



Venturing ChatGPT's lens to explore human values in software artifacts: a case study of mobile APIs

Davoud Mougouei, Saima Rafi, Mahdi Fahmideh, Elahe Mougouei, Javed Ali Khan, Hoa Khanh Dam, Arif Nurwidyantoro & Michel Chaudron

To cite this article: Davoud Mougouei, Saima Rafi, Mahdi Fahmideh, Elahe Mougouei, Javed Ali Khan, Hoa Khanh Dam, Arif Nurwidyantoro & Michel Chaudron (03 May 2025): Venturing ChatGPT's lens to explore human values in software artifacts: a case study of mobile APIs, Behaviour & Information Technology, DOI: [10.1080/0144929X.2025.2478278](https://doi.org/10.1080/0144929X.2025.2478278)

To link to this article: <https://doi.org/10.1080/0144929X.2025.2478278>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 03 May 2025.



Submit your article to this journal [↗](#)



Article views: 242



View related articles [↗](#)



View Crossmark data [↗](#)

Venturing ChatGPT's lens to explore human values in software artifacts: a case study of mobile APIs

Davoud Mougouei^a, Saima Rafi^b, Mahdi Fahmideh^c, Elahe Mougouei^d, Javed Ali Khan^e, Hoa Khanh Dam^f, Arif Nurwidyanoro^g and Michel Chaudron^h

^aSchool of Information Technology, Deakin University, Burwood, Australia; ^bSchool of Computing Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, United Kingdom; ^cSchool of Business, University of Southern Queensland, Springfield, Australia; ^dFaculty of Computer Engineering, Islamic Azad University, Esfahan, Iran; ^eSchool of Physics, Engineering & Computer Science, University of Hertfordshire, Hertfordshire, United Kingdom; ^fSchool of Computing and IT, University of Wollongong, Wollongong, Australia; ^gDepartment of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta; ^hDepartment of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands

ABSTRACT

Software is designed for humans and must account for their values. However, current research and practice focus on a narrow range of well-explored values, e.g. security, overlooking a more comprehensive perspective. Those exploring a broader array of values rely on manual identification, which is labour-intensive and prone to human bias. Moreover, existing methods offer limited reliability as they fail to explain their findings. In this paper, we propose leveraging the reasoning capabilities of Large Language Models (LLMs) for automated inference about values. This allows for not only detecting values but also explaining how they are expressed in the software. We aim to examine the effectiveness of LLMs, specifically ChatGPT (Chat Generative Pre-Trained Transformer), in automated detection and explanation of values in software artifacts. Using ChatGPT, we investigate how mobile APIs align with human values based on their documentation. Human evaluation of ChatGPT's findings shows a reciprocal shift in understanding values, with both ChatGPT and experts adjusting their assessments through dialogue. While experts recognise ChatGPT's potential for revealing values, emphasis is placed on human involvement to enhance the accuracy of the findings by detecting and eliminating convincing but inaccurate explanations provided by the language model due to potential hallucinations or confabulations.

ARTICLE HISTORY

Received 21 August 2024
Accepted 3 March 2025

KEYWORDS

Human values; software;
ChatGPT; LLMs

1. Introduction

Considering human values is a critical imperative in software development (Mougouei et al. 2018; Perera et al. 2020). Breaching values may lead to user dissatisfaction, reputational and financial losses, and societal consequences (Bellamy et al. 2019; Hussain, Mougouei, and Whittle 2018; Ian 2023; Mougouei 2020; Perera et al. 2019). While not all violations of human values are publicised, some have made their way to the first page of the news articles. Social media platforms like X (Twitter) failing to curb cyberbullying against a particular gender (Maevé Duggan n.d.) resulted in users feeling disempowered and fostered a climate of distrust (Haleema Bhatti n.d.). Facial recognition software exhibiting biases and erroneous detection has compromised individual privacy and perpetuated racial biases, eroding trust among users (Joanne Cavanaugh Simpson n.d.). Another example is the failure of the CovidSafe

application sponsored by the Australian Government, which not only raised concerns about privacy and transparency but ultimately led to its removal due to user dissatisfaction and distrust (Jason Cartwright n.d.). The consequences can extend to the entire software ecosystem when value breaches occur in frequently used software libraries (APIs).

While there have been efforts to integrate human values into software, a significant obstacle, as acknowledged in existing literature, lies in the absence of practical understanding of values (Mougouei et al. 2018; Nurwidyanoro et al. 2022; Perera et al. 2020). This arises from the intricate and subjectively interpreted nature of values and their manifestations. This resembles a chicken-and-egg problem: understanding values is crucial for detecting their representations in software artifacts, yet without knowledge of these (practical) representations, a comprehensive understanding remains elusive. It is, therefore, important

CONTACT Davoud Mougouei  dmougouei@gmail.com  School of Information Technology, Deakin University, Burwood, Australia

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

to connect practitioners' understanding of values with the practical representations of those values. The existing state of research and practice is confined either to a limited subset of extensively examined values, with a significant focus on security and privacy (Baldassarre et al. 2020), overlooking the broader spectrum of human values (Schwartz 2012), or relies on the manual identification of values (Khan et al. 2022; Krishtul et al. 2022; Nurwidiantoro et al. 2023, 2022; Obie et al. 2021; Perera et al. 2020). However, a manual approach is laborious and susceptible to human bias due to subjective interpretations of values. Furthermore, current methods often lack transparency as they fail to explain their reasoning about values.

On the other hand, Large Language Models (LLMs), like ChatGPT, open a promising pathway in reasoning about human values in software artifacts. Their evident proficiency in automated analysis (Chen et al. 2021; Colavito et al. 2024) of large-scale textual content and the ability to learn from limited samples (Brown et al. 2020) positions them as noteworthy tools. LLMs can also explain their reasoning processes about human values, enabling a thorough evaluation of the findings about these values. Trained on a diverse corpus of text, these models cultivate nuanced language understanding, thereby enabling the recognition of values across diverse contexts (Brown et al. 2020). Although LLMs have demonstrated effectiveness in detecting patterns in software artifacts, e.g. in automated issue labelling (Colavito et al. 2024), there has not been any effort to utilise them for reasoning about values. Our study explores the potential of ChatGPT in unveiling the alignment of standard and third-party Android and iOS APIs with human values based on their documentation. The study aims to address the following research questions:

RQ1: What patterns of human values does ChatGPT reveal in mobile APIs?

RQ1.1: How do mobile APIs align with socially and personally focussed values?

RQ1.2: How do value alignments vary across different API themes?

RQ1.3: How do value alignments vary across Android/iOS APIs?

RQ2: How accurate are value alignments revealed by ChatGPT?

RQ2.1: How do experts evaluate value alignments across different values?

RQ2.2: How do experts evaluate value alignments across API themes?

RQ2.3: How do experts evaluate value alignments across Android/iOS APIs?

RQ2.4: What are the primary causes of disagreement among experts?

RQ2.5: How do experts rate the (endorsed) value alignments?

RQ3: How does ChatGPT's prompt completion vary for different mobile APIs?

RQ3.1: How does ChatGPT's response time vary for different mobile APIs?

RQ3.2: How does ChatGPT's response size vary for different mobile APIs?

RQ3.3: How does ChatGPT's prompt size vary for different mobile APIs?

RQ3.4: What is the interplay between prompt characteristics and ChatGPT's responses?

Note 1: This study does *not* use ChatGPT as a classifier that merely specifies the presence or absence of human values; instead, it utilises the reasoning capabilities of ChatGPT not only to identify but more importantly to explain the rationale for the value alignments.

Note 2: Currently, there is no automated technique for detecting and explaining human values and their diverse manifestations in software artifacts; we have evaluated the accuracy of ChatGPT findings through human assessment.

To address the research questions, we employ a methodology that combines ChatGPT with domain expertise to uncover the alignment of mobile APIs with human values. The study comprises three phases: Data Acquisition, Pilot, and Main Study. In the Data Acquisition phase, we sampled 3,000 standard and third-party Android and iOS APIs with diverse functionalities. ChatGPT mediated disagreements in the pilot study and, subsequently, revealed human values. The Main Study used insights from the pilot study to develop an automated framework for detecting value alignments in API documentation. The framework capitalises on ChatGPT's ability to justify its findings, fostering discussions about values. To ensure accuracy, a panel of six software engineering experts evaluated a substantial portion of value alignments identified by ChatGPT, striving to reconcile disagreements with the language model through discussions.

An intriguing finding revealed a significant reciprocal shift in understanding human values between experts and ChatGPT during discussions, resolving the majority (96%) of disagreements. Following the debates, experts changed their opinions to align with a notable percentage of ChatGPT's identified value alignments, highlighting ChatGPT's ability to reveal subtle

Sensitive Design (VSD) (Friedman et al. 2013), Continual Value(s) Assessment (CVA) (Perera et al. 2020), and the Human Values Dashboard (Nurwidyantoro et al. 2023), have been proposed to integrate human values into the software development process. However, a significant challenge, as emphasised in the literature (Hussain, Mougouei, and Whittle 2018; Mougouei et al. 2018; Mougouei 2020; Perera et al. 2019, 2020), lies in the practical understanding of human values. Values are context-dependent and manifest in various forms, making them hard to define. While certain values like security and privacy have been studied, the broader spectrum of human values remains under-explored, making it challenging to define and operationalise them in software development (Mougouei et al. 2018). Moreover, human values are often implied rather than explicitly stated (Hanel et al. 2018). This further compounds the complexity of making inferences about values; reasoning is needed and a simple keyword-based search does not suffice to capture the complexity of values and their diverse manifestations (Jamieson, Yamashita, and Foong 2024). Last, automated approaches are needed to make inferences on human values in software artifacts as a manual approach will not be scalable and is prone to human error and personal biases.

2.2. Large language models for reasoning about human values

Large Language Models (LLMs) such as ChatGPT are shaping the future of software engineering, showcasing robust capabilities as dialogue systems and leaving a notable impact on qualitative studies utilising NLP techniques (Guo et al. 2023). Comparative analyses with fine-tuned LLMs like BERT and RoBERTa across various tasks reveal that ChatGPT outperforms RoBERTa in specific scenarios (Qin et al. 2023; Zhong et al. 2023). Noteworthy achievements include ChatGPT's proficiency in categorising implicit hate speech and excelling in automatic genre identification tasks, outperforming humans in providing explanations (Huang et al. 2023; Kuzman, Mozetič, and Ljubešić 2023). This positions ChatGPT as a promising tool for data annotation, particularly in the identification of implicit hateful tweets. In Software Engineering, ChatGPT has proven its significance, displaying effectiveness in diverse software development tasks, including software requirements generation (Beganovic, Jaber, and Almisreb 2023; Khojah et al. 2024). Experimental studies comparing its performance with human participants underscore its potential integration into development teams, leading to a transformative shift in how virtual members

could contribute to, coach, and even execute portions of development work (Bera, Wautelet, and Poels 2023). However, researchers caution for guided integration of AI (Artificial Intelligence), emphasising structured and supervised usage (Daun and Brings 2023).

Large Language Models (LLMs), like ChatGPT, can go beyond their traditional role in software engineering to uncover and make inferences about human values in textual software artifacts (Fischer, Luczak-Roesch, and Karl 2023). Recent advancements, especially in models such as ChatGPT, demonstrate their ability to navigate complex aspects of human values. The automated reasoning capabilities of LLMs make them valuable for revealing implicit values in software artifacts (Chen et al. 2021). Trained on diverse corpora, LLMs develop complex language understanding, enabling the recognition and interpretation of values in code, documentation, and user interfaces (Brown et al. 2020). The combination of contextual reasoning and pattern recognition in LLMs plays a pivotal role in reducing human bias and promoting transparency in integrating values into software artifacts (Brown et al. 2020). Moreover, the ability of LLMs to explain their findings enhances the reliability and transparency of any findings about human values, facilitating comprehension for developers and stakeholders. This positions LLMs as promising tools for automated reasoning about human values (Alshami et al. 2023).

While the existing literature has recognised the potential of LLMs for various applications in software engineering (Beganovic, Jaber, and Almisreb 2023; Ma et al. 2023; Rahmaniar 2023), including requirements analysis (Zhang et al. 2023), improving code quality (Kou et al. 2024; Surameery and Shakor 2023; White et al. 2024), software evaluation (Jalil et al. 2023; Khojah et al. 2024; Ronanki, Cabrero-Daniel, and Berger 2022; Wang et al. 2024), supporting agile development (Bera, Wautelet, and Poels 2023), and software engineering education (Daun and Brings 2023; Lehtinen, Koutcheme, and Hellas 2024), a key gap remains in leveraging LLMs for reasoning about human values despite the promise of LLMs in ethical judgments (Hendrycks et al. 2021).

Despite their remarkable capabilities, LLMs face challenges, particularly with hallucinations (Bellini-Leite 2024; Smith, Greaves, and Panch 2023) and confabulations (Smith, Greaves, and Panch 2023). Hallucinations occur when a language model generates convincing yet ungrounded arguments or presents a selective, misleading view of reality (Huang et al. 2023). For example, LLMs may provide seemingly plausible yet misleading explanations rooted in patterns from their training data, erroneously presenting them as

factual (Smith, Greaves, and Panch 2023). These issues are especially prevalent in multilingual translation models, where hallucinations often arise in low-resource language pairs or translations from English (Guerreiro et al. 2023). For instance, Tamil translations exhibit a concerning rate of hallucinations, with over one in eight cases containing toxic text linked to training data, highlighting critical safety risks (Guerreiro et al. 2023). Similarly, ChatGPT has been observed to produce off-target translations, such as unrelated responses or excessive generation (Guerreiro et al. 2023). Some of the existing works favour the term ‘confabulation’ over ‘hallucination’ (Smith, Greaves, and Panch 2023), arguing this term more accurately describes false but coherent information generated by the language models. LLMs also exhibit limitations in terms of bias. Studies have shown that LLM-generated content demonstrates gender and racial biases (Fang et al. 2024). These biases stem from various sources throughout the LLM lifecycle, including training data, model specifications, and policy decisions (Ferrara 2023). In educational contexts, the integration of LLMs raises concerns about exacerbating inequalities, as biases can originate from multiple steps in the development and application process (Lee et al. 2024). Furthermore, LLMs tend to exploit dataset biases and artifacts as shortcuts for prediction, hindering their generalizability and adversarial robustness (Dogra et al. 2024). Addressing these biases requires a multifaceted approach including the development of more equitable and transparent AI systems, as well as improved methods for identifying, quantifying, and mitigating biases in language models (Ferrara 2023). These limitations underscore the importance of incorporating a human-in-the-loop approach, where human oversight can ensure the contextual relevance and accuracy of LLM inferences. Integrating human oversight remains a critical component to addressing challenges such as hallucinations and confabulations, enabling LLMs to deliver outputs that are not only reliable but also contextually and ethically informed.

This study demonstrates how ChatGPT, a large language model, can be utilised to make inferences about human values in mobile API documentation to not only identify the manifestations of human values but more importantly explain them to enhance the credibility of such findings. We emphasise the importance of human oversight by performing a manual evaluation of a subset of ChatGPT findings through an interactive dialogue-based approach to discuss with the language model its findings and identify potential hallucinations or confabulations.

3. Methodology

This study aims to investigate the effectiveness of large language models, specifically ChatGPT, in automated detection and explanation of human values in software artifacts. To explore the potential of large language models in revealing human values in software artifacts, we have conducted a case study to identify the alignment of Android and iOS APIs with human values based on their documentation. Human values are examined in the API documents from the app user’s perspective using ChatGPT in combination with expert opinions. The study comprises three phases as outlined in Figure 2: Data Acquisition, Pilot Study, and Main Study.

3.1. Data acquisition

We gathered a dataset of around 10,000 APIs comprising both standard and third-party Android and iOS APIs. The dataset comprises 647 standard Android APIs (Java) sourced from the Android API Reference¹ and 148 iOS APIs (Swift) retrieved from the Apple Developer Documentation². The dataset also includes 9,291 open-source third-party APIs, comprising 6,916 for Android and 2,375 for iOS, gathered from GitHub. Prior to analysis, an initial data cleaning was performed to eliminate APIs lacking proper documentation. Subsequently, a subset of 3,000 APIs was randomly and proportionately (stratified) selected.

3.2. The pilot study

We designed a pilot study to evaluate the alignment of APIs with human values from a user perspective. Initially, a set of API documents was chosen, and two coders (SE experts – coauthors) labelled them based on their alignment with values. The labelling process revealed concerns, with the only a fair level of agreement (Cohen’s Kappa = 0.33) between the coders and issues related to laborious manual labelling. Subjective judgments by coders made justifications challenging. Coder availability added complexity to manual labelling. To address these issues, ChatGPT was used as a mediator to reconcile coder disagreements. Disagreements were discussed with ChatGPT, and its justifications were used for reconciliation. Coders followed a defined criterion for determining API alignment. As the study progressed to a larger-scale analysis, automation concerns emerged due to the labour-intensive nature of providing prompts to ChatGPT. Error handling became crucial for consistency and robustness in detecting human values’ patterns in API documents.

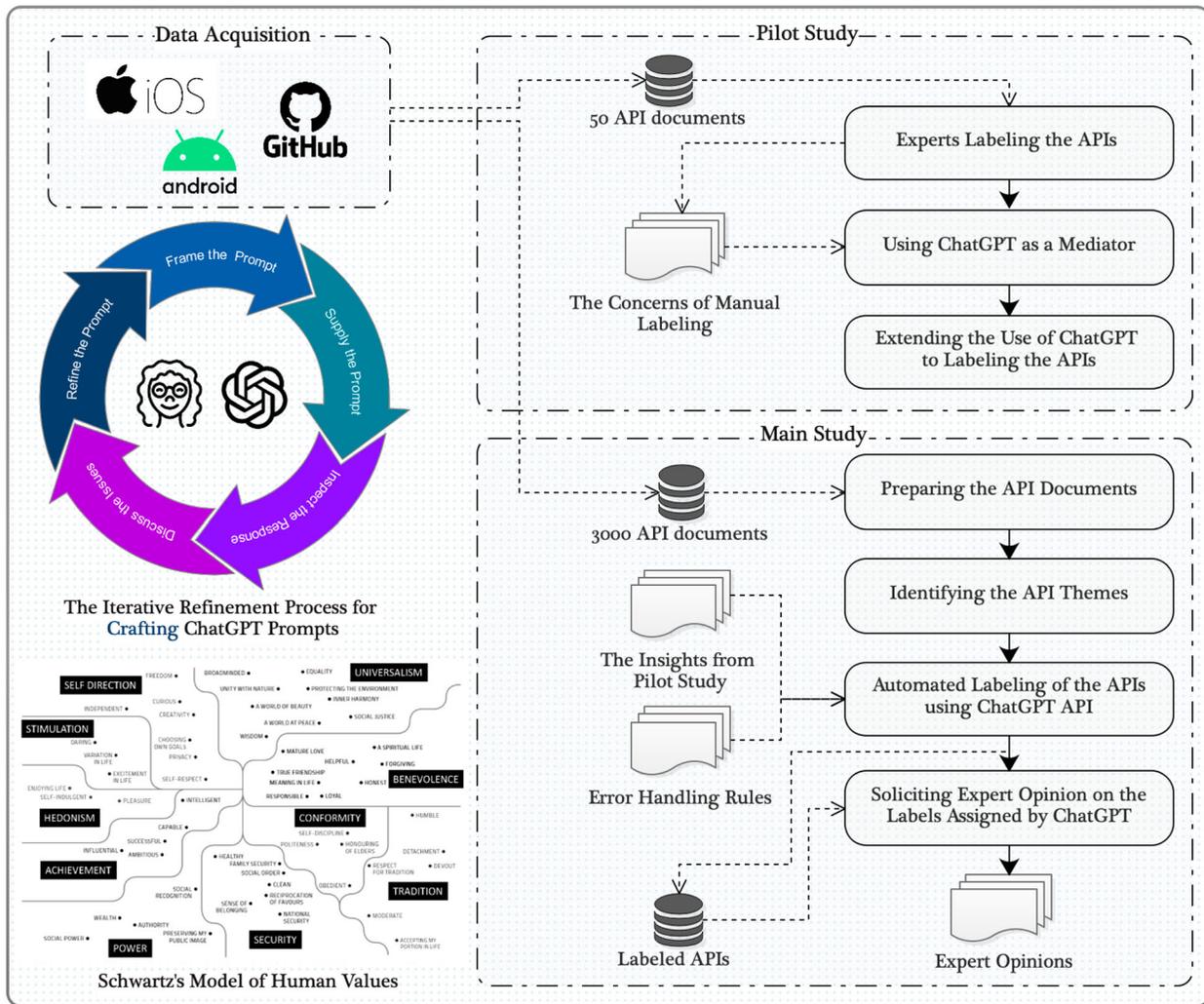


Figure 2. An overview of the methodology.

