino, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. 2022. Using transfer learning for code-related tasks. *IEEE Transactions on Software Engineering* 49, 4 (2022), 1580–1598.
- [211] Antonio Mastropaolo, Luca Pascarella, and Gabriele Bavota. 2022. Using deep learning to generate complete log statements. In *Proceedings of the 44th International Conference on Software Engineering*. 2279–2290.
- [212] Antonio Mastropaolo, Simone Scalabrino, Nathan Cooper, David Nader Palacio, Denys Poshyvanyk, Rocco Oliveto, and Gabriele Bavota. 2021. Studying the usage of text-to-text transfer transformer to support code-related tasks. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 336–347.
- [213] Meta. 2023. Code Llama: Open Foundation Models for Code. <https://ai.meta.com/research/publications/code-llama-open-foundation-models-for-code/>.
- [214] Microsoft. 2023. Bing Chat. <https://www.microsoft.com/en-us/edge/features/bing-chat>.
- [215] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. 2020. Understanding the role of training regimes in continual learning. *Advances in Neural Information Processing Systems* 33 (2020), 7308–7320.
- [216] Ambarish Moharil and Arpit Sharma. 2022. Identification of intra-domain ambiguity using transformer-based machine learning. In *Proceedings of the 1st International Workshop on Natural Language-based Software Engineering*. 51–58.
- [217] Ambarish Moharil and Arpit Sharma. 2023. TABASCO: A Transformer Based Contextualization Toolkit. *Science of Computer Programming* (2023), 102994.
- [218] Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*. 220–224.
- [219] Sebastian Moss. 2021. Google Brain unveils trillion-parameter AI language model, the largest yet. <https://aibusiness.com/nlp/google-brain-unveils-trillion-parameter-ai-language-model-the-largest-yet>.
- [220] Manisha Mukherjee and Vincent J Hellendoorn. 2023. Stack Over-Flowing with Results: The Case for Domain-Specific Pre-Training Over One-Size-Fits-All Models. *arXiv preprint arXiv:2306.03268* (2023).
- [221] Vijayaraghavan Murali, Chandra Maddila, Imad Ahmad, Michael Bolin, Daniel Cheng, Negar Ghorbani, Renuka Fernandez, and Nachiappan Nagappan. 2023. CodeCompose: A Large-Scale Industrial Deployment of AI-assisted Code Authoring. *arXiv preprint arXiv:2305.12050* (2023).
- [222] Ramin Nafisi. 2021. GoldMax, GoldFinder, and Sibot: Analyzing NOBELIUM’s layered persistence. <https://www.microsoft.com/en-us/security/blog/2021/03/04/goldmax-goldfinder-sibot-analyzing-nobelium-malware/>.

- [223] Ramin Nafisi and Andrea Lelli. 2021. GoldMax, GoldFinder, and Sibot: Analyzing NOBELIUM's Layered Persistence. Forrás: <https://microsoft.com> [Accessed: 2022.05.14].
- [224] Stefan Nagy, Anh Nguyen-Tuong, Jason D Hiser, Jack W Davidson, and Matthew Hicks. 2021. Breaking through binaries: Compiler-quality instrumentation for better binary-only fuzzing. In *30th USENIX Security Symposium (USENIX Security 21)*. 1683–1700.
- [225] Nathalia Nascimento, Paulo Alencar, and Donald Cowan. 2023. Comparing Software Developers with ChatGPT: An Empirical Investigation. *arXiv preprint arXiv:2305.11837* (2023).
- [226] Muhammad U Nasir, Sam Earle, Julian Togelius, Steven James, and Christopher Cleghorn. 2023. LLMatic: Neural Architecture Search via Large Language Models and Quality-Diversity Optimization. *arXiv preprint arXiv:2306.01102* (2023).
- [227] Anh Tuan Nguyen, Michael Hilton, Mihai Codoban, Hoan Anh Nguyen, Lily Mast, Eli Rademacher, Tien N Nguyen, and Danny Dig. 2016. API code recommendation using statistical learning from fine-grained changes. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. 511–522.
- [228] Anh Tuan Nguyen and Tien N Nguyen. 2017. Automatic categorization with deep neural network for open-source java projects. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*. IEEE, 164–166.
- [229] Phuong T Nguyen, Juri Di Rocco, Claudio Di Sipio, Riccardo Rubel, Davide Di Ruscio, and Massimiliano Di Penta. 2023. Is this Snippet Written by ChatGPT? An Empirical Study with a CodeBERT-Based Classifier. *arXiv preprint arXiv:2307.09381* (2023).
- [230] Liming Nie, He Jiang, Zhilei Ren, Zeyi Sun, and Xiaochen Li. 2016. Query expansion based on crowd knowledge for code search. *IEEE Transactions on Services Computing* 9, 5 (2016), 771–783.
- [231] Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309* (2023).
- [232] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474* (2022).
- [233] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. A conversational paradigm for program synthesis.
