

Towards More Trustworthy and Interpretable LLMs for Code through Syntax-Grounded Explanations

DAVID N. PALACIO, William & Mary, USA

DANIEL RODRIGUEZ-CARDENAS, William & Mary, USA

ALEJANDRO VELASCO, William & Mary, USA

DIPIN KHATI, William & Mary, USA

KEVIN MORAN, University of Central Florida, USA

DENYS POSHYVANYK, William & Mary, USA

Trustworthiness and interpretability are inextricably linked concepts for LLMs. The more interpretable an LLM is, the more trustworthy it becomes. However, current techniques for interpreting LLMs when applied to code-related tasks largely focus on accuracy measurements, measures of how models react to change, or individual task performance instead of the fine-grained explanations needed at prediction time for greater interpretability, and hence trust. To improve upon this status quo, this paper introduces *ASTrust*, an interpretability method for LLMs of code that generates explanations grounded in the relationship between model confidence and syntactic structures of programming languages. *ASTrust* explains generated code in the context of *syntax categories* based on Abstract Syntax Trees and aids practitioners in understanding model predictions at *both* local (individual code snippets) and global (larger datasets of code) levels. By distributing and assigning model confidence scores to well-known syntactic structures that exist within ASTs, our approach moves beyond prior techniques that perform token-level confidence mapping by offering a view of model confidence that directly aligns with programming language concepts with which developers are familiar.

To put *ASTrust* into practice, we developed an automated visualization that illustrates the aggregated model confidence scores superimposed on sequence, heat-map, and graph-based visuals of syntactic structures from ASTs. We examine both the *practical benefit* that *ASTrust* can provide through a data science study on 12 popular LLMs on a curated set of GitHub repos and the *usefulness* of *ASTrust* through a human study. Our findings illustrate that there is a *causal* connection between learning error and an LLM's ability to predict different syntax categories according to *ASTrust* – illustrating that our approach can be used to interpret model *effectiveness* in the context of its syntactic categories. Finally, users generally found *ASTrust*'s visualizations useful in understanding the trustworthiness of model predictions.

CCS Concepts: • Software and its engineering → Software development techniques.

Additional Key Words and Phrases: Causality, Interpretability, LLMs for Code, Trustworthiness

ACM Reference Format:

David N. Palacio, Daniel Rodriguez-Cardenas, Alejandro Velasco, Dipin Khati, Kevin Moran, and Denys Poshyvanyk. 2024. Towards More Trustworthy and Interpretable LLMs for Code through Syntax-Grounded Explanations. *ACM Trans. Softw. Eng. Methodol.* 1, 1 (July 2024), 26 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Authors' addresses: David N. Palacio, danaderpalacio@wm.edu, William & Mary, Williamsburg, Virginia, USA, 23185; Daniel Rodriguez-Cardenas, William & Mary, Williamsburg, Virginia, USA, 23185, dhrodriguezcar@wm.edu; Alejandro Velasco, William & Mary, Williamsburg, Virginia, USA, 23185, svelascodimate@wm.edu; Dipin Khati, William & Mary, Williamsburg, Virginia, USA, 23188, dkhati@wm.edu; Kevin Moran, University of Central Florida, Orlando, Florida, USA, kpmoran@ucf.edu; Denys Poshyvanyk, William & Mary, Williamsburg, Virginia, USA, 23185, dposhyvanyk@wm.edu.

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 the author(s) 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.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1049-331X/2024/7-ART \$15.00

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

1 INTRODUCTION

The proliferation of open-source software projects and rapid scaling of transformer-based Large Language Models (LLMs) has catalyzed research leading to the increased effectiveness of automated Software Engineering (SE) tools. LLMs have demonstrated considerable proficiency across a diverse array of generative SE tasks [7, 62], including, but not limited to, code completion [10, 50], program repair [3, 9], and test case generation [64]. Current research in both designing LLMs for code and applying them to programming tasks typically makes use of existing *benchmarks* (e.g., CodeSearchNet [22], or HumanEval [8]) and *canonical metrics* (by *canonical*, we refer to metrics that reflect an *aggregate performance* across many model predictions, for example, percentage accuracy). These canonical metrics have been adapted from the field of Natural Language Processing (NLP) to evaluate the performance of deep code generation models.

Recent work has illustrated the limitations of benchmarks such as HumanEval [30] and there has been growing criticism of canonical metrics within the NLP community due to the lack of an *interpretable context* that allows for a deeper understanding of LLMs' predictions or outputs [13, 27, 32, 40, 61]. While code-specific metrics such as CodeBLEU [51] may provide more robust aggregate pictures of model accuracy, they cannot provide the fine-grained context required to truly explain model predictions. The general lack of widely adopted interpretability or explainability tools is a barrier to the adoption of any deep learning model, and in particular LLMs of code, as practitioners are skeptical of models' trustworthiness [34]. This deficiency largely stems from the fact that such benchmarks and canonical metrics are often aimed at evaluating functional correctness or standard performance of generated code *at a glance*. That is, the evaluation is reduced to a single aggregate metric in which relevant information related to individual predictions is obfuscated [6].

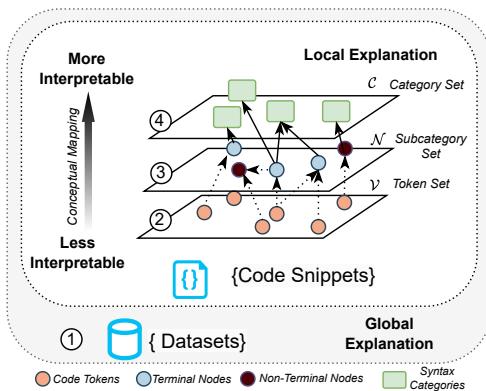


Fig. 1. The Conceptual Framework of Syntax-Grounded Interpretability

addresses challenges related to uncertainty and the potential for bugs or vulnerabilities but also plays a pivotal role in transforming a model perceived as untrustworthy into one deemed as reliable [53].

We assert that a LLM for code is *interpretable*, and hence more trustworthy, if the reasoning behind its predictions is easy for a practitioner to comprehend. In other words, a useful interpretability technique must provide a *conceptual mapping* between descriptions of a model's reasoning process and concepts inherently understood by programmers. In this paper, we explore the possibility of using a model's *confidence* in its predictions as a proxy for describing its reasoning process and

Methods for *interpreting* and *trusting* LLMs for code are inextricably linked. A trustworthy LLM for code requires some degree of interpretability of its predictions, such that model behavior can be understood at a fine-grained enough level to judge which parts of the output are correct or not, and why. The more interpretable an LLM for code is, the higher the confidence and trust in the deployment and use of the model [15, 23]. Notably, interpretability has been identified as an important component for enhancing trustworthiness in various studies [29, 35, 65]. When evaluating trustworthiness, a clear understanding of how and why a model reaches specific predictions is critical. This transparency not only

develop a technique, which we call **ASTrust** that automatically *aligns* and *clusters* model confidence measures with groups of tokens based on syntactic categories derived from Abstract Syntax Trees (ASTs) that we call *Syntax Categories* (*SCs*). This method enables a fine-grained understanding of the correctness of model predictions rooted in **syntax-grounded explanations**. As illustrated by the overview of our approach in Fig. 1, **ASTrust** enables two different granularities of interpretability, *local explanations* at the code snippet level, and *global explanations* for large collections of code. **ASTrust** also makes two main contributions: (i) a statistical technique for aligning and aggregating confidence scores to syntactic code structures of different granularities, and (ii) an automated technique for generating visualizations of these aligned confidence scores. At the *local* level these visualizations take the form of model confidence scores overlaid on both sequence and graph-based illustrations of ASTs and different syntactic structures. At the *global* level, these take the form of a heat map with confidence values clustered around higher-level syntactic categories. An example of the type of explanation that a developer may derive from **ASTrust**'s visualizations is as follows, “*The model’s prediction of the type of the character parameter may be incorrect due to low confidence.*”

Grounding explanations of model confidence in code syntax provides an informative context to practitioners allowing for interpretability. This is due to the fact that code semantics and syntax are tightly coupled. That is, descriptions of code meaning, or semantics, are often *grounded* in syntax. For instance, consider the following example of a developer describing program behavior in numpy in which the description of functionality is grounded in terms of data structures, “*Convert an array representing the coefficients of a Legendre series,*”¹ where the underlined word refers explicitly to the syntactic category of a data structure. One may ask “*why not ground explanations in code semantics directly?*” However, such semantic-based grounding is difficult to achieve, as it requires reasoning among model confidence, input code, predicted code, and widely variable interpretations of code meaning – leading to the potential for *incorrect* explanations that would undermine a technique ultimately meant to build trust. However, as we illustrate in this paper, it is possible to directly map measures of model confidence to different syntactic categories of code, providing a *statistically sound* method of understanding the potential correctness of model predictions rooted in concepts that developers can easily understand.

We explore the *practical benefit* of **ASTrust** through a large-scale data science study examining the relationship between model effectiveness and global explanations and evaluate the *usefulness* of our method through a human study targeted at local explanations of code snippets using **ASTrust**'s different visualizations. The context of our empirical evaluation includes 12 popular LLMs for code and a curated set of code taken from recent commits of the 200 most popular Python projects on GitHub. Using a carefully crafted causal inference study, our analysis illustrates *causal* connections between learning error and a model’s ability to predict different syntax categories according to **ASTrust** – showing that our approach can be used to interpret model *effectiveness* in the context of its syntactic categories. Our human study included 27 participants who examined code snippets completed by GPT 3 and one of four of **ASTrust**'s visualization techniques for local explanations. Our results illustrate that developers generally found **ASTrust** and its visualizations useful in understanding model predictions.

The results of our studies illustrate that mapping token-level predictions of LLMs to segregated Syntax Categories are of considerable practical benefit to SE researchers and practitioners because it allows them to **interpret** and **trust** parts of generated code based on the structural functionality, which contextualizes model predictions beyond the canonical evaluation (*i.e.*, measuring intrinsic and extrinsic metrics). We hope other researchers build upon our method to create new types of

¹<https://github.com/numpy/numpy/blob/main/numpy/polynomial/legendre.py#L152C5-L152C72>

interpretability techniques in the future, and we provide an online appendix with the code for *ASTrust*, and our data and experimental infrastructure to facilitate replication [44].

2 BACKGROUND & RELATED WORK

In this section, we present background on interpretability and trustworthiness as complementary terms for generating syntax-grounded post hoc (e.g., generated after training) explanations for LLMs of code.

Interpretability. The brittleness of LLMs can be formulated as an *incompleteness* in problem formalization [14], which means that it is insufficient that models only infer predictions for certain tasks (the **what?**). The models must also explain how they arrive at such predictions (the **why?**). To mitigate such incompleteness in problem formalization, the field of *interpretability* has risen to encompass techniques and methods that aim to solve the *why* question. Although authors in this field generally use the terms *explainability* and *interpretability* interchangeably, these definitions are inconsistent throughout the literature [16]. We distinguish between the terms to avoid confusion with the purposes of our approach. We will use *explainability* for methods whose goal is to understand how a LLM operates and comes to a decision by exploring inner mechanisms or layers. Conversely, we will use *interpretability* for methods that define *conceptual mapping mechanisms* whose goal is to contextualize models' predictions by associating them with an understandable concept, which in this paper is the syntax of programming languages.