3.3. The main study

As depicted in Figure 2, the main study builds upon insights from the pilot study to develop a comprehensive framework for identifying the alignment of APIs with values based on their documentation. Leveraging the potential of ChatGPT, the framework prioritises automation to handle the large-scale labelling of APIs while handling occasional errors in responses. Implemented in Python³, the framework utilises the ChatGPT API (gpt-3.5-turbo) to automate labelling, addressing errors in content-related outputs and technical interactions. We employ extractive text summarisation to respect ChatGPT's token limitation constraints (El-Kassas et al. 2021). Error handling, including filtering out invalid justifications and implementing a retry mechanism for technical issues, ensures the study's findings' quality and accuracy. To validate ChatGPT's labels, 10% are assessed by a panel of 6 SE experts (coauthors of the papers), who provide opinions and

reconcile their disagreements with the language model when possible (Figure 2).

3.4. Iterative refinement process

The ChatGPT prompts used in the pilot and main studies were developed through an iterative refinement process referred to as Prompt Crafting, as outlined in Figure 2. Similar processes are referred to by literature as Prompt Engineering (Giray 2023). However, we prefer the term 'crafting' over 'engineering' as it highlights the creative, complex, and trial-and-error nature of developing prompts for language models. This terminology also recognises the ad-hoc characteristics of the interplay between prompts and responses from the language model, highlighting how sensitively the model's responses depend on the framing of the prompts. This is particularly important in the context of the GPT models as their behaviour keeps changing

online not only through the stochastic nature of these models but also due to the fact that even the same model, e.g. GPT-3.5, keeps being updated with unknown impacts on the interpretations of the prompt; the same prompt may produce accurate responses today but not tomorrow. This lack of transparency is a major limitation of large language models, which restricts the reproducibility of any findings revealed using the exact same prompt with a given model (Liao and Vaughan 2023; Zhao et al. 2024).

The prompt crafting process began with framing a prompt. This prompt was submitted to ChatGPT, and its response was inspected for accuracy based on whether the identified value alignments were consistent with Schwartz's theory of basic values, clearly justified, aligned with provided examples, traceable to the API documentation, and reflected user perspective, not developers. Potential issues, such as misinterpretations of values, vague (or unsupported) reasoning, failure to identify implicit connections, and confusion between user and developer perspectives, were discussed. Through several iterations, the following actions were taken to refine prompts for higher accuracy: clarifying instructions, explicitly providing the definition of human values from Schwartz's work (Schwartz 2012; Schwartz et al. 2012) in the prompt, providing sample explanations linking API documentation to values, and emphasising prioritisation of more relevant and well-supported value alignments.

For instance, we noticed that if ChatGPT was asked to identify expressions of values without emphasising the directness of these expressions, the language model sometimes overextended its reasoning to establish connections between the API document and values, even when unrealistic. Conversely, emphasising only direct links limited the model's reasoning capabilities, preventing it from uncovering implicit links, sometimes even ignoring evident links. To address this dilemma, we: explained to ChatGPT what we mean by direct links to ensure tractability, provided sample explanations directly linking API documentation to values, and instructed the model to identify all value alignments while prioritising the top three most relevant ones. These refinements produced more balanced responses, though some inconsistencies remained. Another issue revealed during the inspection was ChatGPT's confusion between the API user (the developer who uses the API to create a mobile app) and the end user (the app user who interacts with the API as part of the app). This issue was largely mitigated by explicitly defining the app user in the prompt and emphasising that we sought manifestations of values from the app

user's perspective, not the developer's. Nevertheless, occasional confusion persisted.

These enhancements occurred through multiple iterations. The process was repeated until sample responses from ChatGPT met the above-listed criteria, including relevance, consistency, and clarity in linking API documentation to values. Responses were also qualitatively evaluated for completeness and logical reasoning.

To mitigate biases associated with subjective prompt choices, the iterative refinement process was conducted collaboratively, involving multiple co-authors to ensure diverse input. Once the prompts were finalised, they were fixed and uniformly applied without any further modifications. In the pilot study, experts consistently used these finalised prompts. In the main study, all prompts were executed exactly as finalised, using ChatGPT API (not the online chat interface), ensuring uniformity.

4. Pilot study

In the pilot study, a set of 50 Android and iOS APIs was randomly selected and two coders (SE experts, coauthors) evaluated the alignment of the APIs with values (e.g. Self-Direction) manifested by their associated value items (e.g. Privacy) (Schwartz 2012). An API was deemed aligned with a value if, based on its documentation, it contributed to one of its manifestations (value items) from the user's perspective. ChatGPT was initially used to help reconcile the disagreements between the coders. The reconciliation process involved asking ChatGPT's opinion on the alignment of the APIs with the values (value items) where the coders could not agree. ChatGPT's argument was put forth to the coders instigating further discussion. The use of ChatGPT was then extended to identify the value alignments primarily.

4.1. Experts labelling the APIs

Each coder was a software expert (coauthor) with substantial practical experience in the field. Before the labelling process, the coders underwent a training phase to establish a shared understanding of Schwartz's model of values and the labelling process.

The training process comprised four main components. First, the coders studied seminal works that explain Schwartz's theory of basic values (Schwartz 1992, 2012), including the definition of values and their interrelationships. Second, the coders reviewed several publications about human values from flagship software engineering venues (Hussain, Mougouei, and

Whittle 2018; Hussain et al. 2020; Mougouei et al. 2018; Mougouei 2020; Nurwidyanoro et al. 2023, 2022; Perera et al. 2019, 2020) to understand human values in the context of software. Third, the coders watched a video tutorial that demonstrated how to detect human values in API documentation and how to record their opinions in a spreadsheet; this video was prepared by one of the coauthors (author X), who had previous experience in detecting human values in software artifacts and did not participate in the pilot study. Fourth, the coders initially coded a small sample of the APIs, after which feedback was provided by author X only regarding compliance with the definitions of values from Schwartz's theory.

The inter-rater agreement level was calculated using Cohen Kappa; despite the initial training, only a fair level of agreement (Kappa = 0.33) was reached, overall. A moderate level of agreement was calculated for the value of Universalism (Kappa = 0.48) while the coders only slightly agreed on the alignment of the APIs with Self-Direction and Stimulation. The coders did not declare any API aligned with Tradition. A follow-up discussion revealed the following concerns.

Concern 1: Labeling is a tedious task. The assessment of APIs' alignment with values proved to be a laborious task. In particular, the practicality of evaluating the alignment of the APIs with a large number of values and value items without sacrificing accuracy was questioned by the coders. Similar works have echoed this concern (Nurwidyanoro et al. 2023, 2022; Perera et al. 2020).

Concern 2: Arduous labelling leads to bias. The coders often expressed feelings of monotony, leading to a decline in their attentiveness compared to the initial stages of labelling. As a result, they resorted to making assumptions and drawing conclusions to expedite the labelling process. For instance, they deduced that certain types of graphical user interfaces fostered creativity and, thus, were relevant to the value of Creativity. Such conclusions, based on arguable justifications, raised concern about the reliability of the labels.

Concern 3: Subjective judgments and lack of justification make it hard to verify. When coders were asked to provide justification for the labels, they often found it difficult; the justifications were in many cases arguable, lacking strong evidence in the API documentation. Expressions such as 'It might ... but I am not sure' were frequently used.

Concern 4: Direct versus Indirect links. The coders documented that the alignment of the APIs with values was not entirely straightforward in many cases and necessitated a certain level of reference. A seemingly straightforward remedy might involve solely considering the direct links (alignments). However, this gives rise to a

more intricate predicament for the coder to determine: 'What degree of directness qualifies as direct?'

The primary aim of the pilot study was to demonstrate the inherent difficulties for humans in evaluating 59 value items (under 10 value categories) with subjective interpretations and diverse manifestations in software artifacts, while accurately detecting and explaining them. This critical challenge has somehow been downplayed in existing literature, where studies report a high level of interrater agreement (Nurwidyanoro et al. 2022; Perera et al. 2020) achieved through discussions among raters—yet fail to document the rationale behind their evaluations or the reasons for agreement or disagreement. This lack of transparency makes their results difficult to verify. In such cases, reporting a high Kappa score may present a misleading picture by masking the inherent subjectivity involved in interpreting values and overlooking the significant limitations of manual approaches in accurately detecting and explaining values at scale. The pilot phase in our study was structured to highlight the limitations of manual approaches in detecting and explaining value alignments, rather than to establish final interrater reliability. Consequently, the Kappa value was computed before attempting to resolve expert disagreements. Based on related studies (Nurwidyanoro et al. 2022; Perera et al. 2020), it is foreseeable that a higher Kappa score could have been achieved had we pursued such resolutions. However, doing so would have masked the challenges inherent to manual labelling—specifically those outlined in Concerns 1–4—including the high degree of subjectivity that varied across different values, as highlighted.

To maintain objectivity and surface these challenges, disagreements were intentionally not reconciled during the pilot phase. Instead, the focus was on exposing the limitations of manual coding approaches and revealing the natural variations in interpretation that arise when evaluating human values. These challenges were then systematically addressed in the main study. Rather than avoiding the subjectivity inherent to human values, we proposed an approach that embraces it through human-LLM dialogue—enabling disagreements to be debated and clarified, rather than concealed. This methodology mirrors real-world value evaluations, where interpretations often evolve through deliberation rather than static agreement. This interactive approach naturally led to higher levels of agreement between experts and ChatGPT, as well as substantial inter-rater reliability (Kappa = 0.78) and agreement rates (86%) in the main study. These metrics were calculated based on the evaluation of ChatGPT's explanations for value alignments—not merely detecting the presence or absence of values,

as often done in prior work. As such they cannot be compared against the metrics from the pilot study as they are computed under different settings.

4.2. ChatGPT as a mediator

We employed the language model of ChatGPT as a mediator for reducing ambiguities in the coders' comprehension of values and exploring the possibility of leveraging ChatGPT to facilitate the resolution of their disagreements. This was motivated by Concerns 1–4 articulated in Section 4.1. ChatGPT's opinion was sought in the presence of both coders to ensure impartiality. Following several refinement cycles, we devised a template for the prompts, exemplified in the following instance.

ChatGPT's opinion about aligning an API with a value item.

 Consider the following API documentation and Schwartz's model. Assess the alignment of the API with the value item of 'Variation in Life' from an app user perspective. Deem an API aligned with a value item if it contributes to supporting, enhancing, or considering that value item in the mobile app, from a user perspective. Provide a brief justification for your response. '<https://github.com/pvasa/cameraview-ex>'.

 The API can be considered aligned with the value item of Variation in Life. The API allows users to capture and explore various moments, adding variation to life. The API enables users to capture photos and videos, which can contribute to experiencing different aspects of life and adding variety to personal experiences ...

ChatGPT was used to reconcile the cases of disagreement with the following breakdown. In around 80% of the cases, both coders agreed with the justification provided by ChatGPT regarding the alignment of the APIs with values. In other cases, the coder did not agree with ChatGPT's initial justification, seeking further clarification. The clarification was evaluated as either 'convincing' or 'arguable'. The coders were, at large, impressed by the level of detail provided in ChatGPT's justifications and described them as 'rationally grounded' and 'worthy of investigation.' Nonetheless, ChatGPT's ability to detect the connections between the API documents and values demonstrated sensitivity to how the prompts were framed. Notably, when terms such as 'directly aligned' were emphasised in the prompts to enhance the labels' precision presumably,

ChatGPT tended to declare nearly all APIs as not directly aligned with values, exhibiting an overly cautious behaviour. This observation also further emphasised the concern raised by the coders during the labelling process about the difficulty of determining the 'degree of directness' (Concern 4 in Section 4.1).

4.3. Extending the role of ChatGPT to labelling the APIs

The use of ChatGPT as an intermediary exhibited the potential to provide reasonable justifications concerning the alignment of the APIs with values, thereby helping reconcile the disagreements between the coders. Nonetheless, in such instances, the use of ChatGPT was subject to partial guidance, primarily from prompting ChatGPT to evaluate the alignment of the API with a specific value which was the place of argument between the coders. We extended the role of ChatGPT beyond a mediator to detect the alignment of the APIs with human values. Therefore, the prompt used in Section 4.2 was revised to allow for labelling the APIs. To ensure a consistent and reliable stream of responses, we executed several iterations of the iterative refinement process on the prompt as outlined in Section 3 (Figure 2). Also, we explicitly provided Schwartz's model in the prompt to avoid inconsistencies.

ChatGPT to detect the alignment of an API with human values.

 Consider the following API documentation. Assess the alignment of the API with the values in the Schwartz model from the app user's perspective. Consider an API aligned with a specific value if it is aligned with at least one value item associated with that value. An API is aligned with a value item if it contributes to supporting, enhancing, or considering that value item for the app user. Provide clear justification for each alignment ... '<https://github.com/recruit-mp/RMPZoomTransitionAnimator>.' 'Schwartz's model: values & value items'

 Sure, here are the value alignments ...

- Self-Direction (Freedom): The API allows users to customise the animation in their mobile app; giving them the freedom to personalise the visual experience.
- Hedonism (Enjoying Life): The API provides a smooth and visually pleasing animation, which enhances the enjoyment and pleasure of the app users.

ChatGPT's responses were found promising, yet exhibited sensitivity to the framing of the prompts, e.g. the presence or absence of certain words led to declaring or not declaring APIs as aligned with values. While the primary focus rested on soliciting justifications exclusively from the app user's standpoint, it was observed that ChatGPT occasionally provided justifications relating to the alignment of the APIs to developer values. For instance, it elucidated, 'The API's provided helper methods simplify the process of sending and receiving data, offering a helpful toolkit for users developing wearable apps.' Such justifications were omitted, highlighting the necessity of accounting for inaccurate justifications. That was considered in crafting the prompt used in the main study. Although manual execution of the queries in the ChatGPT's online platform (GPT-3.5 2025) was practical for the relatively small number of the APIs (50) we studied in the pilot phase, and the concern was raised by the experts that this approach would be laborious for a more significant number of APIs, hence the need for automation.

5. Main study

The challenges and concerns highlighted in the pilot study served as guiding principles in refining the proposed methodology to better harness the capabilities of ChatGPT in revealing human values in API documentation. One of the key takeaways from the pilot study was the sensitivity of ChatGPT's responses to the way prompts were framed. In light of this, we refined the prompts supplied to ChatGPT to enhance consistency and accuracy in labelling the APIs. Drawing on the iterative refinement process, we crafted an improved prompt to reveal human values in API documentation with more accurate justifications. Also, the need to address potential errors in responses from ChatGPT and the laborious process of manually supplying the prompts to ChatGPT via its online platform underscored the necessity of automation. The main study thus aimed to develop a framework (implemented in Python) based on ChatGPT to automatically unveil human values in the API documents – from a user perspective. We employed ChatGPT API for gpt-3.5-turbo (GPT-3.5 2025) with the default parameter options for consistency with ChatGPT's online platform used in the pilot study, which also utilises the gpt-3.5 architecture. The language model gpt-3.5-turbo offers the latest capabilities of gpt-3.5 (GPT-3.5 2025).

5.1. Preparing the API documents

A Python script was developed to extract the corresponding API documents from the URLs and summarise them

when needed. The need for summarisation was raised from the limitation of the language model gpt-3.5-turbo on the maximum of the allowed tokens in each transaction; around 4,000 tokens were allowed for the prompt and completion combined (GPT-3.5 2025). The prompt size was around 1200 tokens, and we pre-allocated a minimum of 1000 tokens to the response (completion) – based on the sample responses received from ChatGPT; the remaining 1800 tokens were referred to as floating tokens and distributed – as needed – between the API document and response. When the API document neared the maximum allowed tokens, we utilised the LexRankSummarizer, an automatic technique based on the LexRank algorithm (El-Kassas et al. 2021), to summarise it. LexRank employs a graph-based approach for extractive summarisation (El-Kassas et al. 2021), assessing sentence importance through relationships and centrality within the document. This method selects pivotal sentences in the input document(s) and concatenates them for the summary (El-Kassas et al. 2021). Post-summarisation, approximately 96% of the original API document size was retained. We opted for extractive summarisation due to its advantages over alternatives, such as abstractive summarisation (El-Kassas et al. 2021). It preserves original sentences from the input text, maintaining faithfulness to the source material. Additionally, it relies on existing sentences, mitigating the risk of generating inaccurate or contextually incorrect information present in abstractive methods. The simplicity of implementation is another advantage, as it avoids the generation of novel phrases or sentences.

5.2. Identifying the API themes

To better understand the alignment of the APIs with human values, they were grouped under 10 main themes (Table 1) based on their primary functions specified by the publishers as the API topic (category); 902 API topics (categories) were grouped under 10 themes using ChatGPT. The language model provided descriptions of each theme with examples to justify grouping the API topics under the themes and make it explainable (Gunning et al. 2019). These descriptions and the 10 most prevalent topics under reach theme are listed in (Table 1) with examples. We realised that some of the API topics, mainly for standard Android ('Android Platform', 'AndroidX', and 'Android Support Library') and iOS ('App Services' and 'System') libraries, were too generic, hence they and their respective APIs were grouped under the theme 'Development Essentials,' which was meant to capture general-purpose libraries. However, a closer look at those topics revealed that many APIs could be associated with more specific

Table 1. The API themes and their 10 most prevalent topics.