- [234] Changan Niu, Chuanyi Li, Vincent Ng, Jidong Ge, Liguo Huang, and Bin Luo. 2022. Spt-code: Sequence-to-sequence pre-training for learning source code representations. In *Proceedings of the 44th International Conference on Software Engineering*. 2006–2018.
- [235] Marcel Ochs, Krishna Narasimhan, and Mira Mezini. 2023. Evaluating and improving transformers pre-trained on ASTs for Code Completion. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 834–844.
- [236] Theo X Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. 2023. Demystifying GPT Self-Repair for Code Generation. *arXiv preprint arXiv:2306.09896* (2023).
- [237] Philippe Ombredanne. 2020. Free and open source software license compliance: tools for software composition analysis. *Computer* 53, 10 (2020), 105–109.
- [238] OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <https://chat.openai.com>.
- [239] OpenAI. 2022. GPT-3.5. <https://platform.openai.com/docs/models/gpt-3-5>.
- [240] OpenAI. 2023. Code Interpreter. <https://openai.com/blog/chatgpt-plugins#code-interpreter>.
- [241] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774* [cs.CL]
- [242] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [243] Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828* (2023).
- [244] Stack Overflow. 2023. Stack Overflow. <https://stackoverflow.com/>.
- [245] Rangeet Pan, Ali Reza Ibrahimzada, Rahul Krishna, Divya Sankar, Lambert Pouguem Wassi, Michele Merler, Boris Sobolev, Raju Pavuluri, Saurabh Sinha, and Reyhaneh Jabbarvand. 2023. Understanding the Effectiveness of Large Language Models in Code Translation. *arXiv preprint arXiv:2308.03109* (2023).
- [246] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *arXiv preprint arXiv:2306.08302* (2023).
- [247] Emilio Parisotto, Abdel-rahman Mohamed, Rishabh Singh, Lihong Li, Dengyong Zhou, and Pushmeet Kohli. 2016. Neuro-symbolic program synthesis. *arXiv preprint arXiv:1611.01855* (2016).
- [248] Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. 2023. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334* (2023).

- [249] Rishov Paul, Md Mohib Hossain, Masum Hasan, and Anindya Iqbal. 2023. Automated Program Repair Based on Code Review: How do Pre-trained Transformer Models Perform? *arXiv preprint arXiv:2304.07840* (2023).
- [250] Rishov Paul, Md. Mohib Hossain, Mohammed Latif Siddiq, Masum Hasan, Anindya Iqbal, and Joanna C. S. Santos. 2023. Enhancing Automated Program Repair through Fine-tuning and Prompt Engineering. *arXiv:2304.07840 [cs.LG]*
- [251] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2021. Examining zero-shot vulnerability repair with large language models. *arXiv preprint arXiv:2112.02125* (2021).
- [252] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2023. Examining zero-shot vulnerability repair with large language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2339–2356.
- [253] Tommaso Pegolotti, Elias Frantar, Dan Alistarh, and Markus Püschel. 2023. QIGen: Generating Efficient Kernels for Quantized Inference on Large Language Models. *arXiv preprint arXiv:2307.03738* (2023).
- [254] Kexin Pei, Zhou Xuan, Junfeng Yang, Suman Jana, and Baishakhi Ray. 2020. Trex: Learning execution semantics from micro-traces for binary similarity. *arXiv preprint arXiv:2012.08680* (2020).
- [255] Long Phan, Hieu Tran, Daniel Le, Hieu Nguyen, James Anibal, Alec Peltekian, and Yanfang Ye. 2021. Cotext: Multi-task learning with code-text transformer. *arXiv preprint arXiv:2105.08645* (2021).
- [256] Benjamin C Pierce and David N Turner. 2000. Local type inference. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 22, 1 (2000), 1–44.
- [257] Julian Aron Prenner and Romain Robbes. 2021. Making the most of small Software Engineering datasets with modern machine learning. *IEEE Transactions on Software Engineering* 48, 12 (2021), 5050–5067.
- [258] Rohith Pudari and Neil A Ernst. 2023. From Copilot to Pilot: Towards AI Supported Software Development. *arXiv preprint arXiv:2303.04142* (2023).
- [259] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. *arXiv preprint arXiv:2307.07924* (2023).
- [260] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [261] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [262] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
- [263] Sajjad Rahmani, AmirHossein Naghshzan, and Latifa Guerrouj. 2023. Improving Code Example Recommendations on Informal Documentation Using BERT and Query-Aware LSH: A Comparative Study. *arXiv preprint arXiv:2305.03017* (2023).
- [264] Aurora Ramirez, Jose Raul Romero, and Christopher L Simons. 2018. A systematic review of interaction in search-based software engineering. *IEEE Transactions on Software Engineering* 45, 8 (2018), 760–781.
- [265] Sami Ramly. 2023. Preventing Abuse of LLMs’ Alignment Deficit by Injection Neutralization (PALADIN). <https://medium.com/@SamiRamly/prompt-attacks-are-llm-jailbreaks-inevitable-f7848cc11122>.
- [266] Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya, and Monojit Choudhury. 2023. Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks. *arXiv preprint arXiv:2305.14965* (2023).