Related Work on Interpretability in NLP. There are existing techniques in both natural language processing (NLP) and SE literature focused on interpretability, including LIME [52], Kernel SHAP [35], Integrated Gradient [58] and Contextual Decomposition [42]. These techniques generally try to approximate an interpretable model that either attempts to attribute meaning to hidden representations of neural networks, or illustrate the relationship between input features and model performance. However, we argue that such techniques are difficult to make practical in the context of LLMs for code, given the lack of conceptual mappings explained earlier. However, the most closely related interpretability technique to *ASTrust*, and one of the only to have adapted to LLMs of code is that of *probing* which is a supervised analysis to determine which type of parameters (e.g., input code snippets, tokenization process, number of hidden layers, and model size) influence the learning process in ML models [60]. Probing aims to assess whether hidden representations of LLMs encode specific linguistic properties such as syntactic structures of programming languages. Given our generated visualizations, there may be an inclination to characterize *ASTrust* as a *probing technique*. However, it is important to note that *ASTrust* is focused on estimating the *correctness* of predicted syntactic code elements rather than mapping meaning to internal model representations of data.

Related Work on Interpretability in SE. In the realm of SE research, prior work has taken two major directions: (i) techniques for task-specific explanations [17, 31, 49], and (ii) empirical interpretability studies using existing NLP techniques [33, 38, 59]. Previous authors have proposed techniques for explaining specific tasks including vulnerability explanation [17], vulnerability prediction for Android [31], and defect prediction models [49]. More recently Liu et al. conducted large empirical study using existing explainability techniques for global explanations of code to better understand generative language models of code [33]. Mohammadkhani et al. conducted a study using LLM's attention mechanism to interpret their performance on generating code. Finally, one paper that proposed a code-specific interpretability technique is that of Cito et al. [11] who formulated a method to generate explanations using counterfactual reasoning of models. Our work on *ASTrust* complements this body of past work by developing a *new, generally applicable interpretability method* that can be applied to *both* local and global explanations of code, which no prior study or technique has done.

Trustworthiness. This research is inspired by definitions of *trust* from automated systems, SE, and NLP. In automated systems, trust is defined as “*the attitude that an agent will help achieve an individual’s goal in a situation characterized by uncertainty and vulnerability*” [28]. Bianco et al. define software trust as the degree of confidence when the software meets certain requirements [12]. In NLP, Sun et al. argue that LLMs must appropriately reflect truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability for them to be trustworthy [57]. We define trust as the confidence that practitioners and researchers have in LLMs’ code prediction, anticipating that these predictions will effectively align with their intended goals. Trustworthiness in LLMs implies a sense of interpretability in a given LLM’s performance, instilling confidence among practitioners in their abilities to perform code-related tasks. To the best of our knowledge, no paper proposes a concrete definition of trust based on interpretability within the SE research community. Yet, several researchers have called for the importance of trustworthiness in LLMs for code [34, 56]. In our work we present a concrete definition of trustworthiness, highlight its importance, and show how syntax-grounded explanations such as *ASTrust* contribute to more trustworthy LLMs.

3 SYNTAX-GROUNDED EXPLANATIONS

At a high level, *ASTrust* queries a LLM for probabilities per token, estimates the median across tokens that are part of one AST node, and presents those averages as **confidence performance** values segregated by hand-assigned syntax categories. We also refer to this confidence performance as *ASTrust Interpretability Performance*.

ASTrust consists of four steps depicted in Fig. 1. In step ①, a code snippet for local or a testbed for global explanations is the starting point of the *interpretability process*. Each sequence within the snippet or the testbed is processed by a tokenizer (e.g., Byte-Pair Encoding (BPE)). In step ②, the tokenizer sets a vocabulary we named **token set**. Once code sequences are preprocessed, an LLM under analysis generates **token-level predictions (TLP)** for each position in a sequence. Next, in step ③, the generated token-level predictions are aligned with the associated Abstract Syntax Tree (AST) terminal nodes. Terminal nodes only store *TLP*, while non-terminal nodes hierarchically store clustered and aggregated *TLP*. Terminal and non-terminal nodes comprise the **subcategory set**. For example, consider `if_` BPE token from the *token set*. This token is aligned with the ‘if’ terminal AST node while clustered in the ‘if_statement’ non-terminal node. Finally, in step ④, ten *syntax categories* are proposed to summarize a model’s predictions. *Syntax Categories* aim to group the sub-categories into higher-level, more human-understandable categories. These syntax categories are a fixed **category set** that comprises more interpretable elements and include:

- Decisions
- Data Structures
- Exceptions
- Iterations
- Functional Programming
- Operators
- Testing
- Scope
- Data Types
- Natural Language

For instance, the sub-categories ‘if_statement’ and ‘if’ are both clustered into one syntax category *Decisions*. In the end, *ASTrust* generates an averaged score per category for global explanations and an AST tree visualization with stored scores at each node for local explanations. In essence, we propose that *syntax elements* contain semantic information that contextualizes predicted probabilities. However, this semantic information varies across the granularity of these elements. We can claim, for example, that token-level elements carry less interpretable information than category-level elements.

ASTrust produces *post-hoc* local and global explanations of generated code snippets. A local explanation intends to interpret the generation of a code snippet by decomposing it into AST elements. Conversely, a global explanation uses a set of generated snippets (or existing benchmark dataset) to interpret a given model holistically into *Syntax Categories (SCs)*. The following sub-sections introduce the building blocks of syntax-grounded explanations. Sec. 3.1 defines the interpretable sets (*e.g.*, *Token*, *Subcategory*, and *Category*) that contain the *syntax elements* employed for the interpretability process. Sec. 3.2 formalizes two function interactions that communicate previously interpretable sets. Such communication consists of aligning and clustering elements from code tokens to syntax categories. Finally, Sec. 3.3 shows the process of generating local and global explanations.

3.1 Interpretable Syntax Sets

Token Set \mathcal{V} . Although *ASTrust* was designed to be compatible with different types of LLMs, this paper concentrated on *Decoder-Only* models due to their auto-regressive capacity to generate code [67] by preserving *long-range dependencies* [26]. A Decoder-only model can be employed as a generative process such as any token w_i is being predicted by $\hat{w}_i \sim P(w_i|w_{<i}) = \sigma(y)_i = e^{y_{w_i}} / \sum_j e^{y_j}$. The term y_j represents the *non-normalized log-probabilities* for each output token j (see Fig. 3). We extracted and normalized these log-probabilities from the last layer of LLMs to estimate *Token-Level Predictions (TLP)*. This estimation relies on the softmax function. The softmax σ_i returns a distribution over predicted output classes, in this case, the classes are each token in the *token set \mathcal{V}* . The predictions σ_i are expected to be influenced by previous sequence inputs $w_{<i}$.

Subcategory Set \mathcal{N} . This set comprises elements of Context-Free Grammars (CFGs). Such elements are rules containing the syntax and structural information of a programming language [21]. CFGs define instructions that specify how different tokens (*i.e.*, Lexemes) are assembled to form valid statements for each language. Formally, a *CFG* is defined as $\mathbb{G} = (\alpha, \lambda, \omega, \beta)$ where α denotes the finite set of non-terminal nodes, λ the finite set of terminal nodes, ω the finite set of production rules and β the start symbol. CFGs use the terminal and non-terminal nodes (or subcategories) to define the production rules ω for any statement (*e.g.*, conditional, assignation, operator). Furthermore, these terminal and non-terminal nodes retain different meanings. Note that these nodes are the elements of the subcategory set $\lambda, \alpha \in \mathcal{N}$.

Category Set C . The steps three and four in Fig. 1 illustrate the binding of α and λ into a category $c \in C$. We pose the term **Syntax Categories (SCs)** as the elements within the Category Set C . We propose ten different SCs based on tree-sitter bindings [5] for Python. SCs are the semantic units to enable the syntax interpretability of LLMs. As such, *ASTrust* allows for Token-Level Predictions (*TLP*) to be explained in a developer-centric way. In summary, each token in a sequence s can be mapped to a category $c \in C$. With *ASTrust*, practitioners can easily associate LLMs' code predictions to specific structural attributes. For instance, ‘identifier’ and ‘string’ nodes correspond to a common *Natural Language* category in Fig. 4. As such, we can group nodes λ and α into semantically meaningful *categories C*.

3.2 Alignment and Clustering Formalism

The previous subsection describes the syntax elements for enabling LLMs interpretability (*i.e.*, *token-set*, α and λ subcategories, and *Syntax Categories (SCs)*). This section elaborates on the interaction among these elements. Two interactions in the form of a function are defined. First, the alignment function δ links code tokens from the *Token Set \mathcal{V}* to terminal nodes λ . Second, the clustering function θ groups the subcategories λ (terminal nodes) and α (non-terminal nodes) by syntax categories (SCs) from the *Category Set*. Fig. 2 showcases both function interactions δ and θ respectively.

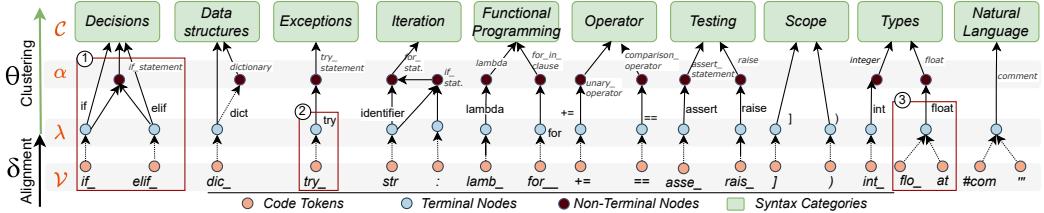


Fig. 2. *Alignment & Clustering Interactions*. The δ function aligns tokens w_i to terminal nodes λ . Terminal and Non-terminal nodes $\lambda, \alpha \in \mathcal{N}$ are clustered by Syntax Categories C .

Alignment Interaction. Fig. 2 illustrates the process of aligning the terminal nodes δ in the AST to their corresponding code tokens w_i . This alignment starts by decomposing an input snippet s into tokens $w_{\leq i} \in \mathcal{V}$. For instance, Fig. 2-② depicts the alignment of `try_` token to the terminal λ ‘try’ node. Note that the alignment ignores the character ‘_’ from `try_`. A tokenizer may produce a sequence in which each token does not necessarily match one-to-one with a terminal λ node, e.g., Fig. 2-③ illustrates the tokens `flo_` and `at` are aligned with the λ node ‘float’. Formally, $\delta(\text{flo_at}) \rightarrow [\text{float}]$ in a many-to-one interaction. Consequently, the alignment between code tokens and terminal nodes is certainly many-to-one, including one-to-one, but never one-to-many or many-to-many.

DEFINITION 1. Alignment (δ). The function $\delta : w_{\leq i} \rightarrow \vec{\lambda}$ where $w_{\leq i}$ corresponds to a code sub-sequence whose tokens are many-to-one associated to the corresponding terminal node vector $\vec{\lambda}$ of syntax subcategories.

Clustering Interaction. A clustering function θ estimates the **confidence performance** of λ and α nodes (subcategories) from an AST by hierarchically aggregating the Token-Level Predictions (*TLP*) to a *Category* $c \in C$. Once the tokens are aligned with their corresponding nodes using δ from Def.1, ASTrust clusters them into their respective category or non-terminal α node according to the AST representation. Some terminal λ nodes can directly be aggregated into a category without considering intermediate non-terminal α nodes. A terminal λ node can initiate a block sentence (i.e., a category) and a block sequence parameters (i.e., non-terminal `if_statement` node). For instance, Fig. 2-① depicts the terminal λ ‘if’ node aggregated into the *Decisions* category and also starts the non-terminal α ‘`if_statement`’ node. To estimate the confidence performance, we traverse the entire AST and aggregate the *TLP* probabilities of respective tokens.

The θ function can adopt average, median, or max aggregations depending on the user configuration. Fig. 3 shows the clustering function applied to a concrete code generation sample. This application constitutes a local post hoc explanation: the parent node ‘parameters’ has a 0.23 associated confidence performance. This parent node average was aggregated with its terminal values: ‘(’ with 0.07, ‘identifier’ with 0.4 and 0.1, ‘,’ with 0.5, and ‘)’ with 0.1. Formally, $\theta(\vec{\lambda} = [0.07, 0.4, 0.1, 0.5, 0.1]) \rightarrow [(parameters, 0.23)]$. If a sample snippet does not contain any particular syntax element (i.e., token, subcategory, or category), such an element is therefore never considered for clustering. An absent syntax element is reported as a null value to avoid biased syntax-grounded explanations.