Theme	Description	Topic: Sample API
User Interface Design	Encompasses libraries and frameworks providing a wide range of user interface components and design elements to enhance the visual appeal and user experience of mobile apps.	AndroidX: <code>androidx.drawerlayout.widget</code> ; Android Support Library: <code>android.support.cursoradapter</code> ; Material Components: <code>com.google.android.material.circularreveal.cardview</code> ; Android Platform: <code>android.view.inputmethod</code> ; App Services: <code>UserNotificationsUI</code> ; Maps: <code>AnyMaps</code> ; Graphics: <code>ProcessingKit</code> ; Cards: <code>TisprCardStack</code> ; Switch: <code>TKSwitcherCollection</code> ; Button: <code>MultiToggleButton</code>
Development Essentials	Includes general-purpose libraries and tools for mobile app development. That includes standard platform-specific functionality (e.g. <code>android.support</code> , core services), support libraries, architectural patterns, IO, and utilities (e.g. <code>java.util</code>).	Android Platform: <code>android.system</code> ; Android Support Library: <code>android.support.v13.view</code> ; Play Core Library: <code>com.google.android.play.core.install.model</code> ; System: <code>Objective-C Runtime</code> ; Functional Programming: <code>reactive-tiny</code> ; AndroidX: <code>androidx.arch.core.util</code> ; Architecture Components: <code>android.arch.persistence.room.testing</code> ; Localization: <code>android-localisation-helper</code> ; FHSTwitterEngine Twitter API for Cocoa developers.: <code>PPEventRegistryAPI</code> ; Extensions (Xcode 8+): <code>XcodeEquatableGenerator</code>
Connectivity & Networking	Includes APIs for network communication, covering HTTP requests, sockets, and RESTful APIs, social media, and news-related functionalities. Also, APIs related to Facebook, Twitter, and Instagram as well as displaying news.	Android Platform: <code>org.xml.sax.helpers</code> ; AndroidX: <code>androidx.browser.customtabs</code> ; App Services: <code>WatchConnectivity</code> ; System: <code>SystemConfiguration</code> ; Bluetooth: <code>MultiPeer</code> ; Web: <code>MapKit JS</code> ; Location: <code>PhoneNumber</code> ; GCD: <code>GrandSugarDispatch</code> ; Networking: <code>Retrofit-DSL</code> ; Android Support Library: <code>android.support.v4.hardware.display</code>
Developer Productivity	Focuses on libraries and tools that enhance developer productivity, including developer tools, logging frameworks, code generation, and performance optimisation tools.	Keyboard: <code>KeyboardObserver</code> ; Debugging: <code>AEConsole</code> ; Logging: <code>LoggerPreferences</code> ; Android Platform: <code>android.util</code> ; Tools: <code>ForecastIO</code> ; Developer Tools: <code>ScriptingBridge</code> ; Updaters: <code>AppUpdater</code> ; Other Xcode: <code>dsnip</code> ; Decompilers: <code>TTDeDroid</code> ; Event Buses: <code>Eventex</code>
Multimedia & Content	Focuses on APIs and tools for handling multimedia-related functionalities, such as audio, voice (e.g. <code>android.service.voice</code>), video, speech, animations, image processing, and TV (e.g. <code>android.media.tv</code>). Examples also include Spruce iOS Animation Library and <code>VisualEffectView</code> .	Media: <code>ImageCaptureCore</code> ; AndroidX: <code>androidx.dynamicanimation.animation</code> ; Android Platform: <code>android.media.browse</code> ; GIF: <code>Android GIF Loader</code> ; Android Support Library: <code>android.support.media.compat</code> ; Camera: <code>CameraBackground</code> ; PDF: <code>UllImagePlusPDF</code> ; Video: <code>ios</code> ; Audio: <code>AudioKit</code> ; Image Loaders: <code>CatKit</code>
Data Management & Analysis	Includes libraries and tools for managing and analyzing data, including working with databases, filesystems, storage. Examples include <code>SQLite</code> , <code>android.os.storage</code> , and ORMs.	Android Platform: <code>org.json</code> ; AndroidX: <code>androidx.room</code> ; Database: <code>SQLBrite</code> ; Ratings: <code>RateBottomSheet</code> ; Purchases: <code>Android Beacon Library</code> ; App Services: <code>Core Spotlight</code> ; File System: <code>RxFileObserver</code> ; Data Structures: <code>Agrona</code> ; Dex: <code>Dex 65536</code> ; ORM: <code>Slim Repo</code>
Testing & Debugging	Includes libraries, frameworks, and tools for testing and debugging mobile applications, covering unit testing, UI testing, automation, performance testing, and code analysis. Examples include <code>Android Testing</code> .	AndroidX Test: <code>androidx.test.core.view</code> ; Test Support Library: <code>android.support.test.espresso.intent.rule</code> ; Android Platform: <code>android.app assist</code> ; Code Quality: <code>spacecommander</code> ; Analytics: <code>Segment</code> ; Benchmarking: <code>TinyDancer</code> ; Crash Reports: <code>BugShaker</code> ; Tag: <code>RKTagsView</code> ; AndroidX: <code>androidx.fragment.app.testing</code> ; UI Testing: <code>ViewInspector</code>
Security & Privacy	Covers libraries that enhance security and privacy in mobile apps, including encryption, authentication, secure network communication, user accounts, and biometrics. It also covers feature that enhance security and privacy of financial transactions and data. Examples include secure payment gateways, in-app purchases, and payment methods in mobile apps.	Android Platform: <code>android.drm</code> ; Permissions: <code>KPermissions</code> ; Security: <code>SwiftPasscodeLock</code> ; Fingerprint: <code>RxBiometric</code> ; Password Inputs: <code>OTP View</code> ; Authentication: <code>LinkedInSignIn</code> ; Credit Cards: <code>EditCredit</code> ; Face Recognition: <code>FaceDetector</code> ; OAuth: <code>AppAuth WebView</code> ; System: <code>LocalAuthentication</code>
Wearable & IoT	Revolves around libraries and tools for wearable devices (including health) and smart gadgets. It includes APIs like <code>HomeKit</code> , <code>MQTT</code> , <code>Zigbee</code> , and others for building applications that communicate with IoT devices. It encompasses APIs such as <code>HealthKit</code> for collecting health and fitness data, connecting with wearable devices, and tracking users' well-being.	Wearable Library: <code>android.support.wearable.media</code> ; Android Things: <code>com.google.android.things.userdriver</code> ; Other Hardware: <code>UIDeviceComplete</code> ; Wearable Data Layer: <code>WearSocket</code> ; Android Support Library: <code>android.support.wear</code> ; AndroidX: <code>androidx.wear.ambient</code> ; App Services: <code>HealthKit</code> ; Android Platform: <code>android.service.vr</code> ; Watch Face: <code>Analog Watch Face</code> ; <code>WearSharedPreferences</code> : <code>Preferences</code>
Gaming & VR	Focuses on libraries and frameworks designed for game development, augmented reality, or virtual reality on mobile apps. Examples include: libraries that enable 2D and 3D graphics and augmented reality integration.	Graphics and Games: <code>Metal</code> ; Games: <code>AssetImportKit</code> ; Physics Engines: <code>Bullet</code> ; 2D Engines: <code>flixel-gdx</code> ; Gamepads: <code>AndroidJoyStickView</code> ; 3D Engines: <code>jMonkeyEngine</code> ; Particle Systems: <code>Leonids</code> ; OpenGL: <code>modern-java6-android-glfwframework</code> ; <code>JustWeEngine</code> : <code>2D Engines</code>

themes such as ‘User Interface Design’ and ‘Connectivity & Networking.’

To achieve a more fine-tuned grouping of the APIs under the themes, we refined our initial grouping – mainly for standard Android and iOS libraries – to consider the API topics listed by the publishers and the

documentation of those APIs. In doing so, a ChatGPT prompt was crafted through the iterative refinement process described in Section 3 (Figure 2). ChatGPT API (gpt-3.5-turbo) then was prompted to specify the themes for the standard Android APIs listed under ‘Android Platform’, ‘AndroidX’, and ‘Android Support

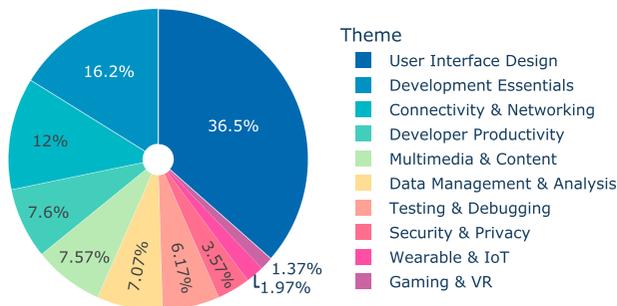


Figure 3. The distribution of mobile API themes.

Library’ as well as for the iOS APIs listed under ‘App Services’ and ‘System’. That led to reassigning themes for 465 APIs; the rest of the APIs preserved their initial themes specified by mapping their listed topics to the themes. Our decision to utilise ChatGPT for fine-tuning the grouping of the generic API topics stems from the nature of our research. Unlike situations where traditional topic modelling techniques such as LDA might excel (Oliveira et al. 2022, 2023), the proposed study involved refining the alignment of the known topics (provided by the API publishers) with their overarching themes rather than uncovering latent ones. On the other hand, ChatGPT’s strength lies in its capacity to comprehend and synthesise textual data in a contextually aware manner. This made it a more fitting choice to leverage the known API topics and documentation to ensure the themes represented the APIs’ functionalities.

Figure 3 shows the overall distribution of the APIs across different themes. The percentages on top of each bar represent the contribution of each theme to the total number of APIs. The themes ‘User Interface and Design’ and ‘Development Essentials’ contain the highest number of APIs. The lowest number are associated with the themes ‘Gaming & VR’ and ‘Wearable and IOT development.’

5.3. Automated labelling of the APIs using ChatGPT-3.5 API

The pilot study highlighted the limitation of manually promoting ChatGPT and the need for error handling. To achieve this, we utilised ChatGPT API, and the prompt supplied to ChatGPT was refined to enable automation and reduce the possibility of erroneous labels. Also, error handling was built into the process to mitigate any residual error in the labels.

5.3.1. Refining the prompt for automated labelling

The refinement of the prompt for automated labelling encompasses two primary considerations: (i) the reduction of erroneous labels characterised by inaccurate justifications, and (ii) the enhancement of the formatting of the

output generated by ChatGPT API to improve the interpretability of the output and optimise token usage, taking into account the constraints imposed by token limitations. Upon conducting an initial analysis of the labels generated by ChatGPT API, instances of labels containing inaccurate justifications came to light. Notably, certain justifications exhibited phrasing suggesting their grounding in the values of developers, notwithstanding the prompt’s explicit focus on identifying user values. Subsequent investigation revealed that the description pertaining to ‘user’ and ‘user value’ lacked the requisite clarity for the language model to consistently discern between the end user of the mobile application and its developer. This confusion resulted in a misunderstanding of perspectives in the results generated by ChatGPT API. To address this, a clear distinction was drawn between the ‘user’ and the ‘developer’. Also, additional contextual information with illustrative examples was incorporated into the prompt to elucidate the expected manner in which justifications are presented. Justifications were required for all labels to ensure the labels were explainable by ChatGPT (Dwivedi et al. 2023).

The prompt supplied to ChatGPT API for automated labelling.



Please follow the instructions listed below:

- (1) Consider the following values and value items in Schwartz’s taxonomy of human values: ‘values & value items’
- (2) The API documentation is: ‘API document’
- (3) A mobile app user is: ‘the definition of app user’
- (4) According to the API documentation, identify and label up to 3 value items and their corresponding values that the API aligns with, enhancing, supporting, or considering them from an app user’s perspective.
- (5) The analysis must be focussed on how the API contributes to the experience of app users, not how the API works for developers.
- (6) In all cases, provide a brief justification for assigning the label.
- (7) Here are some examples: ‘examples’
- (8) Formatting instructions ...



‘Sample output structured as value; value item; justification^a’

- Hedonism; Enjoying Life; The API allows app users to access and utilise various media elements such as images, audio, and video, providing them with the opportunity to enjoy life and experience pleasure through media consumption.
- Self Direction; Curiosity; The API uses dynamic content suggestions and interactive challenges to ignite users’ curiosity, promoting creative exploration.

“Short forms of values and value items are used in the actual responses, as specified in the formatting instructions.

In addressing the challenge of token limitations discussed in Section 5.2, we opted to prompt ChatGPT with the short forms of values and value items to curtail the token count within ChatGPT’s responses, thereby affording more space for the core content, API documentation, and instructions. For example, ‘V1’ represented the value Self Direction, while ‘V1.3’ denoted Curiosity. The responses were formatted into a CSV structure to streamline the output and enhance interpretability. The language model was prompted to prioritise and select a maximum of 3 most relevant value items and their corresponding values (3 labels) for each API. This was to (a) enhance the accuracy (relevance) of the labels, (b) focus on labels with stronger justifications, and (c) enhance the quality of the findings by enabling assessment (Section 5.4) for a larger fraction of the identified labels. Nonetheless, ChatGPT API occasionally assigned more than 3 labels per API. A discussion with ChatGPT revealed the reason: ‘due to the complexity of the content and the availability of strong justifications, sometimes it might be challenging to prioritise the labels clearly’.

5.3.2. Mitigating the residual errors

The necessity for error handling emerges from the dynamic nature of the ChatGPT language model and the inherent complexities of technical interactions with ChatGPT API. A robust error-handling approach became imperative to ensure the validity of the content generated by the model and maintain the technical stability of the interactions with the ChatGPT API. Despite our refinement of the prompt used for automated labelling, the output generated by the model exhibited residual errors. That mainly concerned breaching the expected format of the response or providing uninformative (redundant) justifications for not assigning labels to the APIs: e.g. ‘The API is not relevant to the value item ...’ These errors, while understandable due to the model’s generative nature, could affect the integrity of the findings. Hence, we performed a post-cleaning step to remove such justifications and their corresponding labels. It is essential to mention that post-cleaning did not remove any labels based on the accuracy of the justification; all valid labels were preserved. We further implemented a retry strategy to provide a graceful way to handle errors that might occur when interacting with ChatGPT API, ensuring that the programme doesn’t crash and can

continue functioning when encountering temporary issues (e.g. a rate-limit error). When an error is encountered while querying the ChatGPT API, the programme waits for a specified time and retries the query. We used a decorator function to handle exceptions related to HTTP requests.

5.4. Soliciting expert opinion on ChatGPT’s revelations

To understand expert opinions about 9,000 value alignment labels assigned by ChatGPT, 10% of the labels were randomly selected (stratified) from different API themes and became subject to scrutiny by a panel of 6 experts. The experts (coauthors) held a postgraduate degree in software engineering or related fields with substantial practical experience in software development. They also went through the training process used in the pilot study (Section 4.1) to become familiar with Schwartz’s model of human values.

Figure 4 shows the flowchart used by the experts to assess the value alignment labels assigned by ChatGPT and provide their opinions. The assessment of each label started with studying the documentation of the API (as provided to ChatGPT in the prompt) associated with that label and the justification provided by ChatGPT for assigning the label. The expert flagged

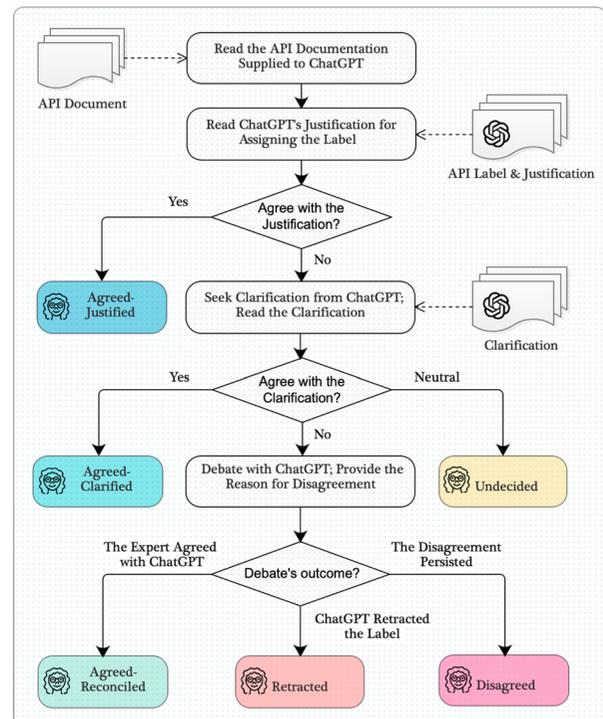


Figure 4. The process used by the experts to provide their opinions on the labels assigned by ChatGPT.

their opinion as ●Agreed-Justified when they agreed with the justification provided by ChatGPT; otherwise, they asked for clarification from ChatGPT. The following template was used to ask ChatGPT (gpt-3.5-turbo) to clarify its justification.

Expert seeking clarification on the justification provided by ChatGPT.

 You were given the following task: ‘automated labelling prompt’ In response, you labelled the API aligned with ‘value item’ and its corresponding value ‘value’ justified as: ‘justification text’. I am not convinced; can you please clarify?

The expert flagged their opinion as ●Agreed-Clarified when they agreed with the clarification provided by ChatGPT; the label was flagged ●Undecided when the expert could neither agree nor disagree. In the cases of disagreement, the expert provided their reason for disagreement, debating ChatGPT (Figure 4). The following prompt was used for the debate.

Expert debating ChatGPT.

 You were given the following task: ‘automated labelling prompt’ In response, you labelled the API aligned with ‘value item’ and its corresponding value ‘value’ justified as: ‘justification’ I was not convinced by your justification, so you provided the following clarification ‘clarification’. I still disagree because ‘reason for disagreement.’ Are you willing to change your opinion? if not, please provide your argument against my reason.

The debate led to three possibilities: (i) the expert agreed with ChatGPT and flagged their opinion as ●Agreed-Reconciled, (ii) Based on the reason provided by the expert, ChatGPT withdrew the label; was flagged ●Retracted, or (iii) the disagreement between the expert and ChatGPT persisted and the label was flagged ●Disagreed. Clarification responses were automatically generated for all value alignments, with experts consulting them only when clarification was needed (based on the process outlined in Figure 4). Debates were also automated by presenting the experts’ reasons for disagreement to ChatGPT and recording its responses for evaluation.

6. Findings from main study

This section presents the findings from the main study to answer (RQ1)–(RQ3) and their specific sub research questions.

6.1. RQ1: what patterns of human values does ChatGPT reveal in mobile APIs?

ChatGPT assigned around 9,000 labels to the 3,000 Android and iOS APIs where each label specified the alignment of an API with a value item (e.g. Privacy) and its corresponding value (e.g. Self Direction) from Schwartz’s model. Each label embodied a justification from ChatGPT explaining (Gunning et al. 2019) its rationale for assigning the label. Figure 5 illustrates the values and value items distribution across the 9,000 value alignment labels. The most prevalent value is Benevolence, which accounts for 21.77% of the labels. This is closely followed by Self Direction, characterised mainly by Independent and Freedom, and Universalism (Protecting the Environment). Achievement is found in 12.71% of the alignment labels, primarily in cases where ChatGPT assessed the API to help users be more successful and capable. Security represents 9.77% of the alignment labels and is mainly represented by Family Security and Health. Stimulation accounts for 7.17% of the labels, typically linked to APIs that enhance excitement in life. At 6.11% of the labels, power is represented by APIs aligned with achieving social recognition. Hedonism, found in 3.95% of the labels, is mainly found in APIs that help users enjoy life. Tradition and Conformity are the least prevalent values, accounting for only 2.03% and 1.74% of the alignment labels, respectively.

6.1.1. RQ1.1: how do mobile APIs align with socially and personally focused values?

Our analysis of the APIs’ alignment with human values (Figure 6) reveals the following insights under two categories Socially Focused and Personally Focused Values.

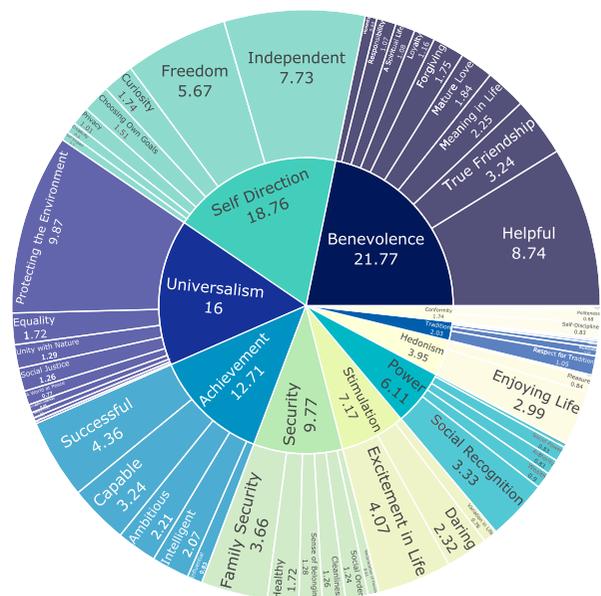


Figure 5. The distribution (%) of value alignments identified by ChatGPT.

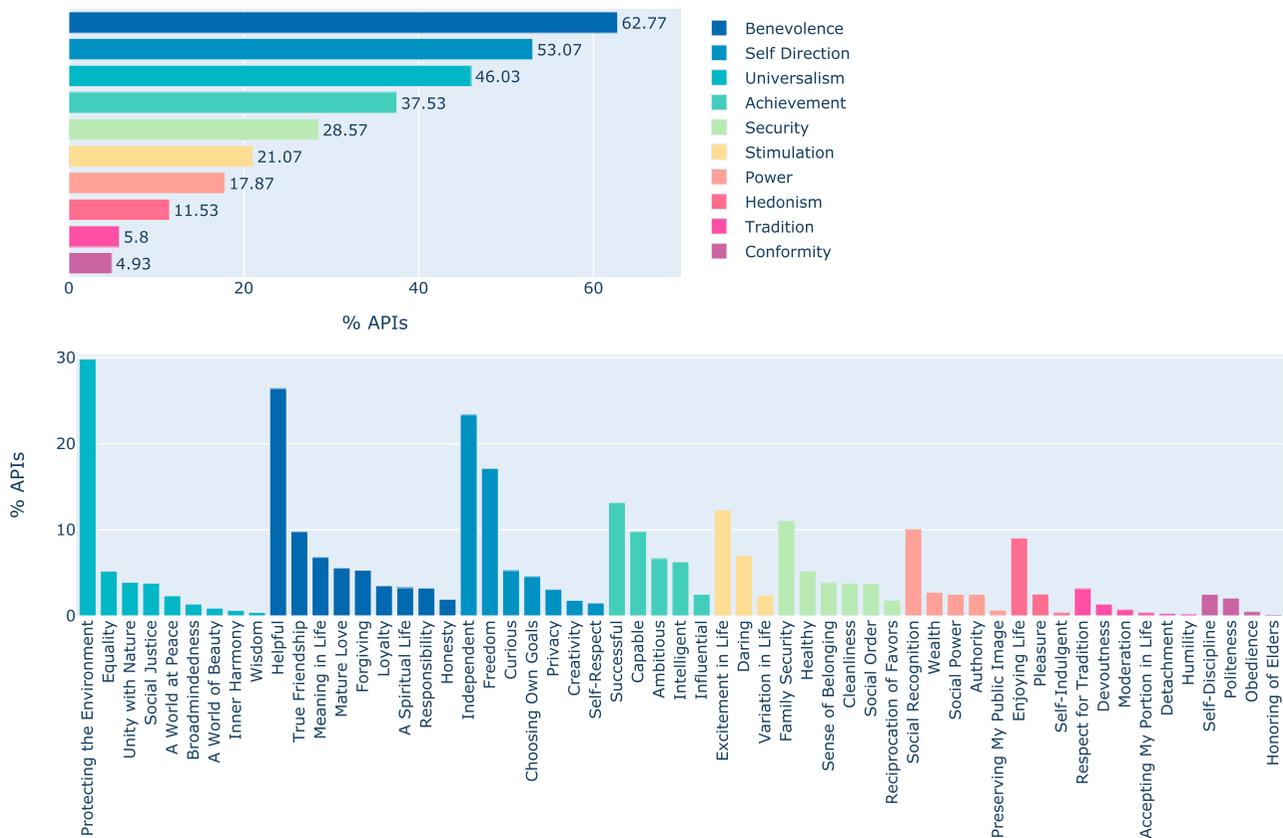


Figure 6. The alignment of mobile APIs with human values and their corresponding value items according to ChatGPT. An API may align with multiple values or value items.