- [267] Nikitha Rao, Jason Tsay, Kiran Kate, Vincent J Hellendoorn, and Martin Hirzel. 2023. AI for Low-Code for AI. *arXiv preprint arXiv:2305.20015* (2023).
- [268] Veselin Raychev, Martin Vechev, and Eran Yahav. 2014. Code completion with statistical language models. In *Proceedings of the 35th ACM SIGPLAN conference on programming language design and implementation*. 419–428.
- [269] Leanna Rierson. 2017. *Developing safety-critical software: a practical guide for aviation software and DO-178C compliance*. CRC Press.
- [270] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology* 57, 9 (2023), 3464–3466.
- [271] Martin P Robillard. 2009. What makes APIs hard to learn? Answers from developers. *IEEE software* 26, 6 (2009), 27–34.
- [272] Martin P Robillard and Robert DeLine. 2011. A field study of API learning obstacles. *Empirical Software Engineering* 16 (2011), 703–732.
- [273] Tobias Roehm, Rebecca Tiarks, Rainer Koschke, and Walid Maalej. 2012. How do professional developers comprehend software?. In *2012 34th International Conference on Software Engineering (ICSE)*. IEEE, 255–265.
- [274] Baptiste Roziere, Marie-Anne Lachaux, Marc Szafraniec, and Guillaume Lample. 2021. Dobf: A deobfuscation pre-training objective for programming languages. *arXiv preprint arXiv:2102.07492* (2021).
- [275] Iman Saberi, Fatemeh Fard, and Fuxiang Chen. 2023. Multilingual Adapter-based Knowledge Aggregation on Code Summarization for Low-Resource Languages. *arXiv preprint arXiv:2307.07854* (2023).

- [276] Iman Saberi, Fatemeh Fard, and Fuxiang Chen. 2023. Utilization of Pre-trained Language Model for Adapter-based Knowledge Transfer in Software Engineering. *arXiv preprint arXiv:2307.08540* (2023).
- [277] Ahmed Sadik, Antonello Ceravola, Frank Joublin, and Jibesh Patra. 2023. Analysis of ChatGPT on Source Code. *arXiv preprint arXiv:2306.00597* (2023).
- [278] Anthony Saieva, Saikat Chakraborty, and Gail Kaiser. 2023. On Contrastive Learning of Semantic Similarity for Code to Code Search. *arXiv preprint arXiv:2305.03843* (2023).
- [279] Fardin Ahsan Sakib, Saadat Hasan Khan, and AHM Karim. 2023. Extending the Frontier of ChatGPT: Code Generation and Debugging. *arXiv preprint arXiv:2307.08260* (2023).
- [280] Pasquale Salza, Christoph Schwizer, Jian Gu, and Harald C Gall. 2022. On the effectiveness of transfer learning for code search. *IEEE Transactions on Software Engineering* (2022).
- [281] Mahadev Satyanarayanan, David C Steere, Masashi Kudo, and Hank Mashburn. 1992. Transparent logging as a technique for debugging complex distributed systems. In *Proceedings of the 5th workshop on ACM SIGOPS European workshop: Models and paradigms for distributed systems structuring*. 1–3.
- [282] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100* (2022).
- [283] Max Schäfer, Sarah Nadi, Aryaz Eghbali, and Frank Tip. 2023. Adaptive test generation using a large language model. *arXiv preprint arXiv:2302.06527* (2023).
- [284] Imanol Schlag, Sainbayar Sukhbaatar, Asli Celikyilmaz, Wen tau Yih, Jason Weston, Jürgen Schmidhuber, and Xian Li. 2023. Large Language Model Programs. *arXiv:2305.05364* [cs.LG]
- [285] Oussama Ben Sghaier and Houari Sahraoui. 2023. A Multi-Step Learning Approach to Assist Code Review. In *2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 450–460.
- [286] Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551* (2022).
- [287] Rishab Sharma, Fuxiang Chen, Fatemeh Fard, and David Lo. 2022. An exploratory study on code attention in BERT. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. 437–448.
- [288] Da Shen, Xinyun Chen, Chenguang Wang, Koushik Sen, and Dawn Song. 2022. Benchmarking Language Models for Code Syntax Understanding. *arXiv preprint arXiv:2210.14473* (2022).
- [289] Jieke Shi, Zhou Yang, Bowen Xu, Hong Jin Kang, and David Lo. 2023. Compressing Pre-Trained Models of Code into 3 MB (ASE '22). Association for Computing Machinery, New York, NY, USA, Article 24, 12 pages. <https://doi.org/10.1145/3551349.3556964>
- [290] Zejian Shi, Yun Xiong, Xiaolong Zhang, Yao Zhang, Shanshan Li, and Yangyong Zhu. 2022. Cross-Modal Contrastive Learning for Code Search. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 94–105.
- [291] Atsushi Shirafuji, Yutaka Watanobe, Takumi Ito, Makoto Morishita, Yuki Nakamura, Yusuke Oda, and Jun Suzuki. 2023. Exploring the Robustness of Large Language Models for Solving Programming Problems. *arXiv preprint arXiv:2306.14583* (2023).