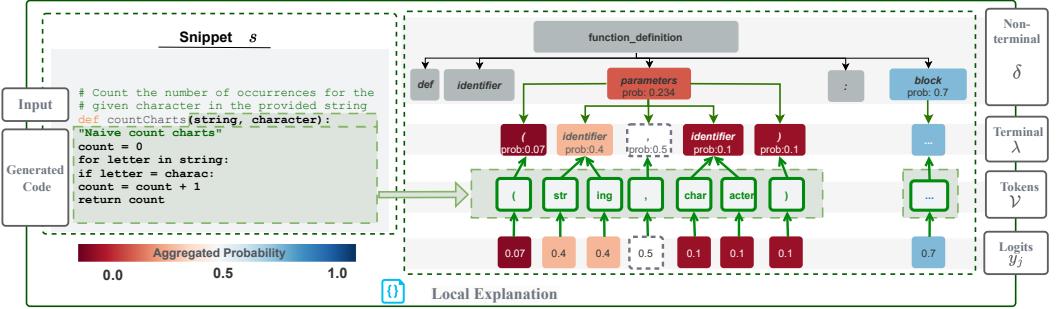


Fig. 3. *Post Hoc Local Explanation*. A snippet is decomposed into code tokens. The highest annotated probabilities (*i.e.*, best predictions) are in blue.

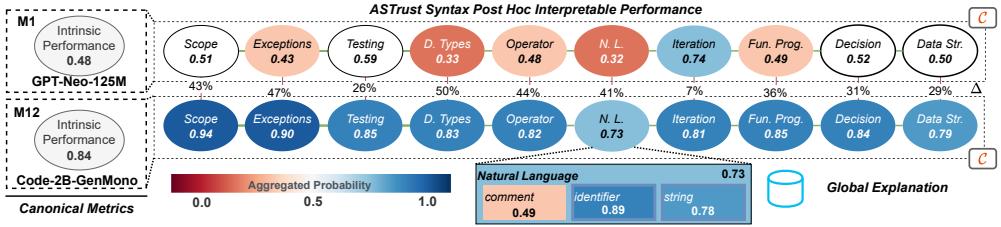


Fig. 4. *Post Hoc Global Explanations Segregated by Categories and Sub-Categories for `gpt-3 [125M]` and `mono-lang [2B]`*

DEFINITION 2. Clustering (θ). The function $\theta : \vec{\lambda} \rightarrow \text{median}(\vec{n})$ where $\vec{\lambda}$ is the resulting vector of a sub-sequence and \vec{n} is the vector of hierarchical associated non-terminal nodes for each terminal λ . The vector \vec{n} , therefore, contains the TLP of non-terminal and the corresponding terminal nodes².

3.3 Post Hoc Local and Global Explanations

LLMs are more understandable when they reflect human knowledge [27]. One way of determining whether an LLM trained on code reflects human knowledge is testing it to see whether or not it operates similar to how a developer would estimate the prediction of a sequence [45]. ASTrust can adopt the form of a post-hoc local or a global explanation to make code predictions humanly understandable.

ASTrust for local interpretability allows us to interpret a single snippet s by generating a visual explanation based on an Abstract Syntax Tree (AST) as illustrated in Fig. 3. A practitioner can explain the code predictions observing the probabilities associated with each element on the AST. In other words, we use θ from Def. 2 to cluster around AST nodes across all levels (*i.e.*, AST probability annotations). Therefore, the syntax-grounded local explanation comprises a conceptual mapping from the code prediction to a terminal and non-terminal node (or sub-categories). Fig. 3 is a visual representation of the conceptual mapping using code predictions by `gpt-3 [1.3B]` model. The visualization displays a confidence value for each λ and δ sub-categories after parsing the AST. The auto-completed snippet is processed with the clustering θ function.

ASTrust for global interpretability allows us to interpret a LLM by decomposing the canonical performance into segregated confidence performance. This segregated confidence is attached to

²In our study, we set the aggregation $\theta : N \rightarrow \text{median}(\hat{w}_{<=i})$ for a subset of tokens $w_{<=i}$.

Syntax Categories (*SCs*). *SCs* are tied to AST tree-sitter nodes [5] and inspired by common object-oriented programming definitions. Although our approach is focused on the Python syntax, these categories may apply to multiple Programming Languages as they are being processed for similar AST elements.

Syntax-grounded global explanations comprise a *conceptual mapping* from code predictions to ten syntax categories. These predictions are calculated for the entire testbed rather than a particular snippet following Def. 2 about clustering function θ . This clustering has an extra step in which a bootstrapping mechanism is used to estimate the *confidence performance mean* of the category set.

Consequently, practitioners can compare Syntax Categories among models and explain the overall behavior of a given LLM by observing the segregated confidence performance. Note that previous analysis can be enriched with the information provided by canonical metrics. For instance, Fig. 4 depicts *M1* with an intrinsic performance of 0.48 while showing a *ASTrust* interpretability performance of 0.74 for the *iteration* category. Details of this global categorization can be further explored and visualized in our online appendix [44].

It should be noted that the validity of both global and local syntax-grounded explanations is dependent upon the *calibration* of LLMs for code in terms of token probabilities and prediction correctness. That is, we assume that token probabilities are a reasonable proxy for the likelihood that a model is correct about a given token prediction. Recent work on calibration for LLMs of code has illustrated that, for code completion (which subsumes the experimental settings in this paper), LLMs tend to be well calibrated to token probabilities [56]. We also further confirm this finding using causal inference in Section 5.3.

4 EMPIRICAL STUDY DESIGN

We study the applicability of *ASTrust* in interpreting code completion tasks. We conducted a *human study* to investigate the *usefulness* of local explanations in real-world settings. In contrast, we conducted a *data science study* to showcase the *effectiveness* of global explanations on a diverse set of LLMs. Finally, we carried out a *causal inference study* to assess the *validity* of the syntax-grounded explanations as they relate to the statistical learning error of the studied models. The following research questions were formulated:

RQ₁ [Usefulness] *How useful are local explanations in real-world settings?* We validate the extent to which AST probability annotations are useful in locally explaining code predictions. We measure usefulness in three key factors: complexity, readability, and LLMs' reliability.

RQ₂ [Effectiveness] *To what extent do LLMs for code correctly predict different syntactic structures?* We interpret the performance of 12 LLMs on each Syntax Category (*SC*). The conceptual mapping allows us to obtain an interpretable and segregated confidence value per category, so we can detect categories that are easier or harder to predict – moving beyond canonical aggregate metrics.

RQ₃ [Validity] *How do Syntax Concepts impact LLMs' statistical learning error?* We validate the causal connection between learning error and LLMs' ability to predict different syntax categories using *ASTrust*.

4.1 Experimental Context

4.1.1 Model Collection. To perform our global analysis, we conducted an interpretability analysis of 12 open Decoder-only LLMs, selected based on their popularity. The largest among these models boasts 2.7 billion parameters. Tab. 1 categorizes these LLMs into four distinct groups, each aligned with a specific fine-tuning strategy. The initial category comprises GPT-3-based models primarily trained on natural language, exemplified by Pile [19]. The second category encompasses models trained on natural language but constructed upon the *codegen* architecture [43]. Moving to the third

Table 1. Large Language Models Descriptions.

Large Language Models (LLMs)				
Type	ID	Name	Architecture	Size
Natural L. gpt-3	M_1	gpt-neo-125m	<i>gpt-3</i>	125M
	M_2^*	gpt-neo-1.3B	<i>gpt-3</i>	1.3B
	M_3	gpt-neo-2.7B	<i>gpt-3</i>	2.7B
Natural L. codegen	M_4	codegen-350M-nl	<i>codegen</i>	350M
	M_5	codegen-2B-nl	<i>codegen</i>	2B
Multi- Language	M_6	codeparrot-small-multi	<i>gpt-2</i>	110M
	M_7	codegen-350M-multi	<i>codegen-350M-nl</i>	350M
	M_8	codegen-2B-multi	<i>codegen-2B-nl</i>	2B
Mono- Language	M_9	codeparrot-small	<i>gpt-2</i>	110M
	M_{10}	codeparrot	<i>gpt-2</i>	1.5B
	M_{11}	codegen-350M-mono	<i>codegen-350M-multi</i>	350M
	M_{12}	codegen-2B-mono	<i>codegen-2B-multi</i>	2B

*The human study was conducted using M_2 - *gpt-3 [1.3B]*

category, we find models trained on multiple programming languages (PLs) using BigQuery [1], implemented on both the gpt-2 and codegen architectures. The final category consists of both *Multi-Language-Type* models fine-tuned on BigPython [43], denoted as *Mono-Language-Type*, and gpt-2 models such as codeparrot [18]. All the datasets for training the LLMs encompass repositories/files sourced from GitHub up to 2021.

4.1.2 Evaluation Dataset. To ensure the integrity of our *ASTrust* evaluation, it is imperative to avoid data contamination by excluding samples used in the training process of the LLMs. We extended and used *SyxTestbed* [54] to overcome this challenge. *SyxTestbed* exclusively comprises code commits from the top 200 most popular Python GitHub repositories between 01/01/22 and 01/01/23. Notably, *SyxTestbed* incorporates comprehensive data, including commit messages, method comments, the entire AST structure, node count, AST levels, AST errors, whitespace details, lines of code, cyclomatic complexity, and token counts.

4.1.3 Machine Configuration. We performed the experiments using 20.04 Ubuntu with an AMD EPYC 7532 32-Core CPU, A100 NVIDIA GPU with 40GB VRAM, and 1TB RAM. For the model inference process, we used HuggingFace and Pytorch [47, 66]. All models were loaded into the GPU to boost the inference time.

4.2 Human Study for *ASTrust* Usefulness

This section presents a preliminary human study comprising a control/treatment experimental design to assess *ASTrust* usefulness in practical settings. We followed a **purposive sampling approach** [4] since our primary goal was to gather preliminary data and insights from practitioners with expertise in ML and SE combined. We selected our subjects carefully rather than randomly to study the usefulness of *ASTrust* (at local explanations). *ASTrust* is designed to enhance the interpretation of model decisions in code completion tasks for practitioners with diverse backgrounds, including researchers, students, and data scientists. By targeting individuals with specific expertise, we ensured that the feedback received was relevant and informed, thereby enhancing the quality of our preliminary findings.

4.2.1 Survey Structure. Each survey consists of three sections. The *first section* is aimed at gathering participant profiling information. The profiling section aims to collect information related to how proficient the participants are when using Python and AI-assisted tools in code generation tasks. In particular, we asked about their level of expertise in Programming Languages (PL) and how familiar

they are with AST structure. Furthermore, we asked about any challenges they encountered when using AI-assisted tools. This information is relevant because we want to control external factors that may influence their perception in validating *AStTrust*. The *second section* aimed to present four code completion scenarios and ask participants to rate their quality. For each scenario, we presented an incomplete Python method as a prompt, the generated code completing the previous method (using *gpt-3 [1.3B]* in Tab. 1), and a specific type of local explanation (e.g., U_{SEQ} , $U_{AST[p]}$, and $U_{AST[c]}$ in Fig. 5). In all four scenarios, the prompt contained either a *semantic error* (e.g., using an undefined variable, incorrect condition statement(s), calling an undefined function) or a *syntax error* (e.g., missing colon, missing parenthesis, incorrect indentation) that participants needed to reason about after considering a given local explanation type. We aimed to capture the participant's perspective regarding the explanation by asking them to describe the cause of a syntax or semantic error from the generated code. To facilitate the analysis, we highlighted the portion of the code generated by the model (see green highlight Fig. 3). We did not provide any detail about the LLM used to generate the predictions in order prevent introducing bias related to preconceived notions about particular LLMs in the responses. Lastly, the *third section* aim to collect feedback about the effectiveness of the different types of local explanation. Specifically, we asked participants about the complexity of the visualizations and potential opportunities for enhancement.