Socially Focused Values. According to ChatGPT, 62.77% of the studied Android and iOS APIs contribute to enhancing, supporting, or considering Benevolence. It is evident (Figure 6) that Benevolence primarily manifests in the APIs contributing to being Helpful. Universalism is supported by 46.03% of the APIs, with most contributions directed toward Protecting the Environment. However, manifestations of Universalism such as Wisdom, Inner Harmony, and A World of Beauty are underrepresented in less than 1% of the APIs. Security is represented by 28.57% of the APIs, mainly associated with supporting Family Security, while National Security is not reflected in any API alignments. Alignment with Tradition is limited to 5.8% (174 APIs), contributing to Respect for Tradition and Devoutness. Other expressions of Tradition, such as Moderation, Accepting my Portion in Life, and Detachment, appear in less than 0.73% of the APIs. Conformity is represented in 4.93% (148 APIs), primarily through Self-Discipline and Politeness, with Obedience and Honoring of Elders present in less than 0.5%. These findings are consistent with the pilot study, where experts did not find any alignment with Tradition.

Personally Focused Values. Self Direction is supported by 53.07% of the APIs, enabling mobile app users to be more Independent and gain more Freedom. Achievement aligns with 37.53% of the APIs, mainly helping users to be Successful and Capable. Stimulation is represented in 21.07% of the APIs, predominantly supporting Excitement in Life. Power, supported by 17.87% of the APIs, is mostly manifested through Social Recognition, while Preserving my Public Image is represented in only 0.67%. Hedonism, at 11.53%, primarily aligns with Enjoying Life, with Self-Indulgent barely represented (0.37% in APIs).

6.1.2. RQ1.2: how do value alignments vary across different API themes?

In Section 5.2, we detailed the categorisation of the Android and iOS APIs examined in this study (3,000 APIs) across 10 different themes, each representing a specific type of functionality. The distribution of the APIs across these themes is demonstrated in Figure 3. We analysed the alignment labels assigned by ChatGPT across the API themes to identify patterns in how mobile APIs, differing in functionality, aligned with human values. Our examination of ChatGPT's findings,

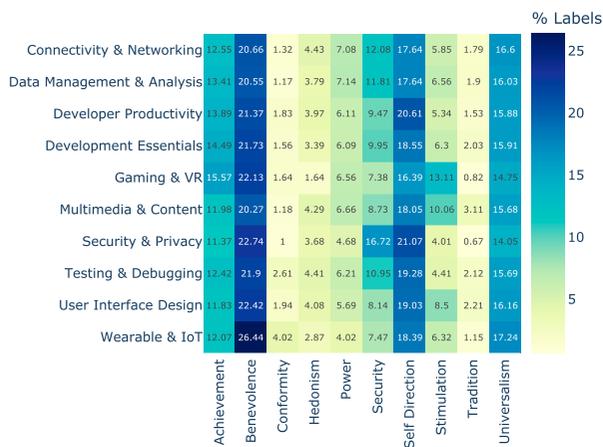


Figure 7. The distribution (%) of value alignment labels assigned by ChatGPT across different API themes, per value. Wearable & IoT APIs have primarily focussed on Benevolence: 26.44% of the value alignments revealed from Wearable & IoT APIs. Conformity and Tradition are rarely expressed, across most API themes.

as illustrated in Figures 7 and 8, shows that mobile APIs spanning all themes primarily align with Benevolence, Self Direction, Universalism, and Achievement, in that order of prominence. However, we note that alignment with Benevolence, particularly its key manifestations of Helpful and True Friendship, is predominantly observed in the context of Wearable & IoT APIs. On the other hand, alignment with Self Direction (emphasizing Freedom and Independence) is most prevalent in the API themes of Security & Privacy and Developer Productivity. Additionally, alignment with Universalism is most frequently found in the API themes of Wearable & IoT, Connectivity and Networking, and User Interface Design. Universalism, mainly represented by Protecting the Environment, also exhibits a notable presence in the Wearable & IoT API theme (Figure 8).

The highest degree of alignment with Achievement is observed for Gaming & VR APIs, particularly those that assist users in achieving greater success. Another aspect of Achievement centres around enabling app users to become more capable. This is mainly seen in the API themes of Development Essential and Developer Productivity, in that order of prevalence. Alignment with Security is noted across all API themes, predominantly in the APIs designed to enhance the Security & Privacy of mobile apps, as shown by Figure 7. We also observe that Power finds substantial support from the mobile APIs associated with Data Management & Analysis, and Connectivity & Networking. In contrast, alignment with Hedonism, particularly manifested as Enjoying Life, is predominantly observed in the API themes of Connectivity & Networking, Testing & Debugging, and Multimedia & Content.

Our analysis also shows that the less common attributes of Conformity and Tradition (Figure 5) are primarily linked to the API themes Wearable & IoT and Multimedia & Content, respectively. Conformity is predominantly characterised by its alignment with Self-Discipline, whereas Tradition is mainly manifested through instances of API alignments with Respect for Tradition. Upon closely examining the manifestations of human values (Figure 8), we observe that mobile APIs aligning with the value items Protecting the Environment, Helpful, Independent, Freedom, and Successful are the most prevalent across all API themes, on average. These values are followed by Family Security, Excitement in Life, True Friendship, Capable, and Social Recognition. In contrast, value items like Preserving My Public Image, Moderation, Inner Harmony, Self-Indulgent, Accepting My Portion in Life, Wisdom, Honoring of Elders, Detachment, and Humility are notably less represented within the API themes, on average.

The API theme of Protecting the Environment, representing Universalism, is predominantly present in APIs associated with Connectivity & Networking, Multimedia & Content, and Data Management & Analysis (Figure 8). Equality, as the second most prevalent manifestation of Universalism, is more commonly observed in the content of Wearable & IoT APIs. Alignment with the value item Helpful, representing Benevolence, is most frequently observed in the Developer Productivity, Testing & Debugging, Wearable & IoT, and Gaming & VR categories. Furthermore, the API themes Security & Privacy, as well as Wearable & IoT, exhibit the highest level of alignment with the value item Independent, representing Self Direction. In addition, the Freedom of mobile app users (indicating Self Direction) receives significant support from the APIs used for User Interface Design and Security & Privacy. On the other hand, Gaming & VR APIs display the highest degree of alignment with the value item Successful, which is the primary manifestation of Achievement.

The primary expression of security, Family Security, is mainly noted among the APIs categorised under the Security & Privacy theme. It is evident that alignment with other security expressions, like Healthy and Social Order, is also prevalent in the APIs. However, Cleanness and Reciprocation of Favor are notably less represented. Furthermore, the presence of Excitement in Life, an expression of Stimulation, is primarily observed in the mobile APIs associated with the Multimedia & Content, Gaming & VR, and User Interface Design themes, in descending order of prominence. Notably, the greatest degree of alignment with the value of True Friendship, representing Benevolence, is evident in the APIs related

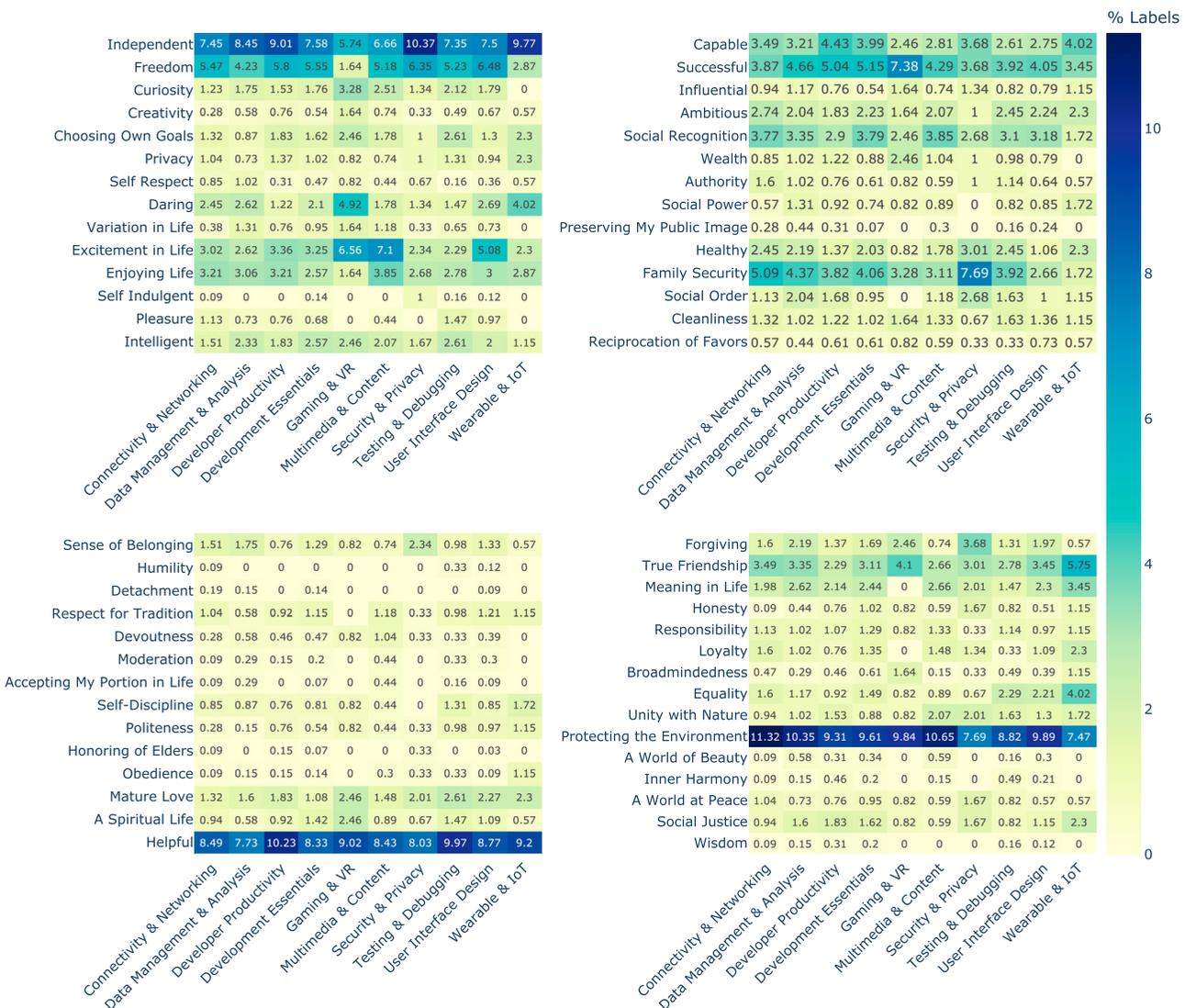


Figure 8. The distribution (%) of value alignment labels assigned by ChatGPT across different API themes, per value item. Expressions of Excitement in Life are most notable in Gaming & VR APIs (5.56% of the alignment labels related to this theme) and Multimedia & Content (7.1% of the alignment labels related to this theme).

to Wearable & IoT. In the context of alignment with the value item Capable (Achievement), it is most commonly encountered within the API themes of Developer Productivity, Wearable & IoT, Development Essentials, and Security & Privacy. Additionally, the manifestation of Social Recognition, representing Power, is primarily supported by the API themes of Multimedia & Content, followed by Development Essentials, Connectivity & Networking, and Data Management & Analysis.

6.1.3. RQ1.3: how do value alignments vary across Android/iOS APIs?

To understand the significance of considering human values in different mobile platforms, we examined the distribution of value alignment labels assigned by ChatGPT within standard Android and iOS APIs and third-party Android and iOS APIs, denoted by Android+ and iOS+,

respectively. Our analysis shows the distribution of value alignment across various mobile platforms, as depicted in Figures 9 and 10. It indicates that Benevolence, Self Direction, Universalism, and Achievement are the most commonly aligned values for standard and third-party Android and iOS APIs. However, the alignment of iOS APIs with these values exceeds that of Android APIs. Additionally, the alignment of third-party iOS APIs with Self Direction and Achievement exceeds that of standard iOS APIs. In contrast, Universalism is more prominent in standard iOS APIs. This identical pattern is also observed in the case of standard and third-party Android APIs.

Figure 9 also shows that the Alignment with Achievement, Benevolence, Hedonism, Self Direction, and Stimulation is more prevalent in third-party iOS and Android APIs compared to the standard APIs. On the

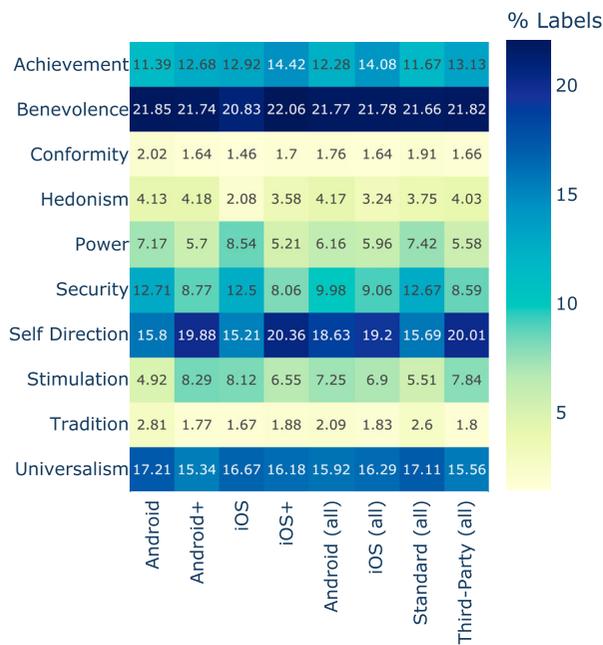


Figure 9. The distribution (%) of value alignment labels across Android and iOS platforms, per value. Third-party Android and iOS APIs are denoted by Android+ and iOS+, respectively. 8.29% of the value alignments revealed from third-party Android APIs, address Stimulation. The focus on Security in the standard APIs (all standard Android and iOS APIs combined) is higher than the third-party APIs.

contrary, alignment with Conformity, Power, Security, Tradition, and Universalism is more commonly found in the standard (iOS and Android) APIs. We also observed that, in comparison to Android APIs, iOS APIs demonstrate a higher level of alignment with Achievement, Benevolence, Self Direction, and Universalism, while Android APIs demonstrate a higher level of alignment with the remaining values, namely Security, Stimulation, Power, Hedonism, Conformity, and Tradition. This implies that, in general, Android APIs exhibit a higher level of alignment with the values that have an overall low representation in the alignment labels assigned by ChatGPT (Figure 5).

Our analysis of the results at the value item level shows (Figure 10) that alignment with the main representations of Self Direction (Freedom and Independent) is significantly higher in the third-party Android APIs compared to the standard Android APIs and all iOS APIs. Similarly, the % labels specifying the alignment of third-party Android APIs with the value items Capable and Successful (primary representations of Achievement) is higher than that of standard Android and all iOS APIs. The same pattern is evident for Benevolence (Helpful and True Friendship) and Universalism (Protecting the Environment and Equality). It is also evident that except for Privacy (manifesting Self Direction), Humility

(manifesting Tradition), and Obedience (manifesting Conformity), third-party Android APIs align more with the value items compared to standard Android APIs. Similarly, third-party iOS APIs align more with the value items compared to standard iOS APIs, except for the value items Humility (an expression of Tradition), Detachment (an expression of Tradition), and Honoring of Elders (an expression of Conformity). It is also worth noting that for all value items, the percentages representing the alignment of standard Android APIs with the value items are higher than those of standard iOS APIs. The same pattern can be observed for third-party Android and iOS APIs.

6.2. RQ2: how accurate are value alignments revealed by ChatGPT?

Considering expert opinion is important for evaluating the accuracy of the value alignment labels revealed by ChatGPT. They provide a comprehensive perspective that enhances the results and aids in recognising and addressing potential biases in the prompt used for revealing values. To gain insight into expert perspectives on the identified value alignments, we assess 10% of the value alignment labels assigned by ChatGPT, as detailed in Section 5.4. As depicted in Figure 4, experts mark their opinion as ● Agreed-Justified when they agree with the assigned label based on the justification provided by ChatGPT. In other cases, the experts seek clarification from ChatGPT and mark ● Agreed-Clarified if they agree with the label. When the expert is unsure, they use ● Undecided to indicate their neutrality. However, when disagreements arise, experts express their reasons for disagreement and engage in a debate with ChatGPT, as shown in Figure 4. Three possibilities are considered: (i) the expert concurs with ChatGPT's assigned label, they marked their opinion as ● Agreed-Reconciled, (ii) in cases where, based on the expert's provided rationale, ChatGPT withdraws the label, it is marked as ● Retracted, or (iii) when the disagreement between the expert and ChatGPT persist, the label is designated as ● Disagreed.

The distribution of expert opinions on the value alignments revealed by ChatGPT shows a relatively low ratio (1.78%) of residual disagreement, marked as ● Disagreed in Figure 11(a). In 75% of the cases, experts agree with ChatGPT's findings, either by providing their opinion as a form of agreement, mainly based on the initial justifications provided by ChatGPT (● Agreed-Justified), or through reading ChatGPT's clarifications (● Agreed-Clarified). In approximately 7.44% of the cases, experts have changed their opinion to agree with ChatGPT (● Agreed-Reconciled) based on the

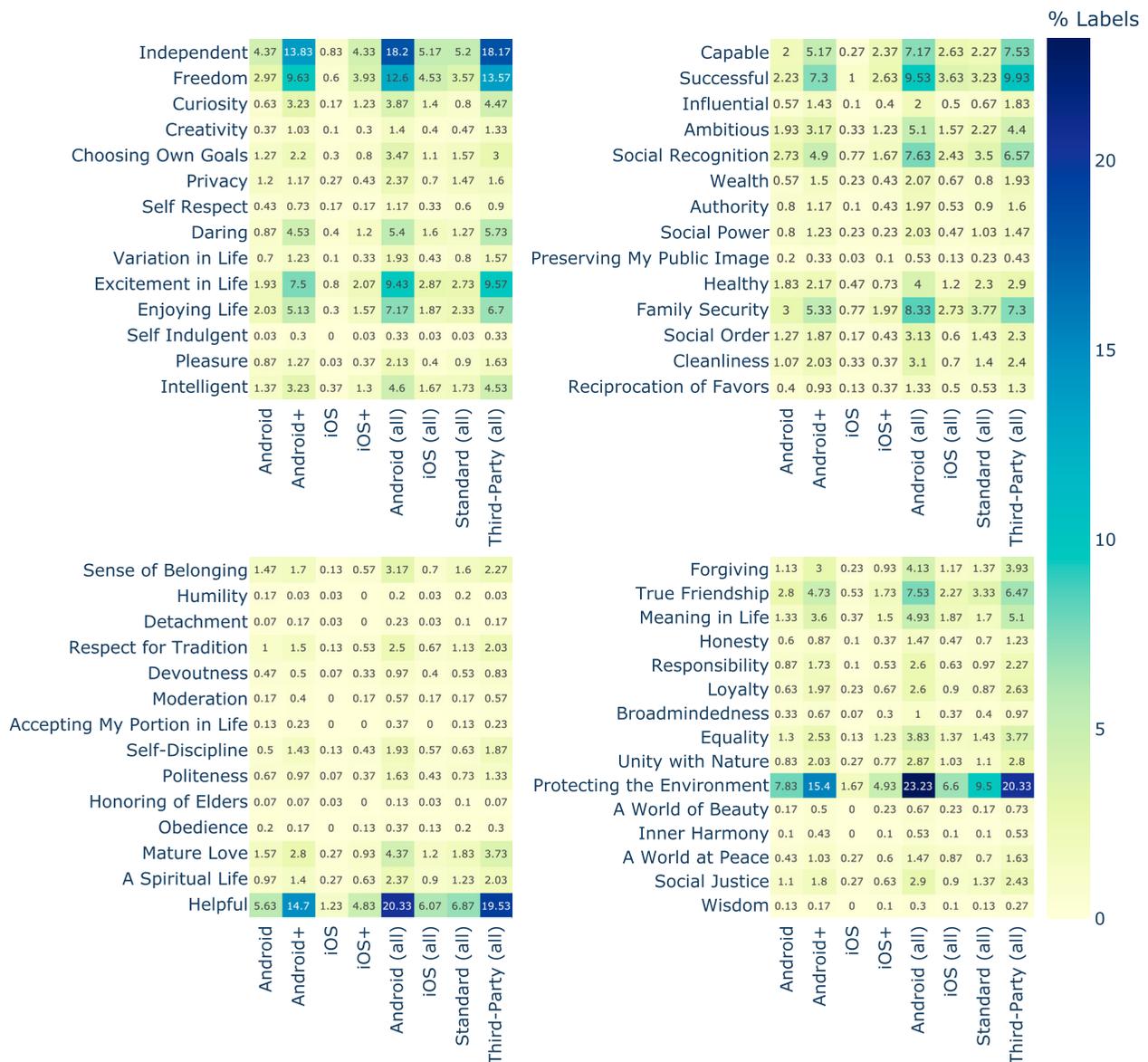


Figure 10. The distribution (%) of value alignment labels across Android and iOS platforms, per value item. Third-party Android and iOS APIs are denoted by Android+ and iOS+, respectively. The ratio of the value alignment labels concerning Excitement in Life in the third-party Android APIs (7.5%) is significantly higher than the standard Android APIs (1.93%).

rational provided by the language model in a debate. That constitutes around 10% of the agreements, suggesting:

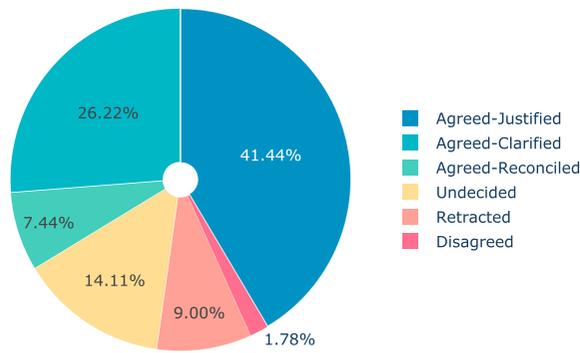
💡 ‘Experts engaging with ChatGPT can influence their understanding of human values and their manifestations’.

