A Framework to Characterize Reporting on Generative AI Use

Agathe Balayn
balaynagathe@microsoft.com
Microsoft Research
New York City, New York, USA

Varun Nagaraj Rao
varunrao@princeton.edu
Princeton University
Princeton, USA

Su Lin Blodgett
Microsoft Research
Montréal, Canada

Aylin Caliskan
University of Washington
Seattle, Washington, USA

Solon Barocas
Microsoft Research
New York City, New York, USA

Abstract

Unlike with traditional predictive AI models, today's generative AI models are increasingly designed to be general-purpose, able to perform a wide range of tasks. This makes it challenging to develop a reliable and useful understanding of the ways in which this technology is and could be used. As a result, academic and policy researchers and generative AI providers have started to publish the results of their own investigations about the use of generative AI. This information is, however, fragmented, potentially incomplete, sometimes ambiguous, and often lacking in methodological specificity. In this paper, we conducted an integrative review to build a multi-dimensional framework that specifies what kind of information about generative AI use could be reported and how, and illustrated its analytical utility by applying the framework to a collection of over 110 industry documents. Our analysis reveals systematic patterns and omissions in current industry reporting and reflects on the narratives this reporting collectively advance about generative AI use.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → *User models*; *Interaction paradigms*; HCI theory, concepts and models; • **Social and professional topics** → User characteristics.

Keywords

Technology use, transparency, documentation, AI, generative AI, general-purpose AI, framework

ACM Reference Format:
Agathe Balayn, Varun Nagaraj Rao, Su Lin Blodgett, Aylin Caliskan, and Solon Barocas. 2026. A Framework to Characterize Reporting on Generative AI Use. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 44 pages. https://doi.org/10.1145/3772318.3791649

1 Introduction

Unlike traditional predictive AI models that are developed to perform a specific task, today's generative AI (GenAI) models are designed to be general-purpose [23], able to perform a wide range of tasks. While it was previously obvious how predictive AI models would be used in practice (because they could only be used to perform the task for which they were developed), it is now much more challenging to anticipate how GenAI models will be used. The open-ended nature of GenAI models means that people are free to try to use these models however they please, limited only by their own imagination. This generality raises a difficult question: how should we go about trying to develop an understanding of the many ways that GenAI can be, will be, or is already being used in practice?

To date, this question has most often been answered by evaluating GenAI models on a range of benchmark datasets designed to measure the degree to which models exhibit certain broad capabilities (e.g., mathematical reasoning) [24]. The reported capabilities of these models are often presented as meaningful evidence of the many *possible* uses to which they could be put (e.g., a model that excels at mathematical reasoning can be used to perform any kind of task that relies on such reasoning). Unfortunately, how models perform on benchmark evaluations may tell us little about the tasks that *actual* users will attempt to perform with these models. Just because a model happens to do well on a benchmark evaluation does not mean that people will use the model to perform tasks that tap into the capabilities that the benchmark evaluation is designed to measure. If real-world uses of GenAI models look very different from the kinds of uses tested in benchmark evaluations, the results of these evaluations will also tell us little about how well such models will perform on these real-world tasks or the kinds of harms to which such real-world use might give rise.

Growing recognition of the limits of benchmark evaluations has led to a wide variety of efforts to develop more reliable accounts of the many different ways people actually use GenAI, including user studies [e.g., 9], surveys [e.g., 71, 95], analyses of user self-reports [e.g., 106], and even log analyses [e.g., 81, 97, 99, 107]. The questions answered by these studies can vary widely. For example, some explore overall adoption rates of GenAI, occasionally broken down by model type, and often without telling the reader anything about what is being done with AI once it has been adopted [e.g., 64]. Others are focused on the activities that people perform with GenAI, often reported in terms of task or topic [e.g., 46, 108]. Still others explore who is adopting GenAI or where it is being adopted, noting differences across different populations and often inferring the likely way GenAI is being used from the fact that adoption is more common in some industries or fields than others [e.g., 13]. These efforts have begun to produce a much richer account of what GenAI use looks like on the ground, but they are still quite rare

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791649

and often conducted in isolation from one another. The result is a fragmented picture of GenAI use—spread across disparate artifacts, ranging from business reports to policy briefs, marketing materials, developer resources, system cards, research papers, and more—often answering different questions using different methods. This variation is perhaps unsurprising since there is, of course, no one way to report on the use of GenAI. And yet, little of this work reflects on the wide range of possibilities for reporting on use, the relative merits of different approaches, and the degree to which different kinds of reporting might be used in concert to provide more useful accounts of use.