4.2.2 Survey Treatments. To collect the perception of practitioners regarding the usability of *AStTrust*, we devised a control survey (U_{CTR}) and three treatments with two types of local explanations: sequential (U_{SEQ}) and AST-based (U_{AST}) explanations. U_{CTR} represents the absence of a local explanation and only collects the participants' perceptions regarding the correctness of LLMs' output. By contrast, treatment surveys U_{SEQ} and U_{AST} yield syntax-grounded explanations. It is worth noting that we define *correctness* as the degree to which the generated code reflects the purpose of the algorithm contained in the prompt. In other words, we ask participants to judge whether the model predicted a valid or closely accurate set of tokens given the information context within the prompt.

Fig. 5 depicts all the explanation types considered in the survey. U_{SEQ} displays the tokens and their corresponding probabilities in a sequential (i.e., linear layout). Linear representations are commonly used by feature-importance explainability techniques such as attention-based [39] and Shapley [24] values. Therefore, U_{SEQ} serves as a *baseline* to determine how our local explanations U_{AST} provide insightful information beyond the linear layout. Conversely, U_{AST} uses an AST visualization, comprising two types of local explanations: AST-Complete (i.e., $U_{AST[c]}$) and AST-partial (i.e., $U_{AST[p]}$). $U_{AST[c]}$ represents the entire sample's AST (i.e., prompt and generated code) including *AStTrust* confidence performance for all nodes. Conversely, $U_{AST[p]}$ is a filtered $U_{AST[c]}$ representation that only exposes the confidence performance of the generated code and omits the nodes from the prompt.

4.2.3 Survey Metrics. When evaluating the usefulness of our approach to answer RQ₁, we measure the qualitative features of local explanations depicted in Fig. 5. More precisely, we proposed five qualitative metrics to evaluate the usefulness of our approach: *Information Usefulness*, *Local Explanation Complexity*, *Local Explanation Readability*, *Visualization Usefulness*, and *LLM's reliability*. We used a Likert scale with three options for quantitatively measuring the responses. Specifically for Information Usefulness: *Agree*, *Neutral* and *Disagree*. For Local Explanation Complexity, Local Explanation Readability and Visualization Usefulness: *Useful*, *Slightly useful* and *Not useful*. Finally, for LLM's Reliability: *Not reliable*, *Highly reliable* and *Impossible to tell*. Each of the survey metrics corresponds to one of the following survey questions.

Metric₁: Information Usefulness - 'Q: How useful was the information for interpreting the model's decisions?' In the treatment surveys, we ask the participants to explain the LLM's behavior when

completing the code of individual samples, and we gauge their perception regarding the usefulness of the provided information to accomplish this task. We anticipate correlations between the explanation types and perceived usefulness.

Metric₂: Local Explanation Complexity - ‘Q: I found the visualization unnecessarily complex’. The local explanation complexity refers to the degree of intricacy of its types. The degree of complexity may affect perceptions of usefulness.

Metric₃: Local Explanation Readability - ‘Q: I thought the visualization was easy to read and use’. We define readability as the degree to which our local explanations are intuitive and easy to understand. We hypothesize that if the explanation fits this criterion, we can consider it useful. Readability accounts for factors such as the amount of consigned information, the arrangement of tokens and categories, and the color scheme.

Metric₄: Visualization Usefulness - ‘Q: I thought the visualization was useful for explaining the model’s behavior’. The visualization is the graphical representation of the local explanation (refer to Fig. 3). Each treatment survey uses a type of visualization (i.e., U_{SEQ} or U_{AST}). We formulate this question to determine which visualization is considered more useful.

Metric₅: LLM’s Reliability - ‘Q: What is your perception of the model’s reliability in generating code?’. We define reliability as the degree to which a user trusts the LLM’s output based on the outcomes from local explanations. We ask the participants to reflect on the LLM’s reliability across our surveys using local explanations. Considering all four code completion scenarios in the surveys include errors, the greater the number of participants in each survey who would not rely on the model, the more valuable the syntax-grounded local explanation.

4.2.4 Open Questions. In addition to survey metrics, we formulated several open-ended questions for collecting the participants’ perception about the correctness of the predictions ($Open_1$) and the most helpful parts of the visual explanations including potential improvement aspects ($Open_2$). Each of these open metrics corresponds to one or more survey questions.

Open₁: LLM’s Prediction Correctness - ‘Q: If the generated code is incorrect, can you explain why the model might have made the mistake? Otherwise, If the generated code is correct, can you speculate on why the model may have been able to correctly predict the above snippet?’. We asked the participants to use the provided information per sample to analyze whether the prompt or the generated code contained any syntax or semantic error. In U_{CTR} , we aimed to assess the extent to which participants could reason about the source code correctness without any type of explanation provided. Conversely, in U_{SEQ} and U_{AST} , we inspected if the layout information somehow contributed to detecting and reasoning about the cause of the error.

Open₂: Importance of visual explanations - ‘Q₁ : What information from the visualization did you find useful in explaining the model’s predictions?’, ‘Q₂ : What information from the visualization did you find useful in explaining the model’s predictions?’, ‘Q₃ : What other information (if any) would you like to see in the visualization?’, ‘Q₄ : What elements of the visualization did you like most?’, ‘Q₅ : What elements of the visualization did you like least?’. We asked the participants to provide overall feedback about the type of representation used in the treatment surveys (U_{SEQ} and U_{AST}). We aimed to identify the most and least useful elements, as well as gather potential ideas for improvement.

To collect, standardize, and analyze the previous group of open-ended questions, two authors independently gathered and reviewed each survey’s responses. Any differences were resolved through discussion to reach a consensus.

4.2.5 Population Profiling. The target population consists of software engineering practitioners experienced in using AI tools for code generation (e.g., ChatGPT, Copilot). Participants were meant to be knowledgeable in Python and understand how algorithms are structured in programming languages and represented in Abstract Syntax Trees (ASTs). While certain knowledge in Deep

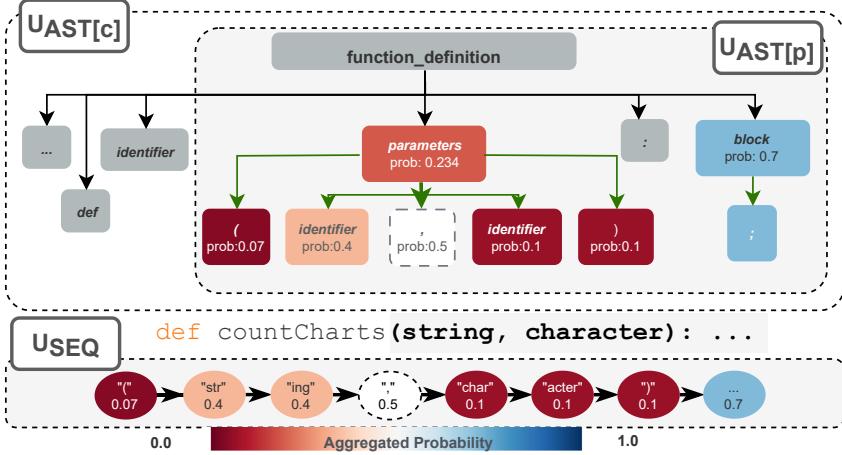


Fig. 5. ASTrust Local Explanation Treatments.

Learning architectures used for Text Generation (e.g., GPT, BERT, T5) is preferred, it was not required. Individuals of any gender were welcome to participate, with a minimum age requirement of 21 years. Participation was entirely voluntary, and no incentives were offered beyond contributing to our efforts to enhance deep learning interpretability for code generation. Furthermore, participants were informed of the voluntary nature of the study during solicitation³.

4.2.6 Data Collection. We reached out to 50 potential participants who were unaware of the purpose of this work, from industrial and academic backgrounds with varying levels of expertise in machine learning and Python. Participants were contacted via email invitations. Out of this group, 27 completed one of the surveys, with the assignment uniformly distributed among the surveys, but we excluded three for low-quality responses, leaving 24 valid submissions. The study was performed on Qualtrics [2] and the anonymized survey data can be found in our appendix [44].

4.2.7 Statistical Analysis. We use U_{SEQ} as a baseline for our study. We expose the participants to ASTrust with two treatments: $U_{AST[p]}$ and $U_{AST[c]}$ (refer to Fig. 5). The result of each question is influenced by these two treatments. To compare the influence of $U_{AST[p]}$ and $U_{AST[c]}$ against U_{SEQ} , we compute the weighted average of the responses from surveys $U_{AST[p]}$ and $U_{AST[c]}$. We refer to the weighted average as U_{AST} . First, we calculate the results of each treatment individually for all the answers. Then, the weight of each answer is estimated by averaging the number of responses per answer across all samples. We then normalize this weight to get the final weighted average for U_{AST} . We use this weighted average for all our statistical analyses in the paper.

4.2.8 Survey Validity. To validate the design of the human study, we conducted a pilot experiment with 10 individuals excluded from the pool of participants. Based on this pilot, the quality and appropriateness of the control and treatment surveys were solidified. Initially, the pilot survey included only the U_{CTR} control and the $U_{AST[c]}$ treatment. However, the pilot revealed the need for an intermediate representation serving as a baseline explanation, which is less complex than an AST visualization, to ensure a fair comparison. Consequently, we introduced U_{SEQ} , inspired by techniques such as SHAP [35], as a baseline treatment with a less complex representation. Additionally, we

³This study was reviewed by the protection of human subjects committee at the College of William & Mary under protocol number PHSC-2023-03-03-16218-dposhyvanyk titled *A Survey Research on Code Concepts for Interpreting Neural Language Models*

introduced $U_{AST[p]}$, a partial representation of $U_{AST[c]}$, as a less complex treatment highlighting only the hierarchical structure of the generated code.

4.3 Data Science Study for ATrust Effectiveness

To answer RQ₂ we implemented a data science study to globally interpret 12 LLMs' performance described in Tab. 1 on the *SyxTestbed* dataset. We performed code completion with different input prompts. The input prompt combines the code completion task, a description, and a partial code snippet. Each prompt has a standard maximum size of 1024 tokens for all considered LLMs.

We first compute the normalized log-probabilities (Sec.3.1) or *TLP* \hat{w}_i for each *SyxTestbed* snippet $s \in S$. These log-probabilities were obtained across the 12 LLMs for every token position. The log-probability distributions maintain a consistent vector size $|V|$ for each token position. Subsequently, these distributions underwent processing to extract the log-probability aligned with the expected token at position i . As a result, each token position corresponds to a stored prediction value \hat{w}_i for constructing the *TLP* sequence $w_{\leq i}$. As discussed earlier, this experimental setting is based on the premise that token probabilities are well-calibrated to model correctness, which has been confirmed in code completion settings by prior work [56]. Additionally, we confirm this finding in answering RQ₃ by observing a causal link between learning error and the probabilities used within ATrust.

We used the alignment function δ to obtain the terminal node λ vector (see Def.1). Next, we traversed the AST for each terminal node λ and clustered them into the corresponding final λ, α node and their correspondent *TLP* by applying the θ function (see Sec.3). The clustering was fixed to generate 32 subcategories and their probability values. We estimated a single **confidence performance** metric (*a.k.a.* ATrust Interpretability Performance) per model by averaging the subcategories probabilities. The confidence performance per model was bootstrapped with the median (size of 500 samplings) to ensure a fair comparison. Lastly, we mapped the subcategories to the SCs obtaining a value per Category C (e.g., Data Structures, Decision, or Scope).