Figure 11(b) depicts the diversity of opinions among the experts labelled as Exp-A to Exp-F. It becomes evident that experts directly engaged in the labelling process during the pilot study (Exp-A to Exp-C) tend to incline towards agreeing with ChatGPT’s findings. In

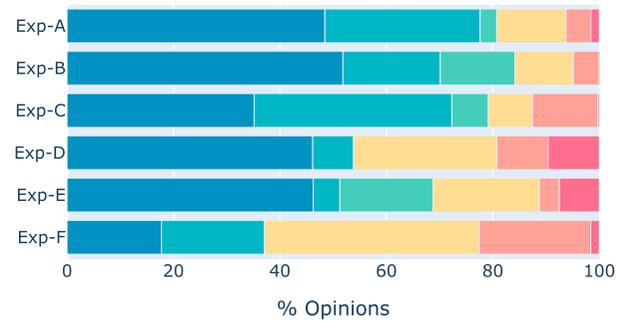
contrast, others lean towards a higher degree of neutrality or disagreement. This inclination may suggest either a greater impartiality in the viewpoints of Exp-D, Exp-E, and Exp-F or a deeper familiarity on the part of Exp-A, Exp-B, and Exp-C with the rationales presented by ChatGPT during the pilot study. It can be entailed:

💡 ‘Understanding ChatGPT’s rationale for revealing values and their manifestations can impact expert opinions’.

The notably higher level of agreement achieved through a debate with ChatGPT (marked ● Agreed-



(a) Overall distribution of opinions.



(b) Distribution of opinions per expert.

Figure 11. Distribution of expert opinions about value alignments revealed by ChatGPT. 9% of the value alignment labels assigned by ChatGPT were retracted by the language model following the debate with the experts. Disagreements persisted in 1.78% cases. (a) Overall distribution of opinions. (b) Distribution of opinions per expert.

Reconciled) for Exp-E lends support to the latter possibility, while the substantial number of labels retracted by ChatGPT (marked ● Retracted) backs the alternative hypothesis; further investigation can examine these possibilities.

6.2.1. RQ2.1: how do experts evaluate value alignments across different values?

Our preliminary analysis of expert opinions on ChatGPT's assigned alignment labels (Figure 12) reveals that the highest level of agreement (● Agreed-Justified, ● Agreed-Clarified, or ● Agreed-Reconciled) is linked with alignment to Hedonism and Stimulation, with Self Direction and Benevolence closely following. Conversely, the lowest level of agreement and the highest degree of neutrality (● Undecided) is evident in cases of alignment with Tradition and Universalism. Experts expressed the least indecisiveness on alignment labels associated with hedonism and Universalism, and the lowest level of indecisiveness is observed for the values Stimulation and Hedonism. Notably, the highest proportion of retracted alignment labels following a debate between experts and ChatGPT (● Retracted) is observed for alignment cases with Tradition, Conformity, and Achievement. In contrast, for the values of Hedonism, Stimulation, and Benevolence, ChatGPT rarely retracts any labels, indicating the confidence of the language model.

The distribution of the opinions marked as ● Agreed-Reconciled indicates (Figure 14) that the highest ratio of opinion change following the debate occurs for the alignment labels associated with Tradition, Self Direction, and Universalism. This implies a significant shift in experts' comprehension of these values and their manifestations due to the reasoning presented by ChatGPT during the debate. It is also evident that, for

Stimulation and Hedonism, a lower ratio of ● Agreed-Reconciled compared to other forms of agreement, specifically ● Agreed-Justified and ● Agreed-Clarified, suggests that the experts' understanding of those values and their manifestations has not been substantially affected by the discussions with ChatGPT. The alignment labels associated with Universalism and Hedonism exhibit the highest level of disagreement. At the same time, there are no instances of residual disagreement (● Disagreed) observed for the alignment labels linked to Power, Conformity, and Tradition.

Upon closely examining expert opinions on alignments with value items (manifestations of values), we observe (Figure 14) that the experts agree (● Agreed-Justified, ● Agreed-Clarified, or ● Agreed-Reconciled) for all instances of alignment with Inner Harmony, Social Justice, and A World at Peace, representing Universalism. In the case of Self-Respect (manifesting Self Direction), the experts mark their opinions as ● Agreed-Justified for all instances of alignments revealed by ChatGPT without requiring any clarification beyond the initial justifications. For alignment labels related to Accepting my Portion in Life (Tradition), however, agreement is only achieved after reading ChatGPT's clarifications on the initial justifications (● Agreed-Clarified). The lowest level of agreement currently exists for value alignment labels related to the value items Respect for Tradition (Tradition), Healthy (Security), A Spiritual Life (Benevolence), Reciprocation of Favors (Security), and Curious (Self Direction). Additionally, it is evident (Figure 14) that the highest level of disagreement is associated with the value items Unity with Nature (Universalism), Responsibility (Benevolence), Healthy (Security), and Curious (Self Direction). It is also worth noting that, for the majority of the value

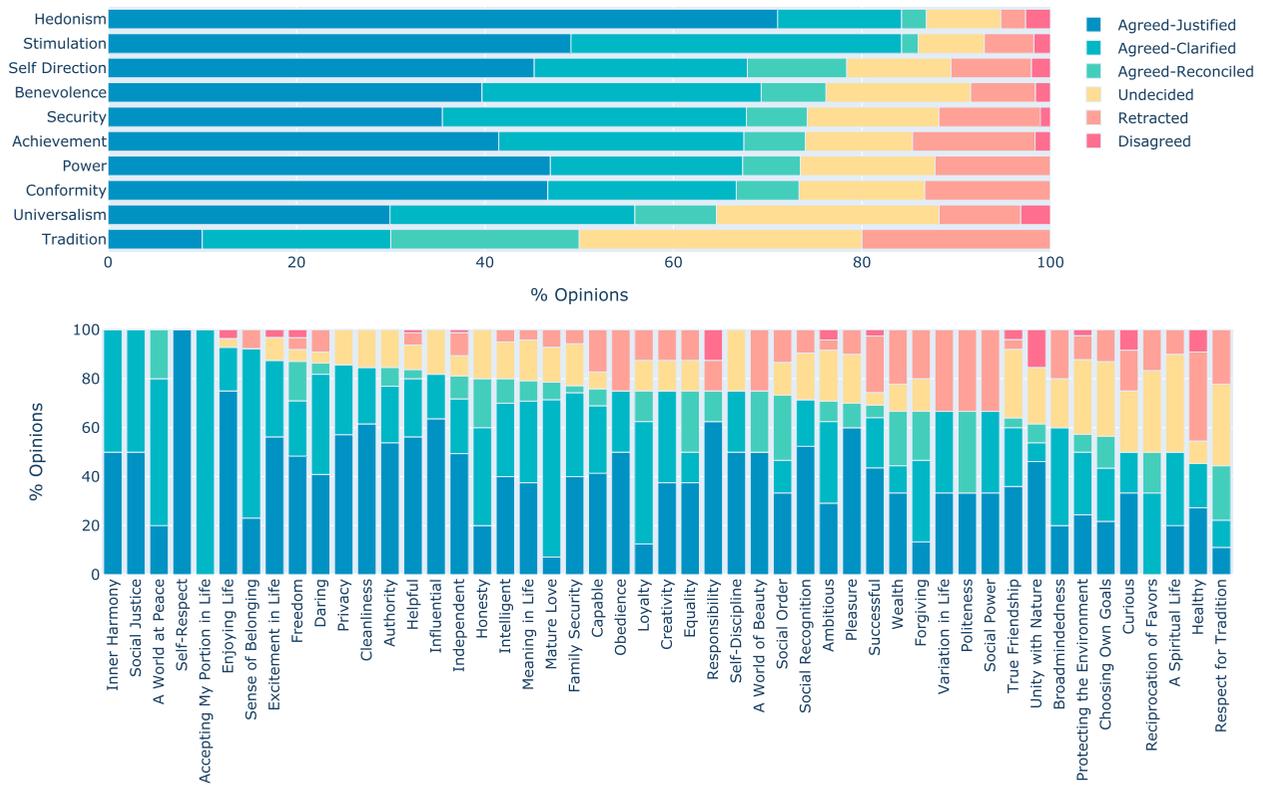


Figure 12. Expert opinions on the value alignments revealed by ChatGPT per value (value item). The highest ratio of experts agreeing with ChatGPT’s justification about value alignments without any need for clarification or debates is observed for value Hedonism. The lowest level of combined agreement is observed for Respect for Tradition.

items, no residual disagreement is found between the experts and ChatGPT. The highest level of neutrality in expert opinions is observed for the value alignment labels related to A Spiritual Life (Benevolence), Respect for Tradition (Tradition), Reciprocation of Favors (Security), Protecting the Environment (Universalism), Choosing Own Goals (Self Direction), and True Friendship (Benevolence).

The proportion of labels retracted by ChatGPT during debates with experts is highest for value items: Healthy (Security), Variation in Life (Stimulation), Social Power (Power), and Politeness (Conformity). Notably, for value items like Unity with Nature (Universalism), Social Justice (Universalism), and Privacy (Self Direction), ChatGPT confidently maintains its initial judgments throughout. In the case of Unity with Nature (Universalism), ChatGPT’s persistence, despite expert reasoning, has resulted in a relatively high level of disagreement, as demonstrated in Figure 14. Figure 14 illustrates that, for the majority of the value items, engaging in debates with ChatGPT leads experts to reconsider their opinions regarding API alignments to agreement (● Agreed-Reconciled). This is most pronounced for Politeness (Conformity), followed by Social Order (Security), a World of Beauty (Universalism), and Equality (Universalism). However, for certain value

items like Variation in Life (Stimulation), Social Power (Power), Obedience (Conformity), and Healthy (Security), expert debates do not result in opinion changes; instead, ChatGPT retracts the labels. In the case of Healthy and some other value items (Figure 14), a degree of residual disagreement persists.

 ‘Experts engaging with ChatGPT can influence the language model’s understanding of human values and their manifestations’.

6.2.2. RQ2.2: how do experts evaluate value alignments across API themes?

Figure 13 illustrates the distribution of expert opinions across different API themes. It shows that experts exhibit the highest level of agreement (● Agreed-Justified, ● Agreed-Clarified, ● Agreed-Reconciled) with the value alignment labels disclosed by ChatGPT in the Security & Privacy APIs documentation. Following closely are the APIs related to Gaming & VR and Connectivity & Networking. In contrast, the API themes Testing & Debugging and Development Essentials display the lowest level of expert agreement for the value alignment labels unveiled by ChatGPT. These two

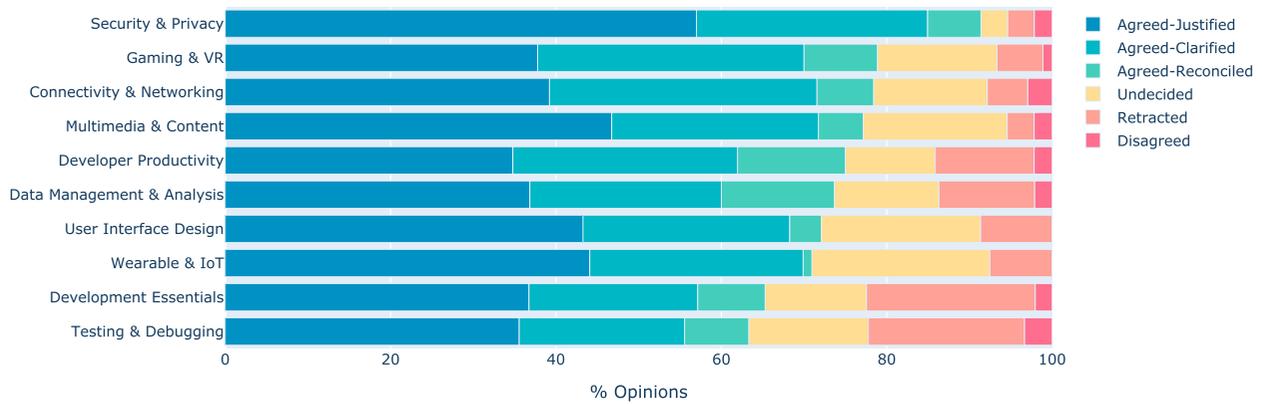


Figure 13. Expert opinions on value alignments, revealed by ChatGPT, across API themes. The highest level of combined agreement between experts and ChatGPT is observed for the value alignments identified in Security & Privacy APIs.

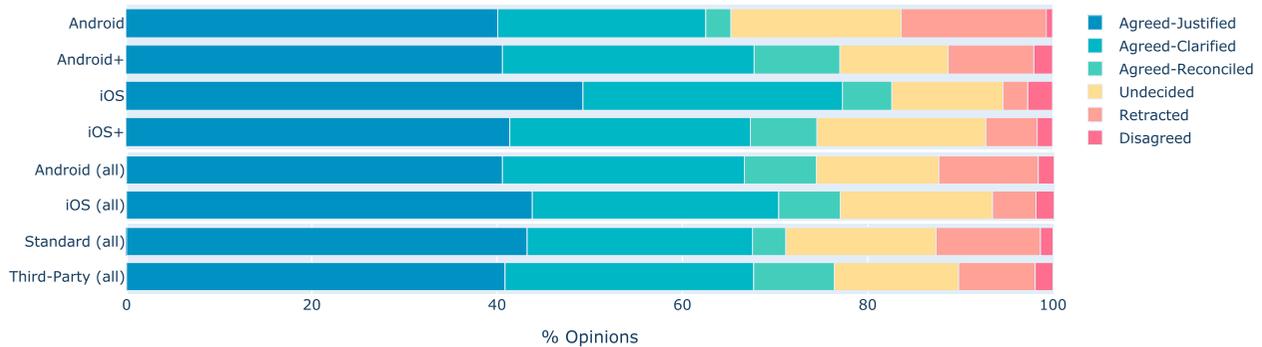


Figure 14. Expert opinions on value alignment labels, revealed by ChatGPT, across different mobile platforms. The + sign denotes third-party APIs. The lowest level of combined agreement is observed for value alignments detected in standard Android APIs.

themes also contain the highest ratio of labels withdrawn by ChatGPT following a debate with the experts (● Retracted), implying less accuracy in the initial judgment of the language model. Furthermore, a higher level of retracted labels is observed in more general-purpose themes, such as Development Essentials, Testing & Debugging, and Developer Productivity, while a lower level of retracted labels is evident in more specific themes like Security & Privacy, Multimedia & Content, and Connectivity & Networking. This pattern is depicted in Figure 13, and it suggests that:

💡 *‘ChatGPT’s accuracy in revealing value alignments is lower for the more general-purpose API themes and higher for the specific ones’.*

Expert agreement, achieved through deliberation with ChatGPT (● Agreed-Reconciled), is most pronounced for the Data Management & Analysis and Developer Productivity themes. This implies that experts have revised their initial judgments in light of the reasoning presented by ChatGPT during the debate.

The most significant disagreement occurs in the API themes of Testing & Debugging and Connectivity & Networking. Additionally, the greatest level of neutrality (● Undecided) in expert opinions is linked to the Wearable & IoT and User Interface Design APIs, respectively.

6.2.3. RQ2.3: how do experts evaluate value alignments across Android/iOS APIs?

Figure 6.2.3 shows the distribution of expert opinions on the value alignment labels across different mobile platforms. It can be viewed that the experts have expressed a higher level of agreement (● Agreed-Justified, ● Agreed-Clarified, and ● Agreed-Reconciled) with the value alignments revealed by ChatGPT in standard iOS APIs. On the contrary, the lowest level of expert agreements corresponds to the value alignment labels revealed in the standard Android API documentation. When standard and third-party APIs are considered together, a higher agreement is reached for the iOS APIs than the Android APIs. We can also see that the agreement level has been higher for the third-party APIs than the standard APIs. The highest ratio of the labels withdrawn by ChatGPT (● Retracted)

following a debate with experts is observed for standard Android APIs. That suggests:

💡 ‘The value alignments revealed by ChatGPT in the documentation of standard Android APIs are generally less accurate’.

Similarly, there is a higher level of retracted labels for the standard Android and iOS APIs than the third-party APIs. That suggests:

💡 ‘The accuracy of the value alignments revealed by ChatGPT in the standard APIs is lower than the third-party APIs’.

We can also see the ratio of ●Agreed-Reconciled is higher in the value alignment labels revealed for the third-party APIs compared with the standard APIs. That implies a higher accuracy in ChatGPT’s findings revealed from the documentation of the third-party APIs. The highest level of residual disagreements is observed for the standard iOS APIs.

6.2.4. RQ2.4: what are the primary causes of disagreement among experts?

Following the debates, we analysed the reasons for residual disagreements that persist (●Disagreed) between experts and ChatGPT. We examined the reasons provided by the experts and ChatGPT for insisting opinions. We identified that ChatGPT sometimes infers connections to values that are hard to understand by the experts as a clear expression of those values in API documentation is hard to see. In the discussion about an API’s ability to measure object size and optimise memory usage, ChatGPT inferred a connection to Excitement in Life (Stimulation). While not explicitly mentioned in the documentation, ChatGPT inferred that the optimisation process can add excitement to the user’s experience by allowing them to experiment with memory optimisation. The expert disagreed with this finding. ChatGPT’s ability to detect indirect connections has also led to a variation of perspective in some cases, where the expert emphasises that the API primarily serves developers. At the same time, ChatGPT considers its indirect impact on app users. In the discussion about an API’s ability to run Android UI tests multiple times, the expert focussed on an API’s impact on developers while ChatGPT examined how it indirectly benefits app users by improving the app’s quality and reliability through enhanced testing capabilities.

Hallucination. There were also cases where ChatGPT justifications sounded persuasive but could not be traced back to the API documents, suggesting possible AI hallucination (Athaluri et al. 2023). This was particularly more visible for poorly documented APIs.

💡 ‘ChatGPT provided made-up yet convincing justifications for imaginary value alignments in the absence of sufficient information’.