- [292] Mohammed Latif Siddiq, Beatrice Casey, and Joanna Santos. 2023. A Lightweight Framework for High-Quality Code Generation. *arXiv preprint arXiv:2307.08220* (2023).
- [293] Mohammed Latif Siddiq, Joanna Santos, Ridwanul Hasan Tanvir, Noshin Ulfat, Fahmid Al Rifat, and Vinicius Carvalho Lopes. 2023. Exploring the Effectiveness of Large Language Models in Generating Unit Tests. *arXiv preprint arXiv:2305.00418* (2023).
- [294] Adish Singla. 2023. Evaluating ChatGPT and GPT-4 for Visual Programming. *arXiv preprint arXiv:2308.02522* (2023).
- [295] Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653* (2023).
- [296] Giriprasad Sridhara, Sourav Mazumdar, et al. 2023. ChatGPT: A Study on its Utility for Ubiquitous Software Engineering Tasks. *arXiv preprint arXiv:2305.16837* (2023).
- [297] Saurabh Srivastava, Sumit Gulwani, and Jeffrey S Foster. 2010. From program verification to program synthesis. In *Proceedings of the 37th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. 313–326.
- [298] Yuqiang Sun, Daoyuan Wu, Yue Xue, Han Liu, Haijun Wang, Zhengzi Xu, Xiaofei Xie, and Yang Liu. 2023. When GPT Meets Program Analysis: Towards Intelligent Detection of Smart Contract Logic Vulnerabilities in GPTScan. *arXiv preprint arXiv:2308.03314* (2023).
- [299] Zhensu Sun, Li Li, Yan Liu, Xiaoning Du, and Li Li. 2022. On the importance of building high-quality training datasets for neural code search. In *Proceedings of the 44th International Conference on Software Engineering*. 1609–1620.
- [300] Jeffrey Svajlenko, Judith F Islam, Iman Keivanloo, Chanchal K Roy, and Mohammad Mamun Mia. 2014. Towards a big data curated benchmark of inter-project code clones. In *2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE, 476–480.

- [301] Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. 2020. Code and named entity recognition in stackoverflow. *arXiv preprint arXiv:2005.01634* (2020).
- [302] Chee Wei Tan, Shangxin Guo, Man Fai Wong, and Ching Nam Hang. 2023. Copilot for Xcode: Exploring AI-Assisted Programming by Prompting Cloud-based Large Language Models. *arXiv preprint arXiv:2307.14349* (2023).
- [303] Wei Tang, Mingwei Tang, Minchao Ban, Ziguo Zhao, and Mingjun Feng. 2023. CSGVD: A deep learning approach combining sequence and graph embedding for source code vulnerability detection. *Journal of Systems and Software* 199 (2023), 111623.
- [304] Yutian Tang, Zhijie Liu, Zhichao Zhou, and Xiapu Luo. 2023. ChatGPT vs SBST: A Comparative Assessment of Unit Test Suite Generation. *arXiv preprint arXiv:2307.00588* (2023).
- [305] Artur Tarassow. 2023. The potential of LLMs for coding with low-resource and domain-specific programming languages. *arXiv preprint arXiv:2307.13018* (2023).
- [306] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).
- [307] Shailja Thakur, Baleegh Ahmad, Hammond Pearce, Benjamin Tan, Brendan Dolan-Gavitt, Ramesh Karri, and Siddharth Garg. 2023. VeriGen: A Large Language Model for Verilog Code Generation. *arXiv preprint arXiv:2308.00708* (2023).
- [308] Chandra Thapa, Seung Ick Jang, Muhammad Ejaz Ahmed, Seyit Camtepe, Josef Pieprzyk, and Surya Nepal. 2022. Transformer-based language models for software vulnerability detection. In *Proceedings of the 38th Annual Computer Security Applications Conference*. 481–496.
- [309] Haoye Tian, Kui Liu, Abdoul Kader Kaboré, Anil Koyuncu, Li Li, Jacques Klein, and Tegawendé F Bissyandé. 2020. Evaluating representation learning of code changes for predicting patch correctness in program repair. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*. 981–992.
- [310] Haoye Tian, Kui Liu, Yinghua Li, Abdoul Kader Kaboré, Anil Koyuncu, Andrew Habib, Li Li, Junhao Wen, Jacques Klein, and Tegawendé F Bissyandé. 2023. The Best of Both Worlds: Combining Learned Embeddings with Engineered Features for Accurate Prediction of Correct Patches. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–34.
- [311] Haoye Tian, Weiqi Lu, Tsz On Li, Xunzhu Tang, Shing-Chi Cheung, Jacques Klein, and Tegawendé F Bissyandé. 2023. Is ChatGPT the Ultimate Programming Assistant—How far is it? *arXiv preprint arXiv:2304.11938* (2023).
- [312] Norbert Tihanyi, Tamas Bisztray, Ridhi Jain, Mohamed Amine Ferrag, Lucas C Cordeiro, and Vasileios Mavroeidis. 2023. The FormAI Dataset: Generative AI in Software Security Through the Lens of Formal Verification. *arXiv preprint arXiv:2307.02192* (2023).