To provide a baseline comparison, we calculated canonical extrinsic metrics *BLUE-4* [46] and *CodeBLEU* [51], and intrinsic performance. Extrinsic metrics evaluate downstream tasks directly (*i.e.*, code completion), while intrinsic metrics assess how well a language model can accurately predict the next word given an incomplete sequence or prompt [25].

Our analysis also includes a corner case experiment that compares the smaller *gpt-3* [125M] to the largest *mono-lang* [2B]. We contrasted the subcategories for each LLM to obtain a segregated global explanation Fig. 9. Since we mapped the subcategories to categories, we can observe the ATrust probability gaps between LLMs at more interpretable levels (see, Fig. 4). The probability values for subcategories and categories are the bootstrapped median.

4.4 Causal Inference Study for ATrust Validity

We validate our ATrust approach using causal inference to answer RQ₃. To accomplish this, we formulated a **Structural Causal Model (SCM)** designed to estimate the impact of SC predictions on the overall learning error of LLMs [48]. We consider that the learning error (*i.e.*, cross-entropy loss) of an LLM is causally impacted by the predicted probabilities of syntax elements. This impact indicates that SCs influence the quality of an LLM. We conducted a causal inference analysis using the *do_{code}* technique [45] to estimate SCs influence. Inherently, a developer mentally rationalizes several things such as the

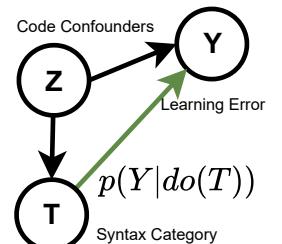


Fig. 6. SCM to estimate Syntax Effect on Learning Error.

concept of the *Iteration* category (see Fig. 9). If an LLM can make a similar mapping, it suggests that it has *statistically learned* some understanding of the syntax cycle structure.

We calculate the causal influence using the *SCM* as the Average Treatment Effect (ATE) with the probability $p(Y|do(T))$ for both *gpt-3 [125M]* and *mono-lang [2B]* models. That is, we estimate the *causal effect* of the variable T on Y after controlling for confounders Z (see Fig. 6). This probability function is estimated using the *doWhy* tool [45, 55]. The proposed treatments (T) embodies Syntax Categories C such as *Decision*, *Natural Language*, or *Iterative*.

The first step of this validity evaluation is to obtain the global intrinsic accuracy. We computed the cross-entropy loss for each snippet s . After obtaining the cross-entropy loss, we estimate Pearson correlation ρ and ATE for 14 sub-categories (λ and α nodes) (Tab. 4). Each sub-category and its cross-entropy loss is correlated with four confounding variables (*i.e.*, Cyclomatic Complexity, AST Levels, #AST Nodes, and Sequence Size) calculating the average value from the set of snippets S from *SyntaxTestbed* dataset [54]. The second step is to validate the obtained ATE by testing the *SCM* robustness (*i.e.*, refutation methods [45]). We limited our exploration to the best and worst models by intrinsic accuracy as we observed similar correlation values across LLMs.

5 RESULTS

In this section, we present our findings for human, data science, and causal studies. The local analysis is focused on answering RQ₁ by using *ASTrust* to interpret concrete snippets. Similarly, we provide insights into our global analysis to answer RQ₂ and RQ₃, which incorporates the interpretation of LLMs' performance segregated by Syntax Categories, a comparison of edge cases, and a causal assessment of *ASTrust* validity.

Before presenting the results, we point out basic stats about AST data processing: The average tree height of the samples in the empirical study was 30, with an average of 104 tokens and 166 nodes. In the human study, the four samples have distinct complexity levels. The smallest sample has 80 tokens, with 47 AST nodes and a tree of height eight. The biggest sample has a token length of 139, with 117 AST nodes and a tree height of 14.

5.1 RQ₁ *ASTrust* Usefulness

Below, we present the results for each survey question as introduced in Sec. 4.2. Quantified responses are detailed in Tab. 2. In addition, we summarize the most relevant feedback received in the open-ended questions. The full human study's results can be accessed in the appendix [44].

Metric₁: Information Usefulness. The data reveals that 67.48% of participants who evaluated U_{AST} explanations, found the presented information useful or slightly useful, with a slight preference for $U_{AST[p]}$ (67.86%) over $U_{AST[c]}$ (62.5%). However, 75% of participants who evaluated U_{SEQ} felt that it was useful, indicating a stronger preference towards it.

Metric₂: Local Explanation Complexity. Participants found U_{AST} explanations slightly more complex (44%) than U_{SEQ} (42%). In particular, $U_{AST[c]}$ was found substantially more complex (67%) than $U_{AST[p]}$. This is not surprising, given that complete ASTs, even for small code snippets can appear complex.

Metric₃: Local Explanation Readability. Both U_{AST} and U_{SEQ} were found to be similarly readable: 35% participants found U_{AST} easy to read and use, compared to 29% for U_{SEQ} . However, between the two AST types $U_{AST[p]}$ (57%) was far preferred in contrast to $U_{AST[c]}$ (17%), again likely due to the complexity of $U_{AST[c]}$.

Metric₄: Visualization Usefulness. U_{SEQ} visualization was found useful by more than half of the participants who evaluated it (57%). Similarly, 49.8% considered the U_{AST} visualizations useful, with an appreciable preference for $U_{AST[c]}$ (50%) over $U_{AST[p]}$ (42%).

Table 2. Survey results for the ASTrust Local Study.

<i>Survey Question</i>		<i>Results (% answers)</i>		
<i>Information Usefulness</i>	U_{SEQ}	50.00	25.00	25.00
	U_{AST}	43.57	23.91	32.52
	$U_{AST[p]}$	39.29	28.57	32.14
	$U_{AST[c]}$	41.66	20.84	37.50
		<i>Agree</i>	<i>Neutral</i>	<i>Disagree</i>
<i>Local Explanation Complexity</i>	U_{SEQ}	42.00	29.00	29.00
	U_{AST}	44.00	28.00	28.00
	$U_{AST[p]}$	14.00	43.00	43.00
	$U_{AST[c]}$	67.00	17.00	16.00
		<i>Agree</i>	<i>Neutral</i>	<i>Disagree</i>
<i>Local Explanation Readability</i>	U_{SEQ}	29.00	29.00	42.00
	U_{AST}	35.00	21.00	44.00
	$U_{AST[p]}$	57.00	29.00	14.00
	$U_{AST[c]}$	17.00	33.00	50.00
		<i>Agree</i>	<i>Neutral</i>	<i>Disagree</i>
<i>Visualization Usefulness</i>	U_{SEQ}	57.00	29.00	14.00
	U_{AST}	49.80	27.78	22.42
	$U_{AST[p]}$	42.00	29.00	29.00
	$U_{AST[c]}$	50.00	33.00	17.00
		<i>Highly Reliable</i>	<i>Not Reliable</i>	<i>Impossible to Tell</i>
<i>LLM's Reliability</i>	U_{SEQ}	29.00	42.00	29.00
	U_{AST}	0.00	62.00	38.00
	$U_{AST[p]}$	0.00	57.00	43.00
	$U_{AST[c]}$	0.00	67.00	33.00
		<i>Highly Reliable</i>	<i>Not Reliable</i>	<i>Impossible to Tell</i>

* **bold**:Highest %, background:Highest % U_{SEQ} or U_{AST}

Metric₅: LLM's Reliability. The high number of participants who judged the LLM as unreliable suggests that all types of explanations helped them assess the quality of the predicted code. However, $U_{AST[c]}$ stood out, with 67% participants favoring it. Meanwhile, 29% participants felt confident about the model based on the information presented by U_{SEQ} , indicating the potential of formulating incorrect interpretations when using this type of explanation. Interestingly, a high percentage of participants (43%) found that $U_{AST[p]}$ cannot help to conclude whether LLM is reliable.

Open₁: LLM's Prediction Correctness. Participants attributed the cause of an incorrect prediction in the model's output to syntax and semantic errors in both treatment and control surveys. The attribution to training bias was prevalent in U_{CTR} as evidenced in answers such as “*Model has not seen enough samples to differentiate between [the characters] = and ==*” or “*Maybe the model is trained in problems with similar error*”. However, in U_{SEQ} and U_{AST} those responses included attribution to the low ASTrust confidence performance, such as “*The probabilities are very low so the predictions are not correct*” and “[the character] = has low probability score of 0.0096”. These results reveal that ASTrust explanations provided insightful information for the participants to judge the model's decisions.

Open₂: Importance of visual explanations. Participants favored the color scheme and the ASTrust confidence performance associated with each token as the most liked elements in the visual explanations. Conversely, they disfavored the inclusion of certain syntax-related tokens, such as white spaces and punctuation marks, in the interpretability analysis. We also encountered contradictory premises: $U_{AST[p]}$ participants believe the explanation missed important details, while in $U_{AST[c]}$ participants criticized the information overload. Participants also suggested improving the navigation in U_{AST} representations by incorporating a mechanism to interactively collapse AST nodes.

Profiling. We found that the participants were well-qualified to take our survey. They all had some background in Machine Learning (Formal or Informal). Similarly, 81.25% of participants were also familiar with the concept of AST.

RQ₁: Although Sequence Explanations contain useful information, AST visualizations were viewed most favorably among explanation types. In fact, AST-based Explanations were found most effective to judge LLM reliability.

5.2 RQ₂ ASt_{rust} Effectiveness

To answer RQ₂ we computed both the canonical intrinsic performance and the ASt_{rust} interpretability performance for 12 LLMs (Tab. 1). Fig. 7 depicts the canonical intrinsic performance for each LLM (*i.e.*, box-plot) and the density canonical intrinsic performance (*i.e.*, density plot) by model type (*e.g.*, NL GPT-3, NL Codegen, Mono-Language-Type, and Multi-Language-Type). The intrinsic performance comprises an aggregated metric that allows us to compare models at a glance. For instance, on average the smallest *mono-lang* [110M] (M_9) has a similar intrinsic performance as the largest GPT-based *gpt-3* [2.7B] (M_3) model with intrinsic performance of 0.61 and 0.62 respectively. After grouping the models by types, we observe that *Mono-Language-Type* models excel in the intrinsic performance with the highest density of 0.9 for performance values between (0.6 – 0.8) and an average intrinsic performance of ≈ 0.7 . Despite the fact canonical intrinsic performance can statistically describe, on average, how the model performs at generating code, these metrics are limited to explaining which categories are being predicted more confidently than others.

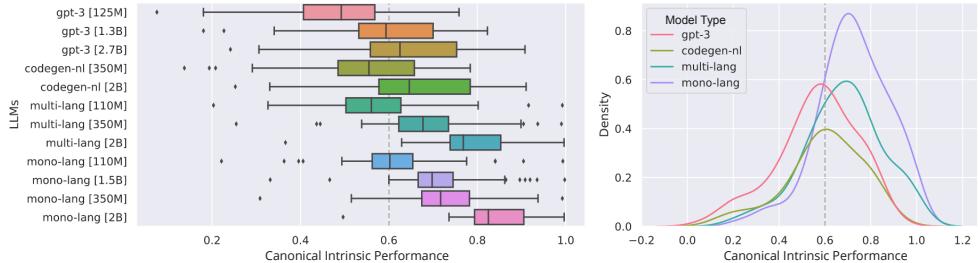


Fig. 7. Canonical intrinsic performance for the models M_1 to M_{12} . Left: box-plots of performance distribution for each model. Right: density plot of performance by model type.