In this paper, we therefore seek to understand the range of possible ways of reporting on GenAI use and the current ways in which GenAI use is commonly reported. In analyzing current reporting practices, we also aim to critically assess the broader narrative about GenAI advanced by such reporting. In particular, we seek to answer:

RQ1: What are the possible ways of reporting on the use of generative AI? What are possible methods that can be employed to generate such reports?

RQ2: How do the major developers of generative AI systems report on the use of their systems? What kinds of narratives does such reporting promote?

To answer the first research question, we reviewed a diverse range of artifacts reporting on GenAI use ($N = 40$), drawn from industry, government, civil society, the academy, and the media, working inductively to identify key axes along which reporting can vary. We summarize these findings in the form of a framework designed to account for all the major dimensions of potential variation in reporting on AI use, organized as a series of distinct questions (e.g., Who is using GenAI? Where is it being used? What is it being used for?). Our framework also maps out the different methods and sampling strategies that might be employed to answer these questions. In so doing, we do *not* propose a novel taxonomy that attempts to enumerate all the possible ways that people could use GenAI; our framework instead offers a meta-taxonomy to account for the many different ways one could go about taxonomizing the use of GenAI.

To answer the second research question, we leveraged our framework to analyze a corpus of publicly available documents ($N = 111$) from major developers of GenAI systems, annotating the types of reporting and methods employed in each document. By design, the corpus includes many different types of documents (e.g., marketing materials, educational resources, developer notes, technical reports, etc.), with the goal of capturing the diverse ways in which industry currently reports on GenAI use for its many different audiences. Our annotations reveal notable patterns in reporting on GenAI use and in the methods used to generate such reports. For instance, we find that the industry documents adopt a diversity of methods to collect information about use, which results in a surprisingly rich landscape of the information covered in our framework. Yet significant gaps remain, including in terms of the questions that are addressed (e.g., *where* and *when* users use GenAI is rarely reported) or the temporalities in which these questions are answered (documents report more often about *potential* uses than *actual* uses). Moreover, many documents tend to under-specify the methods employed to generate the reported findings and make vague and over-generalized statements, especially given the data upon which statements are based, which might be especially misleading for an inattentive reader.

Taken together, the paper makes four main contributions:

- An empirically-informed and -validated framework that maps the space of possibilities for reporting on GenAI use, which can both guide the development of future reporting and ground the analysis of existing reporting;
- An analysis applying this framework to a corpus of documents from the major GenAI providers, spanning a diverse range of publicly available artifacts, revealing current industry reporting practices;
- A critical reflection on the limitations of current reporting practices, the narratives advanced by such reporting, and opportunities to improve on the current state of practice.
- A publicly available dataset of our extensively annotated corpus of documents from the major GenAI providers, which we make available for other researchers to explore further.[1]

2 Background & Related Work

Longstanding efforts across the academy, civil society, and government have urged greater transparency surrounding GenAI. For example, Narayanan and Kapoor [74], drawing lessons from social media governance, argue that GenAI providers should publish platform-style transparency reports to remedy today's "data vacuum" about harms and misuse, while Bommasani et al. [16] formalize that proposal as Foundation Model Transparency Reports and link its schema to the Foundation Model Transparency Index (FMTI) [15], which measures GenAI model provider disclosures. Legislatures have begun to mandate disclosures: the EU's Digital Services Act (DSA) now requires platform transparency reporting and data access, and the EU AI Act sets obligations for GenAI models on a phased timeline; in the U.S., the proposed AI Foundation Model Transparency Act would direct the Federal Trade Commission to set disclosure standards. These calls build on—and have helped accelerate the adoption of—artifacts and guidance like Model Cards [70], Datasheets for Datasets [42], the CLeAR documentation framework [27], Data Nutrition Labels [49], and a human-centered transparency roadmap [61].

While these calls target GenAI broadly, interest in transparency around GenAI *use*—with the purpose of providing lawmakers and regulators with actionable evidence for regulating GenAI—is comparatively more recent. Scholars and advocates have begun to emphasize that effective AI regulation requires understanding how GenAI is actually used, not just its capabilities [14, 22, 41], and to urge companies to share use information [e.g., 22, 74, 76]. While a good deal of information has been released about GenAI systems, due in part to the calls described above, until quite recently little of this information has focused on use. Historically, documents released by AI providers, deployers, and individual practitioners have tended to describe technical specifications of systems and their development processes, as well as assessments of their functionalities and potential impacts [2, 3, 7, 32, 43, 54]. While such documents can

[1] The dataset is publicly available at https://github.com/agathe-balayn/GenAI-Usages-Reports.

provide clues about potential use—e.g., via descriptions of what systems are thought to be capable of or of what developers or providers envision as projected use cases—they do not typically center on the actual use of the systems they describe.