When reasoning about human value alignments in the API documentation, ChatGPT hallucinations manifest in several ways. Hallucinations often arise when the language model extrapolates or infers information that is not supported by the documentation content, leading to misleading claims. Below, we list the hallucination types with examples, which were observed following the human evaluation. It can be seen that the justifications provided by ChatGPT are deceptively convincing.

i. Inferred Connections Without Evidence. This type of hallucination occurs when ChatGPT draws connections between unrelated concepts or functionalities.

- **API:** Palette API^a
- **Value Alignment:** ChatGPT aligned the API with Protecting the Environment.
- **ChatGPT Justification:** ‘The Palette API allows app users to access and select colors from an image’s palette. This promotes the value of protecting the environment as users can create designs that align with eco-friendly principles’.
- **Hallucination:** ChatGPT claimed that selecting colors from an image’s palette promotes environmental protection. This connection is not supported by the API’s documentation; there is no logical link.

^a<https://developer.android.com/reference/androidx/palette/graphics/package-summary.html>

ii. Invented Features or Functionality. In some cases, the LLM invents features or functionalities that are not present in the API documentation.

- **API:** Google Android Things^a
- **Value Alignment:** ChatGPT aligned the API with ‘True Friendship’ under Benevolence.
- **ChatGPT Justification:** ‘The API documentation includes a feature called “True Friendship” that allows app users to connect with other users and share content; enhancing their sense of friendship and social interaction within the mobile app’.

- **Hallucination:** ChatGPT claimed the existence of a feature called True Friendship, which allows users to connect and share content. This feature does not exist in the API documentation, and the justification was entirely fabricated.

^a<https://developer.android.com/reference/com/google/android/things/package-summary.html>

iii. *Ambiguity Due to Insufficient Information.* When the API documentation lacks sufficient detail, the LLM may fill in the gaps with assumptions.

- **API:** JustWeEngine Game Framework^a
- **Value Alignment:** ChatGPT aligned the API with Family Security under *Security*.
- **ChatGPT Justification:** ‘The API supports the implementation of secure authentication mechanisms; ensuring the security of app users’ personal information and data. This feature contributes to the app users’ sense of family security and personal security; as their information is protected ...’
- **Hallucination:** ChatGPT aligned the API with Family Security. The ambiguity in the API’s documentation led the LLM to make an unsupported claim about family security, as there was no explicit evidence to support this alignment.

^a<https://github.com/lfkdsk/JustWeEngine>

iv. *Overgeneralization of Functionality.* The LLM may overgeneralise the functionality of an API, attributing values that are not directly supported by its features.

- **API:** LocaleChanger^a
- **Value Alignment:** ChatGPT aligned the API with Honesty under *Benevolence*.
- **ChatGPT Justification:** ‘The API is designed to persist the configuration of the Locale setting in an app. This feature supports the value of “Honesty” in Schwartz’s Taxonomy; as it ensures that the app accurately reflects the user’s chosen language and region settings; providing a truthful representation of their preferences.’
- **Hallucination:** ChatGPT claimed that persisting in the Locale setting promotes Honesty. This is an overgeneralisation, as the API’s functionality is about maintaining user preferences, not honesty or truthfulness.

^a<https://github.com/franmontiel/LocaleChanger>

v. *Misinterpretation of Technical Details.* In some cases, the LLM misinterprets technical details, leading to incorrect alignments.

- **API:** Wearable Input^a
- **Value Alignment:** ChatGPT aligned the API with ‘Social Power’ under *Power*.
- **ChatGPT Justification:** ‘The documentation mentions the deprecation of the WearableButtons class and recommends using the Jetpack Wear Input library instead. This indicates that the API supports the value item of Social Power by providing app users with the ability to interact with wearable devices and access their features; enhancing their social recognition and influence.’
- **Hallucination:** ChatGPT claimed that deprecating the WearableButtons class and recommending the Jetpack Wear Input library promotes Social Power. This misinterpretation of technical changes as a value-aligned feature demonstrates how the AI can misunderstand the purpose of technical updates, leading to illogical alignments.

^a<https://developer.android.com/reference/android/support/wearable/input/package-summary.html>

The balance between automated analysis and human intervention. Discrepancies in ChatGPT’s responses necessitate human intervention. However, manually evaluating all LLM outputs may not be scalable. Addressing this challenge requires a balanced approach that combines automated reasoning about values with human oversight to improve accuracy while maintaining scalability. Moreover, complicating the prompt used for unveiling values may lead to more confusion and reduce the accuracy of the responses. A viable trade-off involves incorporating self-evaluation prompts to enforce stricter criteria for assessing value alignments, ensuring only outputs that meet these criteria are escalated for human review. Still, the volume of findings may remain too large for exhaustive evaluation. In such cases, assessing a sample – similar to what we did in this study – representing different values offers a practical compromise between rigour and scalability. Fine-tuning the language model is another possible improvement but neither this can replace the need for human evaluation due to the stochastic nature of LLMs as well as the complexity of values and their subtle manifestations in software artifacts.

6.2.5. RQ2.5: how do experts rate the (endorsed) value alignments?

Rating the Endorsed Value Alignments and Computing Interrater Agreement. To rigorously evaluate the accuracy of the inferred value alignments, we conducted a supplementary assessment involving 6 experts who had previously participated in the dialogue-based assessment (Section 5.4). Each of the 676 value alignment labels, previously endorsed by the experts as Agreed-Justified, Agreed-Clarified, or Agreed-Reconciled, was independently rated by two experts, designated as Coder 1 and Coder 2. To maintain objectivity, the coders were assigned labels they had not reviewed during the earlier evaluation stage (the dialogue-based assessment). The coders independently assessed the justifications and clarifications provided by ChatGPT for each endorsed label and rated it as Confirm or Reject. Following the reconciliation of disagreements between the coders, 582 value alignments were confirmed by both, 63 were rejected by both, and 31 cases remained unresolved. This resulted in a Cohen's Kappa value of 0.78, indicating a substantial level of inter-rater agreement. It is important to note that the interrater reliability Kappa of 0.78 is computed based on the evaluation of ChatGPT's explanations (justifications and clarifications) for value alignments – not merely detecting the presence or absence of values, as performed in the pilot study. As such, it cannot be compared against the Kappa value calculated from the pilot study as they are computed under different settings.

Nevertheless, a fine-grained analysis of the coder ratings (Figure 15) revealed significant variability in interrater agreement (Kappa) across different values and value items. Coders achieved near-perfect agreement for the values Conformity and Hedonism, whereas the highest levels of disagreement were observed for Self-Direction (mainly the value item Independent) and Achievement (mainly the value item Successful). As shown in Figure 15, coders reached an agreement worse than random for the value items Successful and Mature Love, highlighting the pronounced subjectivity associated with these value items. Similarly, the value item Forgiving exhibited only a Fair level of agreement among coders.

6.3. RQ3: how does ChatGPT's prompt completion vary for different mobile APIs?

This section discusses the behaviour of ChatGPT's prompt completion. We examine three key parameters that describe prompt completion in language models, both in general and specifically in ChatGPT: prompt completion time (response time), prompt completion size (response size), and prompt size (content size). We analyze the significance of the findings related to

these variables using the Mann-Whitney U test (McKnight and Najab 2010). Additionally, we evaluate the strength and direction of the relationship between each pair using the Spearman rank correlation test. We employ the non-parametric tests, Mann-Whitney U test, and Spearman rank correlation test because the Shapiro-Wilk test (Hanusz, Tarasinska, and Zielinski 2016) does not suggest a normal distribution for response time; it rejects the null hypothesis with a p-value $\approx 0 < 0.05$. The Spearman rank correlation test allows for capturing non-linear relationships.

6.3.1. RQ3.1: how does ChatGPT's response time vary for different mobile APIs?

ChatGPT's prompt completion time, also known as response time, is the duration for a language model to generate a response after receiving a prompt. Response time varies based on multiple factors, including the complexity of the prompts, making it important to understand the behaviour of language models in different applications. The distribution of ChatGPT's response time (seconds) for different mobile API platforms and themes is illustrated in Figure 16(a,d,g). Notably, the response time varies predominantly between 4 and 7 s, with a central tendency around 5 s, which encompasses the analysis of the API documentation, the identification and justification of value alignments, and the formatting and return of the output.

We notice a difference in ChatGPT's response time for third-party and standard APIs. The Mann-Whitney U test shows a considerable value of U-statistic (1, 198, 544), indicating a significant difference between standard and third-party APIs regarding response time. Also, the p-value of $\approx 0.0 < 0.05$ indicates strong statistical significance, thus rejecting the null hypothesis. This, along with the distribution of the response time in Figure 16(a), entails:

 *'There is substantial evidence that ChatGPT has revealed values notably faster for third-party APIs compared to standard APIs.'*

Similarly, we test ChatGPT's response time difference for standard iOS and Android APIs. The high value of the U-statistic (37, 425) indicates a significant difference between standard and third-party APIs regarding response time. Additionally, the p-value of $\approx 0.0 < 0.05$ indicates strong statistical significance and leads to rejecting the null hypothesis. This, along with the distribution of the response time (Figure 16(d)), entails:

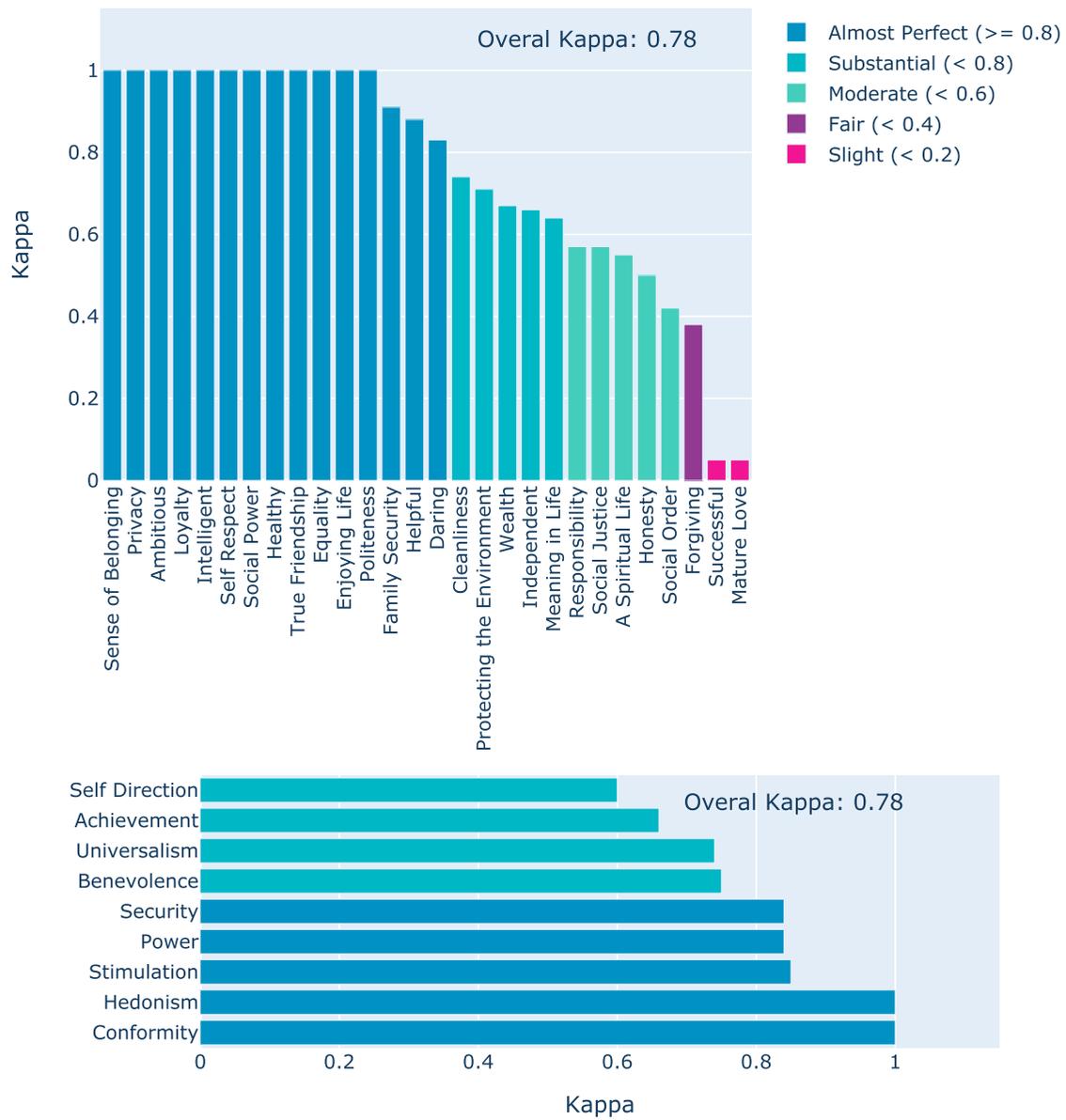


Figure 15. Interrater agreement (Kappa) calculated based on experts rating of the endorsed value alignments, across different values and their respective value items. Kappa could not be computed for Tradition due to insufficient data. Negative Kappa for the value items Successful and Mature Love, indicating agreement worse than random, were minimally adjusted for presentation purposes.

💡 *‘Significant evidence indicates that ChatGPT exhibits notably faster response times for standard iOS APIs in comparison to standard Android APIs’.*

6.3.2. RQ3.2: how does ChatGPT’s response size vary for different mobile APIs?

Prompt completion size (Response Size) refers to the number of tokens used by a language model like ChatGPT to generate a response to a given input prompt. Characterizing a response from ChatGPT in terms of prompt completion tokens is important as

it provides insight into the complexity of the generated response. Generally, a longer response indicates that the language model provides more detailed information or a more comprehensive answer, which may involve more complex reasoning or additional context. We use the term ‘complex’ consistently to account for both possibilities. Since ChatGPT’s responses mainly contain justifications for the value alignments, a larger response size is associated with a more complex justification. We observed variations in ChatGPT’s response size between third-party and standard APIs. The Mann-Whitney U test yielded a substantial U-statistic of 1, 099, 735, highlighting a significant difference in response size between standard and third-

party APIs. Furthermore, with a p-value of approximately 0.0 (< 0.05), strong statistical significance is indicated, leading to the rejection of the null hypothesis. This, combined with the distribution of response sizes (Figure 16(b)), underscores that third-party APIs tend to elicit larger response sizes compared to standard APIs.

💡 ‘Substantial evidence indicates that ChatGPT has provided notably more complex justifications for value alignment in standard APIs than third-party APIs’.

ChatGPT’s response size difference for standard iOS and Android APIs is also being tested. The high value of the U-statistic (44, 340) indicates a significant difference between standard iOS and Android APIs in terms of response size. However, the p-value (0.19) is greater than 0.05, indicating insufficient evidence (statistical significance) to reject the null hypothesis at the conventional threshold of 0.05. This is mainly due to the smaller sample size of the iOS APIs compared to Android APIs. Nonetheless, when looking at the distribution of the response time in Figure 16(d), we can still infer:

💡 ‘Although not substantially evident from the data, the large U-statistic value and the overall distribution of response sizes in Figure 16(d) hint at a greater level of complexity for standard Android APIs in comparison to standard iOS APIs’.

6.3.3. RQ3.3: how does ChatGPT’s prompt size vary for different mobile APIs?

ChatGPT’s prompt size refers to the number of tokens (units of text, e.g. character or word) supplied to the language model for a query. Prompt size is an important factor in determining the complexity of prompt completion (Rao et al. 2023). Since the only variable part of the prompt used for revealing value alignments is the API documentation (Section 5.3.1), a larger prompt means more content in its corresponding API documentation. The Mann-Whitney U test shows a high value of U-statistic (36, 4198), indicating a significant difference between standard and third-party APIs in terms of prompt size (content). Also, the p-value $\approx 0.0 < 0.05$ indicates strong statistical significance, hence rejecting the null hypothesis. This and the distribution of the prompt size (Figure 16(c)) entail:

💡 ‘Substantial evidence indicates that third-party API documentation contains more content than standard API documentation’.

Similarly, we tested the difference in ChatGPT’s prompt size (content size) for standard iOS and Android APIs. The high value of the U-statistic (76, 701) indicates a significant difference between standard and third-party APIs in terms of prompt size. Moreover, the p-value is approximately ≈ 0.0 , which is less than 0.05, indicating strong statistical significance and leading to the rejection of the null hypothesis. These findings, along with the distribution of the prompt size (Figure 16(f)) lead to the following conclusion.

💡 ‘Substantial evidence indicates that the API documentation for standard iOS APIs contains more content than that of standard Android APIs’.

6.3.4. RQ3.4: what is the interplay between prompt characteristics and ChatGPT’s responses?

This section analyses the relationship between the key characteristics of ChatGPT’s prompt completion: response size, response time, and prompt size. We employ the Spearman rank correlation test as it doesn’t assume the normality of the studied parameters, nor does it assume that the relationships are linear. Figure 16 reveals that the distribution of ChatGPT’s response size (prompt completion size in tokens) closely resembles that of ChatGPT’s response time, suggesting a potential link between response time and size. The Spearman rank correlation coefficient is 0.93 with a p-value of less than 0.05, indicating a strong positive relationship between ChatGPT’s response size and time. But we know (Section 6.3.2) that ChatGPT’s response size depends on the complexity of the justifications provided for value alignments.

💡 ‘Substantial evidence indicates that ChatGPT requires notably more time to unveil value alignments with more complex justifications’.

Furthermore, our analysis in Section 6.3.3 reveals that (a) the API documentation of third-party APIs contains more content than the API documentation of standard APIs, and (b) the API documentation of standard Android APIs contains less content than the API documentation of standard iOS APIs. We further concluded in Section 6.3.2 that (a) ChatGPT notably requires more time to reveal value alignments in standard APIs

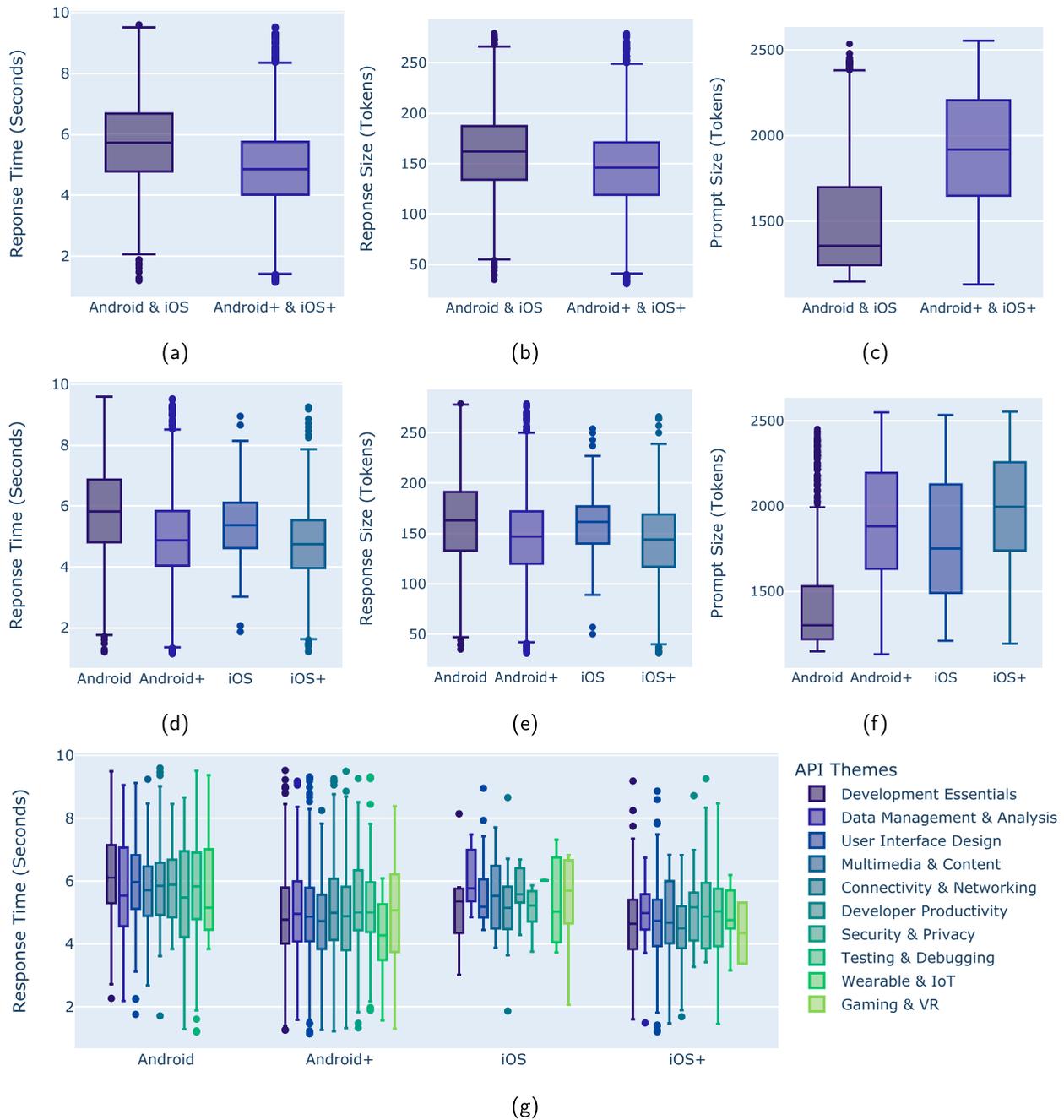


Figure 16. ChatGPT's prompt completion for standard and third-party (+) mobile APIs.

compared to third-party APIs, and (b) ChatGPT needs more time to reveal value alignments for standard Android APIs compared to standard iOS APIs. This raises the question: Is there a meaningful relationship between the amount of content in the API documentation and the time needed for revealing value alignments? We apply the Spearman rank correlation test to examine the relationship between prompt size (documentation content size) and response time. The correlation coefficient of -0.09 suggests a weak negative monotonic relationship between the two variables, and

the p-value of approximately $0 < 0.05$ indicates that this relationship is statistically significant. However, the strength of the relationship is not strong.