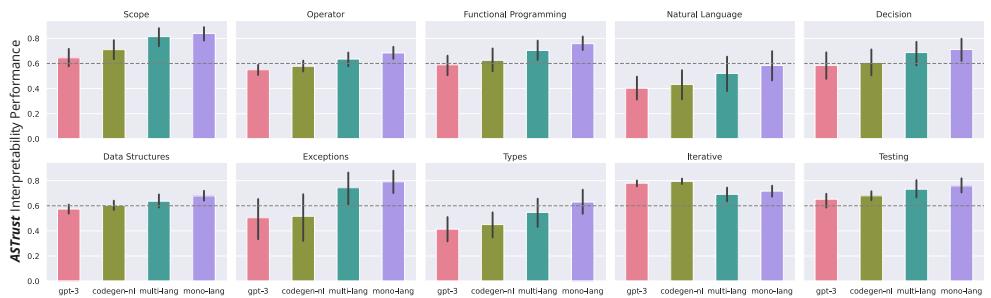


Fig. 8. Segregated ASt_{rust} confidence by Syntax Categories (dotted line is the performance threshold).

To assess the prediction confidence of each Syntax Category (*SC*) for the 12 LLMs we present an empirical ASt_{rust} interpretability performance value (bootstrapped median columns in Tab. 3). Fig. 8 illustrates the ASt_{rust} interpretability performance segregated by Syntax Categories (*SCs*) for each model type. Similarly, Tab. 3 shows bootstrapped median for each model. We set a confidence

prediction threshold of ≥ 0.6 across all analyses. It is worth noting that this threshold is a tunable parameter that can be modified to obtain tailored interpretations of model performance. We easily identify that *Mono-Language-Type* and *Multi-Language-Type* surpass our confidence threshold of 0.6 on all the *SCs* but *Data Types*. Conversely, we observe that *GPT-3-type* models face challenges in *Data Types* categories while excelling in *Iteration* categories.

Table 3. Syntax Concept Empirical Evaluation Results (bold: best, underlined: worst).

LLMs	ASTrust Interpretable Performance (bootstrapped median)										Extrinsic		Intrinsic
	Data Str.	Decision	Except.	F. Prog.	Iter.	NL	Oper.	Scope	Testing	Data Ts.	BLEU-4	CodeBLEU	Perf.
M_1	0.50	0.52	0.43	0.49	0.74	<u>0.32</u>	0.48	0.51	0.59	<u>0.33</u>	0.013	0.139	0.48
M_2	0.60	0.61	0.53	0.62	0.79	<u>0.43</u>	0.57	0.68	0.68	<u>0.44</u>	0.016	0.151	0.59
M_3	0.62	0.63	0.56	0.66	0.81	<u>0.46</u>	0.60	0.74	0.70	<u>0.47</u>	0.015	0.163	0.62
M_4	0.56	0.57	<u>0.45</u>	0.57	0.77	<u>0.39</u>	0.54	0.64	0.64	<u>0.40</u>	0.015	0.151	0.55
M_5	0.65	0.65	0.58	0.68	0.82	0.48	0.61	0.78	0.72	0.50	0.016	0.155	0.65
M_6	0.54	0.55	0.64	0.60	<u>0.60</u>	0.40	0.54	0.71	0.67	<u>0.42</u>	0.010	0.189	0.57
M_7	0.63	0.72	0.75	0.70	0.69	0.51	0.62	0.83	0.73	0.51	0.015	0.171	0.68
M_8	0.74	0.79	0.83	0.81	0.77	0.65	0.74	0.91	0.80	0.71	0.016	0.177	0.79
M_9	0.58	0.58	0.68	0.66	0.63	<u>0.46</u>	0.57	0.73	0.69	<u>0.47</u>	0.011	0.194	0.61
M_{10}	0.67	0.67	0.80	0.76	0.70	0.59	0.66	0.82	0.74	0.64	0.011	0.196	0.71
M_{11}	0.68	0.76	0.78	0.76	0.73	0.57	0.68	0.86	0.77	0.58	0.014	0.179	0.73
M_{12}	0.79	0.84	0.90	0.85	0.81	0.73	0.82	0.94	0.85	0.83	0.016	0.181	0.84

*Erroneous ATrust values are in red. Confident ATrust scores are in blue. Canonical values serve as a baseline.

We found that categories such as *Iteration*, *Except*, and *Scope* surpass our threshold for the majority of our LLMs under analysis. For instance, Tab. 3 shows the *Iteration* category consistently surpasses our threshold for all LLMs, except for *multi-lang [110M]* (M_6), which records an average median ATrust of 0.6 and a highest value of 0.82 for *codegen-nl [2B]* (M_5). Notably, our smaller model *gpt-3 [125M]* (M_1) still outperforms the *Iteration* category prediction with an average median of 0.74. Finally, we note that models trained largely on code, i.e. *codegen-nl [2B]* (M_5), *mono-lang [1.5B]* (M_{10}), *mono-lang [2B]* (M_{12}), could predict the *Data Types* category with an ATrust performance of 0.71, 0.64 and 0.83 respectively.

By contrast, our LLMs struggle to generate good predictions for *Natural language* and *Data Types* categories. We can observe that only *mono-lang [2B]* (M_{12}) surpasses our threshold for *Natural language* with confidence of 0.73, which is still not an outstanding probability. We attribute poor ATrust performance in certain models to the nature of syntax categories like *Natural Language* and *Data Types*, which demand a larger input context for accurate prediction.

Our observations indicate that scaling LLMs' parameters positively influences the prediction of *SCs*. This scaling observation is consistent with canonical scores since our largest models *gpt-3 [2.7B]* (M_3), *codegen-nl [2B]* (M_5), *multi-lang [2B]* (M_8), and *mono-lang [2B]* (M_{12}) report not only intrinsic accuracy that surpasses our threshold but also ATrust confidence over 0.8 for categories such as *Exception*, *Iteration*, and *Scope* (see Tab. 3 in blue). For instance, the largest model, M_{12} exhibits the highest intrinsic accuracy with an avg. median of 0.84 and exceeds our threshold for each category.

By comparing our ATrust against extrinsic metrics, we observe that M_{12} does not achieve the highest *CodeBLUE* score, recording a 0.181. Thus, ATrust offers additional insights into the performance of syntax categories. For example, while *mono-lang [110M]* (M_1) outperforms *mono-lang [2B]* (M_{12}) with a *CodeBLUE* of 0.194, it struggles with inferring *Natural Language* and *Data Types* categories with 0.46 and 0.47 respectively (see Tab. 3).

Corner Case Experiment. Fig. 9, shows a heatmap with the smallest and largest LLMs under analysis, *gpt-3 [125M]* (M_1) and *mono-lang [2B]* (M_{12}) respectively. We selected subcategories and categories with the greatest score jumps for going further into the *SC* analysis. For instance, *Data Types* reported the biggest difference with 0.51 meanwhile the difference between subcategories such as ‘string’ and ‘finally’ are 0.38 and 0.9, respectively. This difference suggests that model scaling

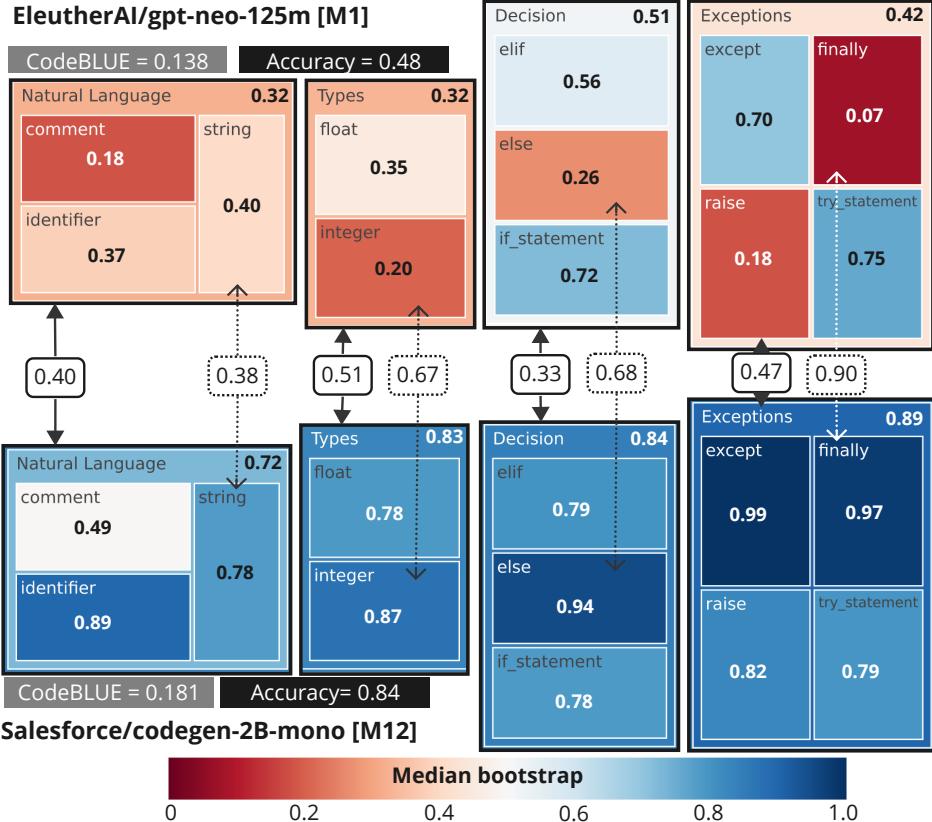


Fig. 9. ASTrust of Syntax Categories and Subcategories (dotted boxes) for corner cases (M_1 and M_{12})

positively impacts the ‘finally’ subcategory, while ‘string’ or ‘if_statement’ subcategories are slightly affected by the model size. We hypothesize that the poor performance of the *Natural Language* is due to limited context windows in the prompt to predict this category. However, a complementary large-scale exploratory analysis of the proportionality types in the training and testing data is required beforehand to determine other causes of poor performance. We also observe that the *Data Types* is prone to errors, especially as these types may frequently appear at the beginning of code snippets, particularly in Python, where dynamic typing is prevalent. This inter-comparison (*a.k.a.* across models) would have been infeasible by just using canonical accuracy metrics.

RQ₂: ASTrust allows us to segregate the prediction performance of LLMs according to Syntax Categories, showing a more interpretable way of comparing models. Syntax-grounded explanations demonstrate, for instance, the struggle of LLMs to statistically learn Natural Language nested within code structures.

5.3 RQ₃ ASTrust Validity

We quantitatively demonstrate that cross-entropy loss of LLMs tends to be negatively impacted by ASTrust probabilities. Therefore, we can explain at syntax category granularity which parts of

the code LLMs perform poorly (see red boxes in Tab. 3). We showcase empirical evidence that the previous statement holds for correlations ρ and causal effects $p(y|do(t))$. Tab. 4 shows, in general, SCs (e.g., Iterative, Scope, or Operator) negatively influence the cross-entropy loss for our best (*i.e.*, M_{12}) and worst (*i.e.*, M_1) models. Negative effects indicate that the better a syntax category is predicted, the lower the learning error associated.

Table 4. ASTrust influence on Learning Error.

[T] Syntax Categories		[Y] Learning Error			
Categories	Sub-Categories	gpt-125		mono-2B	
C	α, λ	ρ	ATE	ρ	ATE
Iterative	<i>for_statement</i>	-0.16	-0.10	-0.07	-0.01
	<i>while_statement</i>	-0.05	-0.11	-0.03	-0.08
Natural Language	<i>identifier</i>	-0.56	-1.78	-0.80	-2.89
Scope	<i>string</i>	-0.31	-0.36	-0.43	-0.55
	<i>return_statement</i>	-0.04	-0.09	-0.22	-0.09
Decision	<i>])</i>	-0.16	-0.04	-0.22	-0.10
	<i>if_statement</i>	-0.37	-0.85	-0.54	-1.49
Operator	<i>comparison_operator</i>	-0.22	-0.21	-0.11	-0.11
Functional Programming	<i>boolean_operator</i>	-0.13	0.02	-0.11	0.00
	<i>for_in_clause</i>	-0.10	0.01	-0.08	-0.09
Intrinsic	<i>if_clause</i>	-0.03	0.09	-0.03	0.04
Programming	<i>lambda</i>	-0.01	0.19	0.01	0.13
	<i>list_comprehension</i>	-0.04	0.20	-0.05	0.06
Baseline		-0.04	0.06	-0.03	0.04
Intrinsic	Avg. Accuracy	-0.38	-1.60	-0.38	-1.78

* **bold**:Highest impact, shadowed: causal effect

The most outstanding finding is that the *Natural Language* category has the largest impact on the cross-entropy loss. For example, the SC ‘*identifier*’ has a causal effect of -1.78 for M_1 and -2.89 for M_{12} . In contrast, *Functional Programming* categories present the lowest impact on cross-entropy loss with a subtle ‘*lambda*’ positive causal effect of 0.2 for M_1 . This subtle positive effect was expected as NL-based LLMs have not been fine-tuned on code corpora with ‘*lambda*’ expressions.