By contrast, other documents directly addressing system use have more recently emerged. These documents come from many different sources—e.g., academic, industrial, and policy researchers and organizations—and take many different forms—e.g., research publications, reports evaluating in-context system use, user discussion forums, and marketing materials (see Table 6). Within the more research-like literature, there are a number of notable trends. A growing body of work has begun to investigate adoption and diffusion of GenAI, initially finding rapid, if uneven, uptake and a later deceleration [8, 13, 93]. Related work has also explored various factors affecting adoption, focusing on psychological, institutional, and broader social dynamics that affect perceived utility and acceptability [20, 68, 105]. A separate body of work has begun to look more closely at the activities that users perform when using GenAI, increasingly drawing on logs of real-world use [25, 26, 45, 46, 66, 81, 86, 97, 99, 106]. Closely related work has also begun to examine the way users interact with these models when performing such tasks, identifying trends in prompting, routine challenges in eliciting the desired model behavior, and potentially sensitive disclosures [55, 69, 73, 84, 88, 89, 107, 108]. Other work has also investigated the factors that affect the perceived utility, acceptability, and risks of different GenAI use cases [71, 72]. Building on data from real-world use, a number of researchers have developed new benchmark datasets designed to better reflect realistic use of GenAI, often motivated by a recognition that performance on hand-crafted benchmark datasets tends to give a misleading impression of real-world performance [56, 63, 82, 102, 103]. Other scholars have attempted to study the ultimate impact of adoption and use, sometimes using analytic methods to determine potential exposure to job replacement and at other times using observational or experimental methods to measure real-world downstream effects on users' productivity and psychological well-being [19, 47, 83]. Finally, there is also a growing collection of papers that have attempted to develop methods for GenAI provider-independent post-deployment monitoring of real-world use and impact, often based on new large-scale data collection infrastructure, including crowdsourcing [30, 35, 59, 85].

Despite this steady expansion in the range of ways that people learn about how AI can be and is already being used in practice, the landscape remains fragmented, and the range of possible ways of talking about GenAI use has not been systematically explored. In this work, we take a first step toward mapping this space by systematically characterizing how GenAI use is described across sources, and by proposing a structured vocabulary that helps make these descriptions more comparable and actionable.

3 Method: Framework Development

We set out to develop a framework that could account for the many different ways that it is possible to report on the use of GenAI. We also wanted to identify key dimensions of variation in reporting, including methodological differences, that would enable readers of use-related documents to better appreciate the choices made (and not made) in such reporting, identify when these documents include ambiguous or unsubstantiated claims, and effectively assess the validity of the reported conclusions. With these requirements in mind, we set out to develop an initial framework by completing an integrative review [94] (Figure 1).[2]

An integrative review is preferred when a field is emergent—as is the case here—and the goal is not to comprehensively survey an entire field, but instead to develop a conceptual framework to critically analyze the field. An integrative review first requires us to gather documents in a "creative" manner to make sure that we cover the wide diversity of sources that deal with use. It then requires us to critically analyze these documents in order to develop the framework. Our goal was thus not to gather all documents that tangentially pertain to the use of GenAI systems, as there are seemingly countless documents that report similar types of use information (e.g., capabilities and risks of GenAI). Instead, we only wanted to ensure broad coverage of the diverse types of information that these documents might report.

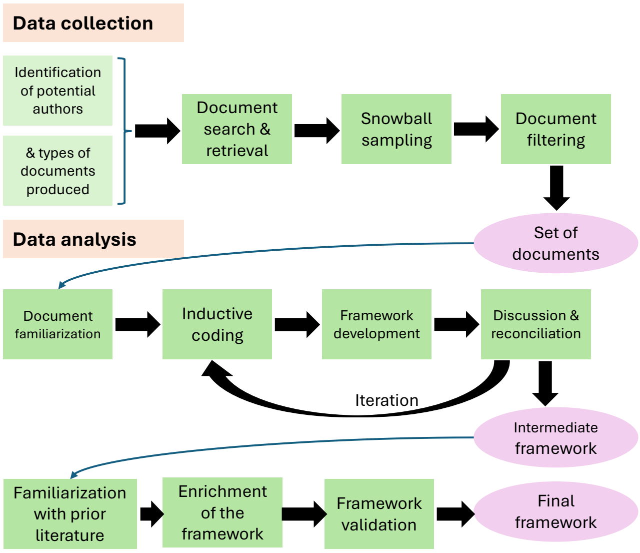


Figure 1: Method for developing our framework.