💡 *'Substantial evidence suggests that ChatGPT requires more time to reveal value alignments when the API documentation contains less content. Nevertheless, the weak correlation implies that other factors might also impact this association.'*

As discussed earlier, ChatGPT responses mainly consist of justifications for value alignments revealed in the APIs. We also clarify that response size is the primary complexity indicator in ChatGPT responses. Considering the strong positive correlation between ChatGPT's response size and response time, we can pose the following question: Does a meaningful relationship exist between the amount of content in API documentation and the complexity of the justifications? We employ the Spearman rank correlation test to investigate the link between prompt size (documentation content size) and response size (justification complexity) to address this. The correlation coefficient of -0.07 indicates a weak negative monotonic relationship between the two variables, and the p-value of approximately $0.0001 < 0.05$ signifies that this relationship holds statistical significance.

 *'Substantial evidence indicates that ChatGPT's justifications for value alignments are more complex in API documentation with less content; however, the weak correlation suggests other factors might be influencing this relationship'.*

As discussed in Section 5.2, we have summarised API documents to meet ChatGPT's token restriction, which raises the question. To statistically eliminate the possibility that summarisation has affected the complexity of the justifications provided for value alignments, we utilise Spearman correlation analysis. The p-value of $0.09 > 0.05$ and the correlation coefficient of -0.03 indicate no statistically significant relationship between response size (complexity) and the text summarisation ratio.

 *'There is no substantial evidence to suggest a connection between text summarisation and the complexity of ChatGPT's justifications'.*

7. Discussion

Trends in Value Alignments. Our findings in Section 6.1 (Answering RQ1) demonstrate significant alignment with Benevolence, Self Direction, and Universalism in mobile APIs. On the other hand, Tradition and Conformity exhibit low alignment. This raises questions about industry priorities and the factors influencing these alignments, necessitating investigation into whether they mirror user preferences, industry trends, or biases within API documentation. Such insights are crucial for developers aiming to create user-centric APIs.

Moreover, exploring patterns across API themes reveals varied alignments with human values, such as Benevolence in Wearable & IoT, Self Direction in Security & Privacy and Developer Productivity, Universalism in Connectivity & Networking and User Interface Design, Achievement in Gaming & VR, and distributed Security. Understanding these associations informs developers about value considerations in different functionalities, guiding the development of guidelines or frameworks for aligning values with specific applications.

Shift in Comprehension of Values. Expert opinion changing following debates with ChatGPT suggests a shift in experts' comprehension, particularly for the manifestations of Tradition, Self Direction, and Universalism. This raises the question of whether ChatGPT can indeed enhance practitioners' understanding of values, fostering mutual understanding with the machine. On the other hand, the substantial number of labels retracted by ChatGPT raises the question of whether the language model's insights could be refined through continuous engagement with the experts. In this regard, exploring the risks of introducing bias in the expert or machine's comprehension of values is crucial. These highlight the need for further investigations into the cognitive aspects of expert-machine interactions.

Missing National Security? The absence of APIs specifically aligned with National Security (Section 6.1) stems from the sensitive nature of national security, which leads to the restriction of API information, preventing public disclosure due to security risks. National security measures require customised solutions for government agencies, diverging from generic APIs. Legal and regulatory factors, along with a preference for secure communication channels not exposed through public APIs, contribute to limited accessibility.

API Themes Specificity and Accuracy of Value Alignments. The finding that ChatGPT's accuracy in revealing value alignments is higher for more specific themes raises questions about why this discrepancy exists. A potential hypothesis is that APIs with more specific purposes are easier to link with human motivational goals and, thereby, values. This prompts a discussion about the clarity in the expert and language model's understanding of values within different themes.

Discrepancies Between Standard and Third-Party APIs. The higher ratio of agreed-reconciled for third-party APIs compared to standard APIs implies higher accuracy in ChatGPT's findings from third-party documentation. Further investigation is needed to understand why discrepancies exist. Exploring potential biases in the model's training data or variations in

documentation styles between standard and third-party APIs could provide valuable insights. Involving software practitioners with expertise in API documentation and social scientists to analyze potential biases in machine-generated insights could contribute to a more comprehensive understanding of these discrepancies.

Complexity and Impact of API Documentation Content. The examination of prompt completion size highlights the complexity of justifications, with larger responses for standard APIs. The content and length of API documentation are suggested to influence ChatGPT's behaviour. Moreover, analyzing relationships among response size, response time, and prompt size reveals that more complex justifications lead to longer response times. Future research should explore the linguistic intricacies leading to complex responses.

Potentials of Using ChatGPT. ChatGPT supports transparency by offering explanations for its findings about values. It also shows potential in identifying hidden relationships between API descriptions and values, sometimes offering insights that human experts may not have considered. ChatGPT can also detect indirect links and reveal manifestations of values that are challenging for humans to detect. Additionally, the language model's ability to provide diverse and alternative answers fosters inclusiveness in understanding the manifestations of values in software artifacts. It also stimulates further discussion, essential for understanding human values' multifaceted nature.

Challenges of Using ChatGPT. ChatGPT's tendency to misinterpret words or phrases in API documentation with limited content sometimes leads to findings that are hard to understand or map into the documentation. Evaluating the suggested value alignment and reasoning provided by ChatGPT can also pose challenges, as it may make assumptions in API documentation that do not align with values, resulting in unexpected findings. The lack of a deep understanding of context and the potential for oversimplification or contradictions add to the difficulties. The observation that ChatGPT retracts labels most frequently for certain value items, such as Healthy, Variation in Life, Social Power, and Politeness, warrants further investigation. It is essential to understand whether these retractions indicate challenges in the model's interpretation or if they reflect nuanced aspects of values that may be difficult to capture accurately. Investigating this may uncover potential areas for improvement.

Expert Recommendations. Some experts recommend using ChatGPT as an initial screening tool to reveal a comprehensive range of values, with expert opinions used for the final judgment. Others suggest using ChatGPT as the primary tool but emphasise the

necessity of expert opinions to address situations where ChatGPT's decisions are based on incomplete information or incorrect identifications. The consensus is that ChatGPT can be valuable as a screening tool to discover human values in software artifacts, making the process more scalable and transparent. However, expert review remains crucial to validate the findings.

Miscellaneous Insights. Third-party mobile APIs are generally found to be more informative than standard APIs. In contrast, Android APIs, in particular, pose challenges due to poorly documented content compared to iOS and third-party APIs. Moreover, ChatGPT's explanations for certain values, such as Protecting the Environment, Reciprocation of Favours, and Unity with Nature, are observed to be difficult to understand, with concerns about the model making assumptions based on the API documentation. An interesting note is the last step of reconciliation between ChatGPT and experts (5.4), highlighting its potential usefulness and the interactive nature of the assessment process.

Generalizability of Findings. This study focussed on value alignments in mobile app reviews. Using Schwartz's theory of basic values, the identified value alignments are grounded in a well-established theoretical model. However, it is important to be careful about generalising any insights reported in this study to other types of software. Similar patterns may or may not be seen depending on how different types of software prioritise values. For instance, mobile APIs prioritising values such as Self-Direction and Benevolence might be due to their interactive nature. This prioritisation aligns with prior studies on mobile apps emphasising usability (Kortum and Sorber 2015). By contrast, safety-critical software may emphasise Achievement and Conformity, reflecting goals related to efficiency, reliability, and regulatory compliance. These differences highlight the need for further empirical studies to investigate value alignments across different domains. Even within the same domain, software products may prioritise values differently depending on their purpose (functionality). For example, a gaming mobile app may emphasise Stimulation and Hedonism, while a banking mobile app may prioritise Security and Conformity. Our thematic analysis of value alignments across different API themes (functionalities), as summarised in Figure 6.1.2), such as Gaming & VR, provides a structured framework that supports systematic comparisons. This framework can be adapted to analyze software systems, including cloud services, machine learning platforms, and IoT with different functionalities. The iterative refinement process and expert validation methodology employed in this study also provide a scalable approach for identifying value alignments across

software artifacts. Our findings highlight the necessity of automated approaches, such as the one proposed in this paper, for identifying numerous manifestations of values across diverse software artifacts in a scalable manner. The explanations provided by ChatGPT for a large number of value alignments in mobile API documentation can also be used to fine-tune the language model and develop more practical definitions of values. These refined definitions could enable more accurate value inference across other domains of software. Future studies should extend this work to various types of software artifacts, such as GitHub issues, user reviews, and code comments, and across different software domains (e.g. web, mobile, IoT, and safety-critical systems). Expanding this research can help build a broader knowledge base for understanding how values are expressed in different software artifacts and application domains. Achieving this goal will require interdisciplinary studies involving researchers and practitioners to further refine and validate automated methods for analyzing value alignments.

8. Limitations and threats to validity

8.1. Limitations

Prompt Framing Challenges. ChatGPT-3.5 demonstrated sensitivity to prompt framing, where small variations in phrasing influenced the outputs, sometimes leading to inconsistencies. This limitation posed challenges for ensuring reliability and consistency in identifying value alignments. In particular, minor changes in prompt wording led to variations in responses, occasionally resulting in overextended or overly constrained interpretations of values. To mitigate these issues, we employed an iterative refinement process referred to as Prompt Crafting. This process included embedding definitions of Schwartz's values directly into prompts, providing illustrative examples of value alignments, and specifying instructions to balance explicit and implicit interpretations. Prompts were further optimised to prioritise the most relevant and well-explained value alignments. Despite these refinements, occasional inconsistencies persisted. For example, ChatGPT sometimes confused perspectives between API developers and end-users, even after explicitly defining these roles in the prompts. Also, in many cases ChatGPT responses still contained inaccurate justifications for value alignments based on assumptions not backed by the API documentation. Future research may seek to develop hybrid approaches that firstly enhance the accuracy of the initial findings and secondly

embed semi-automated evaluation mechanisms that rely on collaboration between machine and human to detect and remove inaccurate findings about human values. Our preliminary findings highlight the importance of human involvement.

Token Limitations. This study relied on ChatGPT-3.5-turbo, which imposes token limitations that constrained the amount of information processed in each prompt. To address this, we employed extractive summarisation techniques to condense lengthy API documentation into manageable inputs. While this approach preserved, on average, 96% of the original content of each API documentation, it may have excluded some details critical to interpreting value alignments for some APIs. Future research may explore using more advanced models, such as GPT-4, with larger token capacities or adopting chunking strategies to better handle lengthy inputs without sacrificing context.

Reliance on ChatGPT-3.5. This study relied solely on ChatGPT-3.5, which may limit reproducibility due to ongoing updates to the model. The same prompts may yield different outputs as the model evolves, reducing consistency. To address this limitation, prompts were finalised and applied uniformly across all experiments, and outputs were verified through human inspection. Nevertheless, future studies should consider integrating hybrid approaches that combine outputs from multiple models, such as GPT-4 and domain-specific classifiers, to improve robustness and interpretability.

Bias and Transparency Concerns in LLMs. Large language models (LLMs) provide chances to uncover human values, but it is crucial to recognise and mitigate potential risks linked to their utilisation. One risk is the possibility of introducing biases into the software despite efforts to promote inclusivity and fairness. LLMs, trained on diverse datasets, may inadvertently perpetuate societal biases present in the training data (Lindahl and Saeid 2023; Pellert et al. 2022), leading to biased interpretations and representations of human values in software. Manual identification of biases in LLM-generated content may be challenging. Although research, e.g. the work by Fischer, Luczak-Roesch, and Karl (2023), has shown promise in LLMs' ability to reveal human values from text with minimal bias, the issue remains to be systematically scrutinised and addressed. Additionally, lack of adequate transparency in LLMs (Liao and Vaughan 2023) can lead to potential biases in understanding and integrating human values into the software. These challenges emphasise the need for a human-centred approach to enhance transparency in LLMs (Liao and Vaughan 2023).

8.2. Threats to validity

Construct Validity. We used extractive summarisation to condense lengthy API documents beyond the token limit of the ChatGPT API. We opted for this approach for its faithfulness to the original API documentation and its simplicity. However, extractive summarisation might yield less fluent summaries, affecting clarity for experts reviewing the labels (El-Kassas et al. 2021). Moreover, despite prompt refinement and error handling to mitigate inaccurate justifications provided by ChatGPT, they occasionally appeared in the labels, necessitating human involvement in assessing and interpreting ChatGPT’s findings. Nonetheless, manually inspecting a large number of value alignment labels may also be cumbersome; a hybrid strategy might involve expert evaluation of labels suspected to have inaccurate justifications.

Internal Validity. It is important to acknowledge that experts’ agreements with justifications generated by ChatGPT in the pilot study and main study might have been influenced by the novelty of the approach and the detailed arguments proffered by the language model. To mitigate this influence, it is recommended that the identified findings be subjected to scrutiny by software practitioners. This cross-validation process would lend additional credibility to the findings. Also, the sensitivity of ChatGPT’s responses to the framing of prompts further underscores a limitation. While the intention was to maintain consistency and accuracy in the responses, the possibility of variations in the labels found by ChatGPT API based on prompt framing must be recognised.

External Validity. It is important to note that the responses elicited from ChatGPT may exhibit variance upon repetition of the same questions. This variability could stem from the inherent probabilistic nature of language generation within the model. We used the default value of 1 for both parameter options *temperature* and *top_p* but changing that may lead to more predictable responses. While the measures undertaken effectively fortify the methodology within the purview of this specific study, their generalizability to alternative contexts or research inquiries might be circumscribed. The efficacy of the error-handling mechanisms and the selected summarisation technique should be evaluated within distinct research frameworks to gauge their broader applicability.

9. Recommendations for using LLMs to reason about values

Grounded in the methodology and findings discussed in this study, this section provides recommendations for

using LLMs like ChatGPT for reasoning about human values in software artifacts. These recommendations focus on prompt design, validation, and mitigating biases and errors.

Prompt Design. The iterative refinement process described in the methodology emphasises the trial-and-error nature of prompt crafting. While there is no guarantee that a prompt provides high accuracy at all times, the following suggestions may help. Our experiments show that explicitly providing the definitions of values and sample value alignments reduces ambiguity and enhances consistency in ChatGPT’s inference about values. To avoid confusion, the prompt must specify whether the focus of the analysis is on user values, developer values, or both; each must be explicitly defined. To maintain accuracy while reducing false negatives, the prompts should encourage prioritisation of direct connections (without overemphasis) while allowing well-justified indirect links. Moreover, instructions should ask ChatGPT to justify each value alignment with traceable explanations, mitigating unsupported claims and enhancing interpretability in automated labelling. Multiple cycles of testing and revising prompts, as demonstrated in the pilot study, addressed weaknesses such as vague outputs, errors in traceability, and inconsistencies. ChatGPT’s ability to detect indirect links and uncover latent value alignments highlights its potential for identifying subtle connections that might otherwise be overlooked during prompt refinement.

Human Involvement. The main study outlines a structured validation process involving expert assessment of ChatGPT outputs, highlighting the importance of human oversight. This shows the necessity of combining the automated reasoning abilities of LLMs with human expertise to ensure the accuracy of the findings about human values. We recommend the use of a dialogue-based assessment that allows multiple experts to engage with the language model and provide reasons rather than simply disagreeing with the findings. This helps reduce the gap between human and LLM understanding of values and provides a basis for future investigations. The ability of ChatGPT to stimulate discussions through diverse and alternative answers played a key role in resolving disagreements, emphasising its potential to uncover value expressions that are challenging for humans to detect. Yet, residual disagreements revealed challenges stemming from speculative explanations and misinterpretations, highlighting the need for continued human involvement.

Minimizing Biases and Errors. Error-handling strategies from the main study emphasise reducing biases and improving reliability through approaches such as self-evaluating prompts for detecting potential biases

inherited from training data, performing manual reviews and fairness checks, or a hybrid approach. Bias-detection prompts can explicitly evaluate fairness and consistency in outputs, allowing biases to be flagged for review. Automated filters flagged outputs with vague explanations or formatting errors, and iterative refinement cycles minimised unsupported responses. Findings stress the importance of periodically testing and updating LLMs to adapt to improvements and new patterns in value alignment.

10. Conclusions and future work

Large language models (LLMs), like ChatGPT, hold the potential to make inferences on human values in textual content. They utilise sophisticated language understanding and contextual reasoning to provide a diverse understanding of values and their expressions. The LLMs' ability to explain the reasoning behind their findings also improves reliability and facilitates discussions about values. In this study we leveraged the reasoning capabilities of ChatGPT, a large language model, to explore the potential of LLMs for revealing human values in software artifacts. We specifically assessed the alignment of standard and third-party Android and iOS APIs with values through a methodology combining ChatGPT with expert evaluation. Our study unveils a reciprocal evolution in understanding values, with both ChatGPT and experts adjusting their assessments through discussions. While experts broadly endorse ChatGPT's findings, indicating its potential for revealing values, a notable proportion of expert neutrality underscores the ongoing need for discussions to establish a common understanding of values between machines and humans.

Summary of the Findings. The summary of the main findings related to each research question is listed below.

Patterns of Human Values (RQ1). ChatGPT revealed around 9,000 value alignments from the documentation of 3,000 APIs. The most frequently identified values are Benevolence, Self-Direction, Universalism, and Achievement, which are expressed mainly through helpfulness, independence, environmental protection, and success. Less prevalent values included Security, Stimulation, Power, and Hedonism, while Tradition (2.03%) and Conformity (1.74%) were least represented. API themes influenced these patterns, with Wearable & IoT APIs showing stronger alignments with Benevolence, Security & Privacy APIs aligning with Self-Direction, and Gaming & VR APIs emphasising Achievement. Hedonism was linked to Multimedia & Content APIs, and Universalism was highlighted in Connectivity & Networking APIs. Differences across

Android and iOS platforms were also observed, with third-party APIs displaying higher alignments with personal values like Self-Direction and Achievement compared to standard APIs.

Accuracy of ChatGPT in revealing Values (RQ2). The findings highlight the role of expert opinions in evaluating ChatGPT's ability to reveal value alignments in API documentation. Overall, experts agreed with 75% of ChatGPT's inferred value alignments, either directly based on initial justifications or after clarifications. About 7.44% of opinions shifted to agreement following debates, underscoring ChatGPT's ability to influence experts' understanding of values. Values like Hedonism and Stimulation had higher agreement levels, whereas Universalism and Tradition showed more neutrality and disagreements. ChatGPT retracted mainly labels associated with Tradition, Conformity, and Achievement. Theme-based analysis of the value alignments indicated higher agreement for Security & Privacy APIs and lower accuracy for general-purpose themes like Testing & Debugging. Persistent disagreements stemmed from ChatGPT inferring indirect or abstract connections to values, leading to hallucinations when documentation lacked clarity. This emphasises the need for human oversight, blending automated reasoning with manual review to balance rigour and efficiency. We completed the human evaluation by two coders rating all of the endorsed value alignments, confirming 86% of the value alignments with substantial agreement between the coders (Cohen's Kappa value of 0.78).