RQ₃: The cross-entropy loss of LLMs tends to be negatively impacted by ASTrust probabilities. This demonstrates that syntax-grounded explanations are indeed representing the syntax learning mechanisms of LLMs at segregated granularity.

6 DISCUSSION

Below, we pose three aspects for discussion: 1) some general insights (*GIs*) from the empirical study, 2) a logical analysis of the connection between trustworthiness and interpretability, and 3) the threats to validity of our approach.

6.1 Empirical Study Insights

GI₁: Token Predictions Reliability. ASTrust relies on logit extraction to generate post-hoc explanations as syntax categories. If logits are wrongly predicted (by over/underfitting), our *causal validation process* detects such inconsistency by reducing the Average Treatment Effect (ATE) of syntax categories on the statistical learning error. Our Structural Causal Model (SCM) was designed

to test the robustness and fidelity of our approach under models' misconfigurations or unreliable performance. Also, as stated earlier in the paper, recent work on calibration for LLMs of code has illustrated that, for code completion (which subsumes the experimental settings in this paper), LLMs tend to be well calibrated to token probabilities/logits [56]. This helps to mitigate issues that may arise due to model confidence and correctness being misaligned.

GI₂: Syntax Aggregations Improves Explanations. Due to its granular nature, token-level predictions are less informative than a hierarchical aggregated level. BPE can make the interpretation of individual tokens much more difficult when code-based sequences are split into tokens that may be meaningless. We posit that practitioners can more easily understand syntax categories rather than individual tokens because these categories are *already defined by context-free grammars*, which are semantically rich. Moreover, our human study provides evidence of this claim since AST-based explanations were found to be easy to read and use by participants. AST-based explanations *also capture semantics* by allowing visualization of the full AST structure. This approach helps practitioners evaluate the model's implementation more effectively by providing a clearer, structured view of the code's semantics.

GI₃: Natural Language Imbalance. Our approach indicates a poor performance on NL sub-categories. We hypothesize this low performance is due to an unbalanced distribution of NL training samples compared to other categories. Before increasing the context window, we believe that a better analysis would be measuring the proportionality of NL sub-categories on the training set and, then, fine-tuning current LLMs to fix possible data bias. Unfortunately, this analysis is currently out of scope since it demands a complementary Exploratory Data Analysis that we envision for future research stages.

GI₄: Foundational Interpretability Research. ASTRust is meant to serve as a more foundational approach required to guide the future development of interpretability tools for users of different backgrounds (e.g., researchers, students, and data scientists). We aimed to not only propose a formal methodology to conduct interpretability in our field but also perform a preliminary assessment of ASTRust's usefulness by conducting a control/treatment experiment (*i.e.*, with and without the approach) on a visualization technique based on ASTRust under clearly defined qualitative metrics.

GI₅: Contradictions about the Usefulness of Explanations. In our human study, we found that AST-based explanations were preferred over sequential-based ones. Results revealed that the AST-partial representation was considered more useful than AST-Complete, as it presents the AST representation and ASTRust confidence performance only for the generated portion of the code. However, the feedback received in the open-ended questions revealed contradictory opinions. Some participants indicated that the AST-partial representation missed important details, while others felt that the AST-Complete representation was excessively detailed. These findings suggest the need for more tailored representations for explanations, aiming to present useful information while maintaining readability. We envision incorporating ASTRust into a tool that adds interactivity to navigate the explanations.

6.2 Trustworthiness & Interpretability Connection

We outline two premises based on state-of-the-art definitions of trustworthiness and direct observations from our quantitative and qualitative analyses. Then, we use *logical deduction* supported by documented and empirical evidence to link the concept of *trustworthiness* with ASTRust highlighting the significance of syntax-grounded explanations.

Premise₁: Interpretability is a cornerstone for trustworthiness in Language Models for Code (LLMs). The interpretability field enhances transparency and provides insights into the decision-making process serving as a key factor in fostering practitioner trust and adoption. In the realm of Deep Learning for Software Engineering (DL4SE) [63], the significance of interpretability

in models to engender trust cannot be overstated. Jiaming et al. [23] underscore interpretability as a pivotal element in aligned models, integral to the RICE principle alongside Robustness, Controllability, and Ethicality. By enhancing transparency in the decision-making process, interpretability plays a crucial role in building trust, a sentiment echoed by Weller et al., who stress the need to extend transparency beyond algorithms to foster trust [65]. The dilemma between accuracy and interpretability, claimed by Lundberg et al. [35], is magnified by the challenge posed by large, complex models that even experts find difficult to interpret. A user study with UX and design practitioners supports this notion, revealing that explanations are sought to gain deeper insights into AI tool decision-making, providing a remedy for the “black box” perception and contributing to user trust and adoption [29]. Therefore, the indisputable importance of interpretability in LLM decision-making lies in its pivotal role in establishing trustworthiness.

Premise₂: ASTrust improves interpretability. It is feasible to segregate intrinsic metrics (*i.e.*, standard accuracy) into interpretable Syntax Categories revealing the LLMs’ inner workings concerning code structure and contributing towards interpretability. By conducting extensive qualitative and quantitative studies involving 12 prominent LLMs, we have demonstrated the effectiveness of *ASTrust* in enhancing interpretability. We do not claim that our set of categories is complete; however, we consider that a good *alignment* of the generated categories by the LLM with the ones expected by humans configures a good explanation [20]. Our *ASTrust* clusters tokens to *meaningful* categories that are easier for human concept association. Furthermore, we uncovered valuable insights, such as the causal influence of AST categories on the cross-entropy loss of LLMs after accounting for confounding factors. Our human study participants attested to the usefulness of our *ASTrust* in explaining the predictions of Python code snippets by a LLM 5.1. By breaking down intrinsic metrics into segregated and interpretable terminal and non-terminal nodes, our approach not only enhances the understandability of LLMs but also unveils crucial insights into the inner workings of syntax elements.

Conclusion. Given the first premise that interpretability is fundamental for trustworthiness in LLMs, supported by several shreds of evidence, and from the second premise asserting that *ASTrust* enhances interpretability by segregating intrinsic metrics into interpretable syntax categories and subcategories, collectively supports the fact that *ASTrust* contributes to the improvement of trustworthiness in LLMs for Code using syntax-grounded explanations.

6.3 Threats to Validity

Threats to **construct validity** concern the intentionality of *ASTrust* in providing useful explanations. Instead of attempting to disentangle information represented between the layers learned by LLMs (*i.e.*, probing [36]), *ASTrust* focuses on conceptually mapping LLMs’ code predictions to present the accuracy in a segregated way. We quantitatively and qualitatively validated the extent to which *ASTrust* is interpretable through causal analyses and a human study. While we cannot claim that the results from our study generalize beyond the population of users that participated in our study, our participants represent a diverse range of backgrounds mitigating this threat. Nonetheless, the purpose of our study was to conduct a preliminary assessment of *ASTrust* representations. As such, all code completion scenarios were designed to include a syntax or semantic error since we assessed how useful our approach is in assisting users in understanding models’ incorrect behavior, increasing the reliability of our findings.

Threats to **internal validity** refer to the degree of confidence in which the *ASTrust* study results are reliable. Firstly, in our causal study, the potential for unidentified confounders in the code may bias the causal relationship between cross-entropy loss and the Syntax Categories. That is why we ensured the robustness of the Structural Causal Model by performing placebo refutations, which involves simulating unrelated treatments and then re-estimating the causal effects. Secondly, we

used rigorous statistical techniques such as bootstrapping to guarantee a consistent comparison between aggregated Token-Level Predictions by syntax elements.

Threats to **external validity** represent the extent to which *ASTrust* can be used to contextualize the performance of other LLMs or datasets. We excluded GPT-4 based models from our empirical experiments due to the constraints of the current OpenAI API, which restricts access to softmax layers values (a key factor in the intrinsic evaluations). While our evaluation relied on decoder-only based models, *ASTrust* can also be used to interpret encoders and other types of auto-regressive architectures. Finally, our *SyntaxTestbed* dataset may not contain enough samples to represent all the syntax categories or Python project attributes fairly. Nonetheless, we designed a data mining pipeline to guarantee the diversity of collected samples.

7 LESSONS LEARNED & CONCLUSIONS

Lesson₁: **Aggregated metrics may give false impressions about LLMs' capabilities.** The research community should incentivize researchers to report AI4SE results in a granular way, as opposed to more traditional aggregated accuracy metrics. After controlling for code confounders, we demonstrated that segregated syntax elements influence the cross-entropy loss of LLMs. This influence persists across models at different parameter sizes and fine-tuning strategies. Syntax information is also relevant for any posterior static analysis of code enabling further evaluations of LLMs in downstream tasks that entail elements of software design (e.g., refactoring).

Lesson₂: **New interpretability methods are required to enable trustworthiness.** In our studies, we have noted an absence of a concrete definition for the term *trust* in the Software Engineering research. However, several researchers have highlighted the importance of establishing trust in work on AI4SE. Research has also shown that interpretability is one of the keys to improving trustworthiness, but at the same time, there is a scarcity of interpretable methods linked to trustworthiness. Despite this limitation, surveyed participants agreed that *ASTrust* was useful to understand why and how a LLM produced certain errors in code-completion tasks.

Lesson₃: **Grounding model explanations in the relationship between syntactic structures and prediction confidence is useful.** It is feasible to segregate intrinsic metrics (*i.e.*, standard accuracy) into interpretable Syntax Categories revealing the LLMs' inner workings concerning code structure and contributing towards interpretability. By conducting extensive qualitative and quantitative studies involving 12 prominent LLMs, we have demonstrated the effectiveness of *ASTrust* in enhancing interpretability. We do not claim that our set of categories is complete; however, we consider that a good *alignment* of the generated categories by the LLM with the ones expected by humans configures a good explanation [20]. Our *ASTrust* clusters tokens to *meaningful* categories that are easier for human concept association. Furthermore, we uncovered valuable insights, such as the causal influence of AST categories on the cross-entropy loss of LLMs after accounting for confounding factors. Our human study participants attested to the usefulness of our *ASTrust* in explaining the predictions of Python code snippets by a LLM 5.1. By breaking down intrinsic metrics into segregated and interpretable terminal and non-terminal nodes, our approach not only enhances the understandability of LLMs but also unveils crucial insights into the inner workings of syntax elements.

Lesson₄: **The usability of proposed techniques must be further evaluated for industry adoption.** We adapted the non-mathematical definition of *interpretability* by Doshi-Velez & Kim [13], Molnar [41] and Miller [37] to the field of AI4SE [63]. However, as our preliminary human study suggests, *ASTrust* solution is incomplete until being extensively evaluated for industry settings.

Artifact Availability: Experimental data, curated datasets, source code, and complementary statistical analysis used in this research are published in an open-source repository [44].