- [313] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [314] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutli Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [315] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.
- [316] Haoxin Tu, Zhide Zhou, He Jiang, Imam Nur Bani Yusuf, Yuxian Li, and Lingxiao Jiang. 2023. LLM4CBI: Taming LLMs to Generate Effective Test Programs for Compiler Bug Isolation. *arXiv preprint arXiv:2307.00593* (2023).
- [317] Michele Tufano, Shubham Chandel, Anisha Agarwal, Neel Sundaresan, and Colin Clement. 2023. Predicting Code Coverage without Execution. *arXiv preprint arXiv:2307.13383* (2023).
- [318] Rosalia Tufano, Simone Masiero, Antonio Mastropaolo, Luca Pascarella, Denys Poshyvanyk, and Gabriele Bavota. 2022. Using pre-trained models to boost code review automation. In *Proceedings of the 44th International Conference on Software Engineering*. 2291–2302.
- [319] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [320] Vasudev Vikram, Caroline Lemieux, and Rohan Padhye. 2023. Can Large Language Models Write Good Property-Based Tests? *arXiv preprint arXiv:2307.04346* (2023).
- [321] Julian Von der Mosel, Alexander Trautsch, and Steffen Herbold. 2022. On the validity of pre-trained transformers for natural language processing in the software engineering domain. *IEEE Transactions on Software Engineering* 49, 4 (2022), 1487–1507.
- [322] Yao Wan, Jingdong Shu, Yulei Sui, Guandong Xu, Zhou Zhao, Jian Wu, and Philip Yu. 2019. Multi-modal attention network learning for semantic source code retrieval. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 13–25.

- [323] Yao Wan, Shijie Zhang, Hongyu Zhang, Yulei Sui, Guandong Xu, Dezhong Yao, Hai Jin, and Lichao Sun. 2022. You See What I Want You to See: Poisoning Vulnerabilities in Neural Code Search. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (Singapore, Singapore) (ESEC/FSE 2022). Association for Computing Machinery, New York, NY, USA, 1233–1245. <https://doi.org/10.1145/3540250.3549153>
- [324] Yao Wan, Wei Zhao, Hongyu Zhang, Yulei Sui, Guandong Xu, and Hai Jin. 2022. What do they capture? a structural analysis of pre-trained language models for source code. In *Proceedings of the 44th International Conference on Software Engineering*. 2377–2388.
- [325] Yao Wan, Zhou Zhao, Min Yang, Guandong Xu, Haochao Ying, Jian Wu, and Philip S Yu. 2018. Improving automatic source code summarization via deep reinforcement learning. In *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*. 397–407.
- [326] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model.
- [327] Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R Lyu. 2022. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 382–394.
- [328] Junjie Wang, Yuchao Huang, Chunyang Chen, Zhe Liu, Song Wang, and Qing Wang. 2023. Software Testing with Large Language Model: Survey, Landscape, and Vision. *arXiv preprint arXiv:2307.07221* (2023).
- [329] Jian Wang, Shangqing Liu, Xiaofei Xie, and Yi Li. 2023. Evaluating AIGC Detectors on Code Content. *arXiv preprint arXiv:2304.05193* (2023).
- [330] Simin Wang, Liguang Huang, Amiao Gao, Jidong Ge, Tengfei Zhang, Haitao Feng, Ishna Satyarth, Ming Li, He Zhang, and Vincent Ng. 2022. Machine/deep learning for software engineering: A systematic literature review. *IEEE Transactions on Software Engineering* 49, 3 (2022), 1188–1231.
- [331] Shufan Wang, Sebastien Jean, Sailik Sengupta, James Gung, Nikolaos Pappas, and Yi Zhang. 2023. Measuring and Mitigating Constraint Violations of In-Context Learning for Utterance-to-API Semantic Parsing. *arXiv preprint arXiv:2305.15338* (2023).
- [332] Shiqi Wang, Zheng Li, Haifeng Qian, Chenghao Yang, Zijian Wang, Mingyue Shang, Varun Kumar, Samson Tan, Baishakhi Ray, Parminder Bhatia, et al. 2022. ReCode: Robustness Evaluation of Code Generation Models. *arXiv preprint arXiv:2212.10264* (2022).
- [333] Wenhan Wang, Ge Li, Bo Ma, Xin Xia, and Zhi Jin. 2020. Detecting code clones with graph neural network and flow-augmented abstract syntax tree. In *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 261–271.
- [334] Wenhan Wang, Ge Li, Sijie Shen, Xin Xia, and Zhi Jin. 2020. Modular tree network for source code representation learning. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 29, 4 (2020), 1–23.
- [335] Xingyao Wang, Hao Peng, Reyhaneh Jabbarvand, and Heng Ji. 2023. LeTI: Learning to Generate from Textual Interactions. *arXiv preprint arXiv:2305.10314* (2023).