ChatGPT Prompt Completion (RQ3). The findings highlight ChatGPT's behaviour in prompt completion based on response time, response size, and prompt size. Response time varies significantly, with faster completions for third-party APIs compared to standard APIs, and for standard iOS APIs over Android APIs. Response size, reflecting complexity, is larger for standard APIs than third-party APIs, although differences between iOS and Android APIs lack statistical significance. Prompt size analysis reveals that third-party API documentation is more content-rich than standard APIs, and standard iOS APIs contain more content than Android APIs. Correlation tests indicate a strong positive relationship between response time and size, implying longer times for more complex justifications. However, weak negative correlations between prompt size and both response time and response size suggest that smaller documentation tends to produce longer and more complex responses, though other factors likely contribute.

Future Work. Future research may expand our analysis to include other LLMs, such as GPT-4o, to compare their reasoning capabilities and reliability in

uncovering latent values in software artifacts. Such comparisons can help identify model-specific strengths and weaknesses, informing best practices for AI-driven evaluations. Additionally, hybrid AI-human frameworks should be explored to optimise reliability and reduce biases inherent in automated systems. Integrating human feedback into model training could further refine LLM performance. Examining the socio-technical dimensions of value disagreement such as cultural biases or variations in domain-specific interpretations would also provide richer insights into human-LLM interactions. Lastly, applying this methodology to other types of software artifacts, including source code, bug reports, and user feedback, can extend its applicability and reveal new patterns in value representations across software ecosystems. These directions underscore the need for interdisciplinary approaches combining AI research, ethics, and software engineering to advance the study of human values in software using large language models.

Notes

1. <https://developer.android.com/reference>.
2. <https://developer.apple.com/documentation>.
3. <https://globalaffects.org/chatgpt>.

Acknowledgments

 Davoud Mougouei conceived the core idea, designed the study, and led its execution. Saima Rafi assisted with data analysis. All authors contributed to drafting and refining the manuscript.

Data availability statement

The data can be accessed from <https://github.com/global-affects/alignments-apis/>.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- Alshami, Ahmad, Moustafa Elsayed, Eslam Ali, Abdelrahman E. E. Eltoukhy, and Tarek Zayed. 2023. "Harnessing the Power of Chatgpt for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions." *Systems* 11 (7): 351. <https://doi.org/10.3390/systems11070351>.
- Athaluri, Sai Anirudh, Sandeep Varma Manthena, V. S. R. Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. "Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing through ChatGPT References." *Cureus* 15 (4): e37432.
- Baldassarre, Maria Teresa, Vita Santa Barletta, Danilo Caivano, and Michele Scalera. 2020. "Integrating Security and Privacy in Software Development." *Software Quality Journal* 28 (3): 987–1018. <https://doi.org/10.1007/s11219-020-09501-6>.
- Beganovic, Adna, Muna Abu Jaber, and Ali Abd Almisreb. 2023. "Methods and Applications of ChatGPT in Software Development: A Literature Review." *Southeast Europe Journal of Soft Computing* 12 (1): 8–12.
- Bellamy, Rachel K. E., Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia. 2019. "Think Your Artificial Intelligence Software is Fair? Think Again." *IEEE Software* 36 (4): 76–80. <https://doi.org/10.1109/MS.2019.2908514>.
- Bellini-Leite, Samuel C. 2024. "Dual Process Theory for Large Language Models: An Overview of Using Psychology to Address Hallucination and Reliability Issues." *Adaptive Behavior* 32 (4): 329–343. <https://doi.org/10.1177/10597123231206604>.
- Bera, Palash, Yves Wautelet, and Geert Poels. 2023. "On the Use of ChatGPT to Support Agile Software Development." In *Short Paper Proceedings of the Second International Workshop on Agile Methods for Information Systems Engineering (Agil-ISE 2023) Co-Located with the 35th International Conference on Advanced Information Systems Engineering (CAiSE 2023), June 13, 2023*, Vol. 3414, 1–9. Zaragoza, Spain: CEUR-WS. org.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, 159. Vancouver, BC, Canada, Red Hook, NY, USA: Curran Associates Inc.
- Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33, 1877–1901.
- Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. "Evaluating Large Language Models Trained on Code." Preprint [arXiv:2107.03374](https://arxiv.org/abs/2107.03374).
- Cheng, An-Shou, and Kenneth R. Fleischmann. 2010. "Developing a Meta-Inventory of Human Values." *Proceedings of the American Society for Information Science and Technology* 47 (1): 1–10.
- Colavito, Giuseppe, Filippo Lanubile, Nicole Novielli, and Luigi Quaranta. 2024. "Leveraging GPT-like LLMs to Automate Issue Labeling." In *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*, 469–480. New York, NY: Association for Computing Machinery.
- Daun, Marian, and Jennifer Brings. 2023. "How ChatGPT Will Change Software Engineering Education." In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*. Vol. 1, 110–116. New York, NY: Association for Computing Machinery.

- Dogra, Varun, Sahil Verma, Marcin Woźniak, Jana Shafi, and Muhammad Fazal Ijaz. 2024. "Shortcut Learning Explanations for Deep Natural Language Processing: A Survey on Dataset Biases." *IEEE Access* 12 (2024): 26183–26195. <https://doi.org/10.1109/ACCESS.2024.3360306>.
- Dwivedi, Rudresh, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian. 2023. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions." *ACM Computing Surveys* 55 (9): 1–33. <https://doi.org/10.1145/3561048>.
- El-Kassas, Wafaa S., Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. "Automatic Text Summarization: A Comprehensive Survey." *Expert Systems with Applications* 165:113679. <https://doi.org/10.1016/j.eswa.2020.113679>.
- Fang, Xiao, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. "Bias of AI-generated Content: An Examination of News Produced by Large Language Models." *Scientific Reports* 14 (1): 5224. <https://doi.org/10.1038/s41598-024-55686-2>.
- Ferrara, Emilio. 2023. "Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models." *First Monday* 28 (2023).
- Fischer, Ronald, Markus Luczak-Roesch, and Johannes A. Karl. 2023. "What Does ChatGPT Return About Human Values? Exploring Value Bias in ChatGPT Using a Descriptive Value Theory." Preprint [arXiv:2304.03612](https://arxiv.org/abs/2304.03612).
- Friedman, Batya, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. "Value Sensitive Design and Information Systems." In *Early Engagement and New Technologies: Opening up the Laboratory*, 55–95. Springer.
- Giray, Louie. 2023. "Prompt Engineering with ChatGPT: A Guide for Academic Writers." *Annals of Biomedical Engineering* 51 (12): 2629–2633. <https://doi.org/10.1007/s10439-023-03272-4>.
- GPT-3.5. n.d. "GPT-3.5." Accessed August 18, 2024. <https://platform.openai.com/docs/models>.
- Guerreiro, Nuno M., Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. "Hallucinations in Large Multilingual Translation Models." *Transactions of the Association for Computational Linguistics* 11:1500–1517. https://doi.org/10.1162/tacl_a_00615.
- Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. "XAI-Explainable Artificial Intelligence." *Science Robotics* 4 (37): eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>.
- Guo, Biyang, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection." <https://arxiv.org/abs/2301.07597v1>.
- Haleema Bhatti. n.d. "Cyberbullying in Pakistan: The Case of Cyber Harassment Against Women." Accessed August 18, 2024. <https://www.paradigmshift.com.pk/cyberbullying-in-pakistan/>.
- Hanel, Paul H. P., Gregory R. Maio, Ana K. S. Soares, Katia C. Vione, Gabriel L. de Holanda Coelho, Valdiney V. Gouveia, Appasaheb C. Patil, Shanmukh V. Kamble, and Antony S. R. Manstead. 2018. "Cross-Cultural Differences and Similarities in Human Value Instantiation." *Frontiers in Psychology* 9:849. <https://doi.org/10.3389/fpsyg.2018.00849>.
- Hanusz, Zofia, Joanna Tarasinska, and Wojciech Zielinski. 2016. "Shapiro–Wilk Test with Known Mean." *REVSTAT-Statistical Journal* 14 (1): 89–100.
- Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. "Aligning AI With Shared Human Values." In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen. 2023. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions." *ACM Transactions on Information Systems* 43 (2023): 42.
- Huang, Hanyao, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin. 2023. "ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model." *International Journal of Oral Science* 15 (1): 29. <https://doi.org/10.1038/s41368-023-00239-y>.
- Hussain, Waqar, Davoud Mougouei, and Jon Whittle. 2018. "Integrating Social Values into Software Design Patterns." In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 8–14. New York, NY: Association for Computing Machinery.
- Hussain, Waqar, Harsha Perera, Jon Whittle, Arif Nurwidyantoro, Rashina Hoda, Rifat Ara Shams, and Gillian Oliver. 2020. "Human Values in Software Engineering: Contrasting Case Studies of Practice." *IEEE Transactions on Software Engineering* 48 (5): 1818–1833. <https://doi.org/10.1109/TSE.2020.3038802>.
- Ian, Blair. 2023. "Mobile App Download and Usage Statistics." Accessed July, 2023. <https://buildfire.com/app-statistics/>.
- Jalil, Sajed, Suzzana Rafi, Thomas D. LaToza, Kevin Moran, and Wing Lam. 2023. "Chatgpt and Software Testing Education: Promises & Perils." In *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 4130–4137. IEEE.
- Jamieson, Jack, Naomi Yamashita, and Eureka Foong. 2024. "Predicting Open Source Contributor Turnover From Value-Related Discussions: An Analysis of GitHub Issues." In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 1–13.
- Jason Cartwright. n.d. "The Government's COVID-19 Tracking App is Called CovidSafe and is Launching Today!" Accessed August 18, 2024. <https://techau.com.au/the-governments-covid-19-tracking-app-is-called-covidsafe-and-is-launching-today/>.
- Joanne Cavanaugh Simpson. n.d. "South Florida Police Widely Use Facial Recognition, Yet Resist Policies To Curb Abuse." That's a Problem for People of Color. Accessed August 18, 2024. <https://pulitzercenter.org/stories/south-florida-police-widely-use-facial-recognition-yet-resist-policies-curb-abuse-thats>.
- Khan, Javed Ali, Affan Yasin, Rubia Fatima, Danish Vasan, Arif Ali Khan, and Abdul Wahid Khan. 2022. "Valuating Requirements Arguments in the Online User's Forum for Requirements Decision-Making: The CrowdRE-VArg Framework." *Software: Practice and Experience* 52 (12): 2537–2573.
- Khojah, Ranim, Mazen Mohamad, Philipp Leitner, and Francisco Gomes de Oliveira Neto. 2024. "Beyond Code Generation: An Observational Study of Chatgpt Usage in

- Software Engineering Practice.” *Proceedings of the ACM on Software Engineering* 1 (FSE): 1819–1840. <https://doi.org/10.1145/3660788>.
- Kortum, Philip, and Mary Sorber. 2015. “Measuring the Usability of Mobile Applications for Phones and Tablets.” *International Journal of Human-Computer Interaction* 31 (8): 518–529. <https://doi.org/10.1080/10447318.2015.1064658>.
- Kou, Bonan, Shengmai Chen, Zhijie Wang, Lei Ma, and Tianyi Zhang. 2024. “Do Large Language Models Pay Similar Attention like Human Programmers when Generating Code.” *Proceedings of the ACM on Software Engineering* 1 (FSE): 2261–2284. <https://doi.org/10.1145/3660807>.
- Krishtul, Sara, Mojtaba Shahin, Humphrey O. Obie, Hourieh Khalajzadeh, Fan Gai, Ali Rezaei Nasab, and John Grundy. 2022. “Human Values Violations in Stack Overflow: An Exploratory Study.” In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, 70–79. New York, NY: Association for Computing Machinery.
- Kuzman, Taja, Igor Mozetič, and Nikola Ljubešić. 2023. “ChatGPT: Beginning of An End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.” <https://arxiv.org/abs/2303.03953v2>.
- Lee, Jinsook, Yann Hicke, Renzhe Yu, Christopher Brooks, and René F. Kizilcec. 2024. “The Life Cycle of Large Language Models in Education: A Framework for Understanding Sources of Bias.” *British Journal of Educational Technology* 55 (5): 1982–2002. <https://doi.org/10.1111/bjet.v55.5>.
- Lehtinen, Teemu, Charles Koutchme, and Arto Hellas. 2024. “Let’s Ask AI About Their Programs: Exploring ChatGPT’s Answers To Program Comprehension Questions.” In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, 221–232. New York, NY: Association for Computing Machinery.
- Liao, Q. Vera, and Jennifer Wortman Vaughan. 2023. “AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap.” Preprint [arXiv:2306.01941](https://arxiv.org/abs/2306.01941).
- Lindahl, Caroline, and Helin Saeid. 2023. “Unveiling the Values of ChatGPT: An Explorative Study on Human Values in AI Systems.”
- Ma, Wei, Shangqing Liu, Wenhan Wang, Qiang Hu, Ye Liu, Cen Zhang, Liming Nie, and Yang Liu. 2023. “The Scope of ChatGPT in Software Engineering: A Thorough Investigation.” Preprint [arXiv:2305.12138](https://arxiv.org/abs/2305.12138).
- Maev Duggan. n.d. “Online Harassment.” Accessed August 18, 2024. <https://www.pewresearch.org/internet/2014/10/22/online-harassment/>.
- McKnight, Patrick E., and Julius Najab. 2010. “Mann-Whitney U Test.” In *The Corsini Encyclopedia of Psychology*, 1–1. John Wiley & Sons, Ltd.
- Mougouei, Davoud. 2020. “Engineering Human Values in Software through Value Programming.” In *ICSEW’20: Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering*, 133–136. New York, NY: Association for Computing Machinery.
- Mougouei, Davoud, Harsha Perera, Waqar Hussain, Rifat Shams, and Jon Whittle. 2018. “Operationalizing Human Values in Software: A Research Roadmap.” In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 780–784. New York, NY: Association for Computing Machinery.
- Nurwidyantoro, Arif, Mojtaba Shahin, Michel Chaudron, Waqar Hussain, Harsha Perera, Rifat Ara Shams, and Jon Whittle. 2023. “Integrating Human Values in Software Development Using a Human Values Dashboard.” *Empirical Software Engineering* 28 (3): 67. <https://doi.org/10.1007/s10664-023-10305-y>.
- Nurwidyantoro, Arif, Mojtaba Shahin, Michel R. V. Chaudron, Waqar Hussain, Rifat Shams, Harsha Perera, Gillian Oliver, and Jon Whittle. 2022. “Human Values in Software Development Artefacts: A Case Study on Issue Discussions in Three Android Applications.” *Information and Software Technology* 141:106731. <https://doi.org/10.1016/j.infsof.2021.106731>.
- Obie, Humphrey O., Waqar Hussain, Xin Xia, John Grundy, Li Li, Burak Turhan, Jon Whittle, and Mojtaba Shahin. 2021. “A First Look at Human Values-Violation in App Reviews.” In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 29–38. IEEE.
- Oliveira, Francisco Braulio, Amanul Haque, Davoud Mougouei, Simon Evans, Jaime Simao Sichman, and Munindar P. Singh. 2022. “Investigating the Emotional Response to COVID-19 News on Twitter: A Topic Modeling and Emotion Classification Approach.” *IEEE Access* 10:16883–16897. <https://doi.org/10.1109/ACCESS.2022.3150329>.
- Oliveira, Francisco Braulio, Davoud Mougouei, Amanul Haque, Jaime Simão Sichman, Hoa Khanh Dam, Simon Evans, Aditya Ghose, and Munindar P. Singh. 2023. “Beyond Fear and Anger: A Global Analysis of Emotional Response to COVID-19 News on Twitter.” *Online Social Networks and Media* 36:100253. <https://doi.org/10.1016/j.osnem.2023.100253>.
- Pellert, Max, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2022. “AI Psychometrics: Assessing the Psychological Profiles of Large Language Models through Psychometric Inventories.” *Perspectives on Psychological Science* 19:17456916231214460.
- Perera, Harsha, Waqar Hussain, Davoud Mougouei, Rifat Ara Shams, Arif Nurwidyantoro, and Jon Whittle. 2019. “Towards Integrating Human Values Into Software: Mapping Principles and Rights of GDPR to Values.” In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 404–409. IEEE.
- Perera, Harsha, Waqar Hussain, Jon Whittle, Arif Nurwidyantoro, Davoud Mougouei, Rifat Ara Shams, and Gillian Oliver. 2020. “A Study on the Prevalence of Human Values in Software Engineering Publications, 2015–2018.” In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, 409–420. New York, NY: Association for Computing Machinery.
- Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. “Is ChatGPT a General-Purpose Natural Language Processing Task Solver?” In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1339–1384.
- Rahmaniar, Wahyu. 2023. “ChatGPT for Software Development: Opportunities and Challenges.” *IT*

- Professional* 26 (3): 80–86. <https://doi.org/10.1109/MITP.2024.3379831>.
- Rao, Arya, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, Keith J. Dreyer, and Marc D. Succi. 2023. “Evaluating GPT as An Adjunct for Radiologic Decision Making: GPT-4 versus GPT-3.5 in a Breast Imaging Pilot.” *Journal of the American College of Radiology* 20 (10): 990–997. <https://doi.org/10.1016/j.jacr.2023.05.003>.
- Ronanki, Krishna, Beatriz Cabrero-Daniel, and Christian Berger. 2022. “ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box? In *Agile Processes in Software Engineering and Extreme Programming -- Workshops*, edited by Kruchten Philippe and Gregory Peggy, 173–181. Cham: Springer Nature Switzerland.
- Schwartz, Shalom H. 1992. “Universals in The Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries.” In *Advances in Experimental Social Psychology*, edited by Mark P. Zanna. Vol. 25, 1–65. Academic Press.
- Schwartz, Shalom H. 2012. “An Overview of the Schwartz Theory of Basic Values.” *Online Readings in Psychology and Culture* 2 (1): 11. <https://doi.org/10.9707/2307-0919.1116>.
- Schwartz, Shalom H., Cieciuch Jan, Vecchione Michele, Davidov Eldad, Fischer Ronald, Beierlein Constanze, Ramos Alice, Verkasalo Markku, Lönnqvist Jan-Erik, Demirutku Kelebek, Dirilen-Gumus Ozlem, and Konty Mark. 2012. “Refining the Theory of Basic Individual Values.” *Journal of Personality and Social Psychology* 103 (4): 663–688.
- Smith, Andrew L., Felix Greaves, and Trishan Panch. 2023. “Hallucination Or Confabulation? Neuroanatomy as Metaphor in Large Language Models.” *PLOS Digital Health* 2 (11): e0000388. <https://doi.org/10.1371/journal.pdig.0000388>.
- Surameery, Nigar M. Shafiq, and Mohammed Y. Shakor. 2023. “Use Chat Gpt to Solve Programming Bugs.” *International Journal of Information Technology & Computer Engineering (IJITC)* 3 (1): 17–22. <https://doi.org/10.55529/ijitc>.
- Thew, Sarah, and Alistair Sutcliffe. 2018. “Value-Based Requirements Engineering: Method and Experience.” *Requirements Engineering* 23 (4): 443–464. <https://doi.org/10.1007/s00766-017-0273-y>.
- Wang, Wei, Huilong Ning, Gaowei Zhang, Libo Liu, and Yi Wang. 2024. “Rocks Coding, Not Development: A Human-Centric, Experimental Evaluation of LLM-Supported SE Tasks.” *Proceedings of the ACM on Software Engineering* 1 (FSE): 699–721. <https://doi.org/10.1145/3643758>.
- White, Jules, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C. Schmidt. 2024. “Chatgpt Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design.” In *Generative AI for Effective Software Development*, edited by Nguyen-Duc Anh, Abrahamsson Pekka and Khomh Foutse 71–108. Cham: Springer Nature Switzerland.
- Zhang, Jianzhang, Yiyang Chen, Chuang Liu, Nan Niu, and Yinglin Wang. 2023. “Empirical Evaluation of ChatGPT on Requirements Information Retrieval Under Zero-Shot Setting.” In *2023 International Conference on Intelligent Computing and Next Generation Networks (ICNGN)*, 1–6. IEEE.
- Zhao, Haiyan, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. “Explainability for Large Language Models: A Survey.” *ACM Transactions on Intelligent Systems and Technology* 15 (2): 1–38. <https://doi.org/10.1145/3639372>.
- Zhong, Qihuang, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. “Can Chatgpt Understand Too? A Comparative Study on Chatgpt and Fine-Tuned Bert.” Preprint [arXiv:2302.10198](https://arxiv.org/abs/2302.10198).