REFERENCES

- [1] 2024. Bigquery Dataset. <https://cloud.google.com/bigquery> [Accessed 23-03-2024].
- [2] 2024. Qualtrics XM: The Leading Experience Management Software. <https://www.qualtrics.com/> [Accessed 23-03-2024].
- [3] Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. [n. d.]. Unified Pre-training for Program Understanding and Generation. arXiv:2103.06333 [cs] <http://arxiv.org/abs/2103.06333>
- [4] Sebastian Baltes and Paul Ralph. 2021. Sampling in Software Engineering Research: A Critical Review and Guidelines. <https://doi.org/10.48550/arXiv.2002.07764> arXiv:2002.07764 [cs].
- [5] Max Brunsfeld, Andrew Hlynksyi, Patrick Thomson, Josh Vera, Phil Turnbull, et al. 2023. tree-sitter/tree-sitter: v0.20.8. <https://doi.org/10.5281/zenodo.7798573>
- [6] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, et al. [n. d.]. Rethink reporting of evaluation results in AI. 380, 6641 ([n. d.]), 136–138. <https://doi.org/10.1126/science.adf6369>
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. 2021. Evaluating Large Language Models Trained on Code. (2021). <http://arxiv.org/abs/2107.03374>
- [8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, et al. [n. d.]. Evaluating Large Language Models Trained on Code. ([n. d.]). arXiv:2107.03374 <http://arxiv.org/abs/2107.03374>
- [9] Zimin Chen, Steve James Kommrusch, Michele Tufano, Louis-Noël Pouchet, Denys Poshyvanyk, et al. 2019. SEQUENCER: Sequence-to-Sequence Learning for End-to-End Program Repair. *IEEE Transactions on Software Engineering* (2019), 1–1. <https://doi.org/10.1109/TSE.2019.2940179>
- [10] Matteo Ciniselli, Nathan Cooper, Luca Pascarella, Denys Poshyvanyk, Massimiliano Di Penta, et al. 2021. An Empirical Study on the Usage of BERT Models for Code Completion. *CoRR* abs/2103.07115 (2021). arXiv:2103.07115 <https://arxiv.org/abs/2103.07115>
- [11] Jürgen Cito, Isil Dillig, Vijayaraghavan Murali, and Satish Chandra. 2022. Counterfactual explanations for models of code. In *Proceedings of the 44th international conference on software engineering: software engineering in practice*. 125–134.
- [12] Vieri del Bianco, Luigi Lavazza, Sandro Morasca, and Davide Taibi. 2011. A Survey on Open Source Software Trustworthiness. *IEEE Software* 28, 5 (2011), 67–75. <https://doi.org/10.1109/MS.2011.93>
- [13] Finale Doshi-Velez and Been Kim. [n. d.]. Towards A Rigorous Science of Interpretable Machine Learning. ([n. d.]), 1–13. Issue Ml. arXiv:1702.08608 <http://arxiv.org/abs/1702.08608>
- [14] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. Ml (2017), 1–13. <http://arxiv.org/abs/1702.08608>
- [15] Finale Doshi-Velez and Been Kim. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. (2018), 3–17. https://doi.org/10.1007/978-3-319-98131-4_1
- [16] Montgomery Flora, Corey Potvin, Amy McGovern, and Shawn Handler. [n. d.]. Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement. arXiv:2211.08943 [physics, stat] <http://arxiv.org/abs/2211.08943>
- [17] Michael Fu, Van Nguyen, Chakkrit Kla Tantithamthavorn, Trung Le, and Dinh Phung. 2023. VulExplainer: A Transformer-Based Hierarchical Distillation for Explaining Vulnerability Types. *IEEE Transactions on Software Engineering* 49, 10 (2023), 4550–4565. <https://doi.org/10.1109/TSE.2023.3305244>
- [18] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, et al. 2021. Codeparrot. (2021). <https://github.com/huggingface/blog/blob/main/codeparrot.md>
- [19] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. arXiv:cs.CL/2101.00027
- [20] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. <http://arxiv.org/abs/1902.03129> arXiv:1902.03129 [cs, stat].
- [21] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Addison-Wesley Longman Publishing Co., Inc., USA.
- [22] 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).
- [23] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, et al. 2024. AI Alignment: A Comprehensive Survey. arXiv:cs.AI/2310.19852
- [24] Ehud Kalai and Dov Samet. 1983. On weighted Shapley values. *International Journal of Game Theory* 16 (1983), 205–222. <https://api.semanticscholar.org/CorpusID:9418424>
- [25] Rafael Michael Karampatsis, Hlib Babii, Romain Robbes, Charles Sutton, and Andrea Janes. 2020. Big code != big vocabulary: Open-vocabulary models for source code. *Proceedings - International Conference on Software Engineering* (2020), 1073–1085. <https://doi.org/10.1145/3377811.3380342>

- [26] Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and Understanding Recurrent Networks. *CoRR* abs/1506.02078 (2015). arXiv:1506.02078 <http://arxiv.org/abs/1506.02078>
- [27] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, et al. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018* 6 (2018), 4186–4195.
- [28] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (Jan 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- [29] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM. <https://doi.org/10.1145/3313831.3376590>
- [30] 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:cs.SE/2305.01210
- [31] Yue Liu, Chakkrit Tantithamthavorn, Li Li, and Yepang Liu. 2022. Explainable ai for android malware detection: Towards understanding why the models perform so well?. In *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 169–180.
- [32] Yue Liu, Chakkrit Tantithamthavorn, Yonghui Liu, and Li Li. 2023. On the Reliability and Explainability of Automated Code Generation Approaches. <http://arxiv.org/abs/2302.09587> arXiv:2302.09587 [cs].
- [33] Yue Liu, Chakkrit Tantithamthavorn, Yonghui Liu, and Li Li. 2024. On the reliability and explainability of language models for program generation. *ACM Transactions on Software Engineering and Methodology* (2024).
- [34] David Lo. [n. d.]. Trustworthy and Synergistic Artificial Intelligence for Software Engineering: Vision and Roadmaps. arXiv:2309.04142 [cs] <http://arxiv.org/abs/2309.04142>
- [35] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:cs.AI/1705.07874
- [36] José Antonio Hernández López, Martin Weyssow, Jesús Sánchez Cuadrado, and Houari Sahraoui. 2022. AST-Probe: Recovering abstract syntax trees from hidden representations of pre-trained language models. <http://arxiv.org/abs/2206.11719> arXiv:2206.11719 [cs].
- [37] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. arXiv:cs.AI/1706.07269
- [38] A. Mohammadkhani, C. Tantithamthavorn, and H. Hemmatif. 2023. Explaining Transformer-based Code Models: What Do They Learn? When They Do Not Work?. In *2023 IEEE 23rd International Working Conference on Source Code Analysis and Manipulation (SCAM)*. IEEE Computer Society, Los Alamitos, CA, USA, 96–106. <https://doi.org/10.1109/SCAM59687.2023.00020>
- [39] A. Mohammadkhani, C. Tantithamthavorn, and H. Hemmatif. 2023. Explaining Transformer-based Code Models: What Do They Learn? When They Do Not Work?. In *2023 IEEE 23rd International Working Conference on Source Code Analysis and Manipulation (SCAM)*. IEEE Computer Society, Los Alamitos, CA, USA, 96–106. <https://doi.org/10.1109/SCAM59687.2023.00020>
- [40] Christoph Molnar. 2019. *Interpretable Machine Learning*. <https://christophm.github.io/interpretable-ml-book/>.
- [41] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. 01 (2020), 1–15. <http://arxiv.org/abs/2010.09337> arXiv: 2010.09337.
- [42] W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453* (2018).
- [43] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, et al. 2023. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. arXiv:cs.LG/2203.13474
- [44] David N. Palacio, Daniel Rodriguez-Cardenas, and Alejandro Velasco. 2024. ASTrust: Github Repository. <https://github.com/WM-SEMERU/CodeSyntaxConcept> [Accessed 23-03-2024].
- [45] David N. Palacio, Alejandro Velasco, Nathan Cooper, Alvaro Rodriguez, Kevin Moran, et al. 2023. Toward a Theory of Causation for Interpreting Neural Code Models. arXiv:cs.SE/2302.03788
- [46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, USA, 311–318. <https://doi.org/10.3115/1073083.1073135>
- [47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems* 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, et al. (Eds.). Curran Associates, Inc., 8024–8035. <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [48] Judea Pearl, Madelyn Glymour, and Nicholas P.Jewell. 2016. *Causal Inference in Statistics, A Primer*.
- [49] Chanathip Pornprasit, Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, Michael Fu, and Patanamon Thongtanunam. 2021. PyExplainer: Explaining the Predictions of Just-In-Time Defect Models. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 407–418. <https://doi.org/10.1109/ASE51524.2021.9678763>

- [50] Veselin Raychev, Martin T. Vechev, and Eran Yahav. 2014. Code completion with statistical language models. *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation* (2014).
- [51] Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, et al. 2020. CodeBLEU: a Method for Automatic Evaluation of Code Synthesis. *CoRR* abs/2009.10297 (2020). arXiv:2009.10297 <https://arxiv.org/abs/2009.10297>
- [52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:cs.LG/1602.04938
- [54] Daniel Rodriguez-Cardenas, David N. Palacio, Dipin Khati, Henry Burke, and Denys Poshyvanyk. 2023. Benchmarking Causal Study to Interpret Large Language Models for Source Code. In *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 329–334. <https://doi.org/10.1109/ICSME58846.2023.00040>
- [55] Amit Sharma, Vasilis Syrgkanis, Cheng Zhang, and Emre Kiciman. 2021. DoWhy : Addressing Challenges in Expressing and Validating Causal Assumptions. (2021).
- [56] Claudio Spiess, David Gros, Kunal Suresh Pai, Michael Pradel, Md Rafiqul Islam Rabin, et al. 2024. Quality and Trust in LLM-generated Code. arXiv:cs.SE/2402.02047
- [57] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, et al. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:cs.CL/2401.05561
- [58] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
- [59] Chakkrit Tantithamthavorn, Jürgen Cito, Hadi Hemmati, and Satish Chandra. 2023. Explainable AI for SE: Challenges and Future Directions. *IEEE Software* 40, 3 (2023), 29–33. <https://doi.org/10.1109/MS.2023.3246686>
- [60] Sergey Troshin and Nadezhda Chirkova. 2022. Probing Pretrained Models of Source Code. <http://arxiv.org/abs/2202.08975> [cs].
- [61] Yao Wan, Wei Zhao, Hongyu Zhang, Yulei Sui, Guandong Xu, et al. 2022. What Do They Capture? – A Structural Analysis of Pre-Trained Language Models for Source Code. <http://arxiv.org/abs/2202.06840> arXiv:2202.06840 [cs].
- [62] Cody Watson, Nathan Cooper, David Nader-Palacio, Kevin Moran, and Denys Poshyvanyk. 2020. A Systematic Literature Review on the Use of Deep Learning in Software Engineering Research. *CoRR* abs/2009.06520 (2020). arXiv:2009.06520 <https://arxiv.org/abs/2009.06520>
- [63] Cody Watson, Nathan Cooper, David Nader Palacio, Kevin Moran, and Denys Poshyvanyk. 2021. A Systematic Literature Review on the Use of Deep Learning in Software Engineering Research. arXiv:cs.SE/2009.06520
- [64] Cody Watson, Michele Tufano, Kevin Moran, Gabriele Bavota, and Denys Poshyvanyk. 2020. On Learning Meaningful Assert Statements for Unit Test Cases. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20)*. Association for Computing Machinery, New York, NY, USA, 1398–1409. <https://doi.org/10.1145/3377811.3380429>
- [65] Adrian Weller. 2019. Transparency: Motivations and Challenges. arXiv:cs.CY/1708.01870
- [66] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [67] Frank F. Xu, Uri Alon, Graham Neubig, and Vincent J. Hellendoorn. 2022. A Systematic Evaluation of Large Language Models of Code. <http://arxiv.org/abs/2202.13169> arXiv:2202.13169 [cs].