- [336] Xin Wang, Yasheng Wang, Fei Mi, Pingyi Zhou, Yao Wan, Xiao Liu, Li Li, Hao Wu, Jin Liu, and Xin Jiang. 2021. Syncobert: Syntax-guided multi-modal contrastive pre-training for code representation. *arXiv preprint arXiv:2108.04556* (2021).
- [337] Yue Wang, Hung Le, Akhilesh Deepak Gotmare, Nghi DQ Bui, Junnan Li, and Steven CH Hoi. 2023. Codet5+: Open code large language models for code understanding and generation. *arXiv preprint arXiv:2305.07922* (2023).
- [338] Yawen Wang, Lin Shi, Mingyang Li, Qing Wang, and Yun Yang. 2020. A deep context-wise method for coreference detection in natural language requirements. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 180–191.
- [339] Yawen Wang, Junjie Wang, Hongyu Zhang, Xuran Ming, Lin Shi, and Qing Wang. 2022. Where is your app frustrating users?. In *Proceedings of the 44th International Conference on Software Engineering*. 2427–2439.
- [340] Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. 2021. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. *arXiv preprint arXiv:2109.00859* (2021).
- [341] Huihui Wei and Ming Li. 2017. Supervised deep features for software functional clone detection by exploiting lexical and syntactical information in source code.. In *IJCAI*. 3034–3040.
- [342] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [343] Moshi Wei, Nima Shiri Harzevili, Yuchao Huang, Junjie Wang, and Song Wang. 2022. Clear: contrastive learning for api recommendation. In *Proceedings of the 44th International Conference on Software Engineering*. 376–387.
- [344] Martin Weyssow, Xin Zhou, Kisub Kim, David Lo, and Houari Sahraoui. 2023. On the Usage of Continual Learning for Out-of-Distribution Generalization in Pre-trained Language Models of Code. *arXiv preprint arXiv:2305.04106* (2023).

- [345] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).
- [346] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint arXiv:2303.07839* (2023).
- [347] Patricia Widjojo and Christoph Treude. 2023. Addressing Compiler Errors: Stack Overflow or Large Language Models? *arXiv preprint arXiv:2307.10793* (2023).
- [348] Man-Fai Wong, Shangxin Guo, Ching-Nam Hang, Siu-Wai Ho, and Chee-Wei Tan. 2023. Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review. *Entropy* 25, 6 (2023), 888.
- [349] Fangzhao Wu, Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, and Xing Xie. 2023. Defending ChatGPT against Jailbreak Attack via Self-Reminder. (2023).
- [350] Tongshuang Wu, Kenneth Koedinger, et al. 2023. Is AI the better programming partner? Human-Human Pair Programming vs. Human-AI pAIr Programming. *arXiv preprint arXiv:2306.05153* (2023).
- [351] Yi Wu, Nan Jiang, Hung Viet Pham, Thibaud Lutellier, Jordan Davis, Lin Tan, Petr Babkin, and Sameena Shah. 2023. How Effective Are Neural Networks for Fixing Security Vulnerabilities. *arXiv preprint arXiv:2305.18607* (2023).
- [352] Chunqiu Steven Xia, Yuxiang Wei, and Lingming Zhang. 2022. Practical program repair in the era of large pre-trained language models. *arXiv preprint arXiv:2210.14179* (2022).
- [353] Chunqiu Steven Xia, Yuxiang Wei, and Lingming Zhang. 2023. Automated Program Repair in the Era of Large Pre-Trained Language Models. In *Proceedings of the 45th International Conference on Software Engineering (ICSE '23)*. <https://doi.org/10.1109/ICSE48619.2023.00129>
- [354] Chunqiu Steven Xia and Lingming Zhang. 2023. Conversational automated program repair. *arXiv preprint arXiv:2301.13246* (2023).
- [355] Chunqiu Steven Xia and Lingming Zhang. 2023. Keep the Conversation Going: Fixing 162 out of 337 bugs for 0.42 each using ChatGPT. *arXiv preprint arXiv:2304.00385* (2023).
- [356] Zhuokui Xie, Yinghao Chen, Chen Zhi, Shuiguang Deng, and Jianwei Yin. 2023. ChatUniTest: a ChatGPT-based automated unit test generation tool. *arXiv preprint arXiv:2305.04764* (2023).
- [357] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 1–10.
- [358] Xiangzhe Xu, Zhuo Zhang, Shiwei Feng, Yapeng Ye, Zian Su, Nan Jiang, Siyuan Cheng, Lin Tan, and Xiangyu Zhang. 2023. LmPa: Improving Decompilation by Synergy of Large Language Model and Program Analysis. *arXiv preprint arXiv:2306.02546* (2023).
- [359] Chengran Yang, Bowen Xu, Junaed Younus Khan, Gias Uddin, Donggyun Han, Zhou Yang, and David Lo. 2022. Aspect-based api review classification: How far can pre-trained transformer model go?. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. IEEE, 385–395.
- [360] Di Yang, Aftab Hussain, and Cristina Videira Lopes. 2016. From query to usable code: an analysis of stack overflow code snippets. In *Proceedings of the 13th International Conference on Mining Software Repositories*. 391–402.
- [361] Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Yiran Xu, Tingting Han, and Taolue Chen. 2023. A Syntax-Guided Multi-Task Learning Approach for Turducken-Style Code Generation. *arXiv preprint arXiv:2303.05061* (2023).
- [362] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712* (2023).
- [363] Lanxin Yang, He Zhang, Haifeng Shen, Xin Huang, Xin Zhou, Guoping Rong, and Dong Shao. 2021. Quality assessment in systematic literature reviews: A software engineering perspective. *Information and Software Technology* 130 (2021), 106397.
- [364] Yanming Yang, Xin Xia, David Lo, and John Grundy. 2022. A survey on deep learning for software engineering. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–73.
- [365] Zhou Yang, Jieke Shi, Junda He, and David Lo. 2022. Natural Attack for Pre-Trained Models of Code. In *Proceedings of the 44th International Conference on Software Engineering (Pittsburgh, Pennsylvania) (ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 1482–1493. <https://doi.org/10.1145/3510003.3510146>
- [366] Zhou Yang, Bowen Xu, Jie M. Zhang, Hong Jin Kang, Jieke Shi, Junda He, and David Lo. 2023. Stealthy Backdoor Attack for Code Models. <https://doi.org/10.48550/ARXIV.2301.02496>
- [367] Jiacheng Ye, Chengzu Li, Lingpeng Kong, and Tao Yu. 2023. Generating Data for Symbolic Language with Large Language Models. *arXiv preprint arXiv:2305.13917* (2023).
- [368] Burak Yetiştiren, Işık Özsoy, Miray Ayerdem, and Eray Tüzün. 2023. Evaluating the Code Quality of AI-Assisted Code Generation Tools: An Empirical Study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT. *arXiv preprint arXiv:2304.10778* (2023).

- [369] Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. *arXiv preprint arXiv:1704.01696* (2017).
- [370] ymcui. 2023. Chinese LLaMA & Alpaca Large Language Models. https://github.com/ymcui/Chinese-LLaMA-Alpaca-2/blob/main/README_EN.md.
- [371] Hao Yu, Bo Shen, Dezhi Ran, Jiaxin Zhang, Qi Zhang, Yuchi Ma, Guangtai Liang, Ying Li, Tao Xie, and Qianxiang Wang. 2023. CoderEval: A Benchmark of Pragmatic Code Generation with Generative Pre-trained Models. *arXiv preprint arXiv:2302.00288* (2023).
- [372] Wei Yuan, Quanjun Zhang, Tieke He, Chunrong Fang, Nguyen Quoc Viet Hung, Xiaodong Hao, and Hongzhi Yin. 2022. CIRCLE: Continual repair across programming languages. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 678–690.
- [373] Zhiqiang Yuan, Junwei Liu, Qiancheng Zi, Mingwei Liu, Xin Peng, and Yiling Lou. 2023. Evaluating Instruction-Tuned Large Language Models on Code Comprehension and Generation. *arXiv:2308.01240* [cs.CL]
- [374] Zhiqiang Yuan, Yiling Lou, Mingwei Liu, Shiji Ding, Kaixin Wang, Yixuan Chen, and Xin Peng. 2023. No More Manual Tests? Evaluating and Improving ChatGPT for Unit Test Generation. *arXiv preprint arXiv:2305.04207* (2023).
- [375] Daoguang Zan, Bei Chen, Yongshun Gong, Junzhi Cao, Fengji Zhang, Bingchao Wu, Bei Guan, Yilong Yin, and Yongji Wang. 2023. Private-library-oriented code generation with large language models. *arXiv preprint arXiv:2307.15370* (2023).
- [376] Daoguang Zan, Bei Chen, Zeqi Lin, Bei Guan, Yongji Wang, and Jian-Guang Lou. 2022. When language model meets private library. *arXiv preprint arXiv:2210.17236* (2022).
- [377] Daoguang Zan, Bei Chen, Dejian Yang, Zeqi Lin, Minsu Kim, Bei Guan, Yongji Wang, Weizhu Chen, and Jian-Guang Lou. 2022. CERT: Continual Pre-training on Sketches for Library-oriented Code Generation. *arXiv preprint arXiv:2206.06888* (2022).
- [378] Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. 2023. Large Language Models Meet NL2Code: A Survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7443–7464.
- [379] Zhengran Zeng, Hanzhuo Tan, Haotian Zhang, Jing Li, Yuqun Zhang, and Lingming Zhang. 2022. An extensive study on pre-trained models for program understanding and generation. In *Proceedings of the 31st ACM SIGSOFT international symposium on software testing and analysis*. 39–51.
- [380] He Zhang, Muhammad Ali Babar, and Paolo Tell. 2011. Identifying relevant studies in software engineering. *Information and Software Technology* 53, 6 (2011), 625–637.
- [381] Jingxuan Zhang, Siyuan Liu, Lina Gong, Haoxiang Zhang, Zhiqiu Huang, and He Jiang. 2022. BEQAIN: An Effective and Efficient Identifier Normalization Approach With BERT and the Question Answering System. *IEEE Transactions on Software Engineering* (2022).
- [382] Jialu Zhang, Todd Mytkowicz, Mike Kaufman, Ruzica Piskac, and Shuvendu K Lahiri. 2022. Using pre-trained language models to resolve textual and semantic merge conflicts (experience paper). In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 77–88.
- [383] Jiyang Zhang, Pengyu Nie, Junyi Jessy Li, and Milos Gligoric. 2023. Multilingual Code Co-Evolution Using Large Language Models. *arXiv preprint arXiv:2307.14991* (2023).
- [384] Jiyang Zhang, Sheena Panthapackel, Pengyu Nie, Junyi Jessy Li, and Milos Gligoric. 2022. Coditt5: Pretraining for source code and natural language editing. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–12.
- [385] Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, and Xudong Liu. 2020. Retrieval-based neural source code summarization. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*. 1385–1397.
- [386] Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, Kaixuan Wang, and Xudong Liu. 2019. A novel neural source code representation based on abstract syntax tree. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 783–794.
- [387] Kechi Zhang, Ge Li, Jia Li, Zhuo Li, and Zhi Jin. 2023. ToolCoder: Teach Code Generation Models to use APIs with search tools. *arXiv preprint arXiv:2305.04032* (2023).
- [388] Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023. Self-Edit: Fault-Aware Code Editor for Code Generation. *arXiv preprint arXiv:2305.04087* (2023).
- [389] Kexun Zhang, Danqing Wang, Jingtao Xia, William Yang Wang, and Lei Li. 2023. ALGO: Synthesizing Algorithmic Programs with Generated Oracle Verifiers. *arXiv preprint arXiv:2305.14591* (2023).
- [390] Quanjun Zhang, Chunrong Fang, Weisong Sun, Yan Liu, Tieke He, Xiaodong Hao, and Zhenyu Chen. 2023. Boosting Automated Patch Correctness Prediction via Pre-trained Language Model. *arXiv preprint arXiv:2301.12453* (2023).
- [391] Ting Zhang, DongGyun Han, Venkatesh Vinayakarao, Ivana Clairine Irsan, Bowen Xu, Ferdian Thung, David Lo, and Lingxiao Jiang. 2023. Duplicate bug report detection: How far are we? *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 1–32.

- [392] Ting Zhang, Bowen Xu, Ferdian Thung, Stefanus Agus Haryono, David Lo, and Lingxiao Jiang. 2020. Sentiment analysis for software engineering: How far can pre-trained transformer models go?. In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE, 70–80.
- [393] Yuwei Zhang, Ge Li, Zhi Jin, and Ying Xing. 2023. Neural Program Repair with Program Dependence Analysis and Effective Filter Mechanism. *arXiv preprint arXiv:2305.09315* (2023).
- [394] Jianyu Zhao, Yuyang Rong, Yiwen Guo, Yifeng He, and Hao Chen. 2023. Understanding Programs by Exploiting (Fuzzing) Test Cases. *arXiv preprint arXiv:2305.13592* (2023).
- [395] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [396] Xu Zhao, Yuxi Xie, Kenji Kawaguchi, Junxian He, and Qizhe Xie. 2023. Automatic Model Selection with Large Language Models for Reasoning. *arXiv preprint arXiv:2305.14333* (2023).
- [397] Yanjie Zhao, Li Li, Haoyu Wang, Haipeng Cai, Tegawendé F Bissyandé, Jacques Klein, and John Grundy. 2021. On the impact of sample duplication in machine-learning-based android malware detection. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 30, 3 (2021), 1–38.
- [398] Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568* (2023).
- [399] Wenqing Zheng, SP Sharan, Ajay Kumar Jaiswal, Kevin Wang, Yihan Xi, Dejia Xu, and Zhangyang Wang. 2023. Outline, then details: Syntactically guided coarse-to-fine code generation. *arXiv preprint arXiv:2305.00909* (2023).
- [400] Shufan Zhou, Beijun Shen, and Hao Zhong. 2019. Lancer: Your code tell me what you need. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1202–1205.
- [401] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *arXiv preprint arXiv:2308.03279* (2023).
- [402] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models Are Human-Level Prompt Engineers. *arXiv:2211.01910* [cs.LG]
- [403] Jie Zhu, Lingwei Li, Li Yang, Xiaoxiao Ma, and Chun Zuo. 2023. Automating Method Naming with Context-Aware Prompt-Tuning. *arXiv preprint arXiv:2303.05771* (2023).
- [404] Jianfei Zhu, Guanping Xiao, Zheng Zheng, and Yulei Sui. 2022. Enhancing Traceability Link Recovery with Unlabeled Data. In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 446–457.
- [405] Behrouz Zolfaghari, Reza M Parizi, Gautam Srivastava, and Yoseph Hailemariam. 2021. Root causing, detecting, and fixing flaky tests: State of the art and future roadmap. *Software: Practice and Experience* 51, 5 (2021), 851–867.