3.1 Data Collection

In order to collect the corpus of documents ($N = 40$) for the integrative review, we took the following steps.

(1) ***Source mapping.*** We identified the types of actors who might produce use-related documents (academia, industry, civil society, government, community forums) and the different document types that each produces (see Appendix Table 6). We refined our initial lists via exploratory Web and Google Scholar searches to locate specific organizations and document types.
(2) ***Targeted retrieval.*** We pulled documents from the websites of relevant organizations and collected relevant publications using academic search engines.

[2] Details about our methodology can be found in Appendix A.

(3) ***Snowballing.*** We expanded the pool via references, citations, outbound links, and related items on organizations' websites.
(4) ***Screening.*** We applied inclusion and exclusion criteria to finalize the corpus (see Appendix A.2).

3.2 Data Analysis

To build the framework, we iteratively analyzed the corpus.

Step 1: Development of an initial version of the framework. To identify the initial dimensions of the framework, we analyzed a purposive sample of 15 documents from the corpus, selected to maximize diversity across author and document types. We focused on the types of information reported about use and on the types of evidence collected that enabled such reporting. After collecting passages from the documents that fit these two themes, we conducted an inductive, reflexive thematic analysis: we first performed open coding of segments related to use and evidence, then iteratively compared and grouped codes into higher-level themes. This led to an initial set of descriptive categories (e.g., use-related questions such as "who" uses GenAI and "when," methods such as log analysis and user studies) that constituted the basis for a tentative framework. Throughout this step, all co-authors participated in joint coding discussions; when interpretations diverged, we revisited the underlying passages and resolved differences by consensus.

Step 2: Improvement of the framework. To make sure that no obvious types of information or methods had been neglected as well as to enrich the proposed categories in the initial framework, the first author familiarized herself with the literature about technology use beyond the context of AI (Appendix A.1). We drew on this broader literature as a sensitizing lens rather than a source of a priori categories. It helped us check whether the inductively derived themes covered known dimensions of technology use (e.g., context of use, interactional patterns, perceived outcomes), while retaining only categories that were meaningful for our corpus. With this knowledge in mind, the author proposed re-organizations of the framework to better distinguish between information about context and interaction while keeping the framework simple by further merging some categories. For instance, while the initial framework separated information about the impact of GenAI from the reasons for using GenAI, we decided to bring these categories together under the "why" as prior literature suggests that the reason one might use a technology is co-created while using it and observing its impacts. Three of the authors also worked on identifying commonalities and differences among the methods that had been annotated in the first step. In particular, they drew a distinction between reporting that takes GenAI model properties and human-model interactions as its object of analysis and between reporting that relies on empirical methods and axiomatic reflections. Conceptually, this corresponds to an axial-coding phase in which relationships between initial themes were made explicit.

Step 3: Identification of additional framework components and categories. While conducting the analysis above, we realized that many assertions in the documents were ambiguous, unsubstantiated, and sometimes incorrect. Hence, we revised the framework to include a component that would allow us to highlight such issues (e.g., unclear evidence source). In particular, we added a third componentto capture whether the reporting describes the study population—that is, the specific systems, users, and interaction contexts that comprise the study sample—and whether the identified study population actually supports the reported findings. This step completed the triadic structure of the framework (information, methods, and sample).

Step 4: Validation of the framework. The first author reviewed the remaining 25 documents in the corpus to make sure that no other type of use information, methods, or sampling had been missed. The first author also used these documents to refine the descriptions of each framework component (see Appendix B for example passages). Across Steps 1–4, all 40 documents in the corpus were thus used either to construct, refine, or validate the framework.

4 Framework Description

Our framework identifies the major dimensions along which reporting on GenAI use might vary. We cluster these dimensions into the three overarching components necessary for reporting on GenAI use (Figure 2): use-related research questions, methods for answering those questions, and populations to which such methods will be applied. In what follows, we describe each dimension of possible variation within each component.

4.1 Research questions

4.1.1 Use questions. Through our coding process, we identified seven types of questions relevant to the use of GenAI systems, which we describe in Table 1 (a more extensive description can be found in Appendix B.1). Inspired by prior literature, we explicitly distinguish between questions that touch upon users' interactions with a GenAI system (including the goals users have when using a system and the specific strategies they employ to prompt it), and questions that deal with the context in which the users adopt the system (including the organizational environment in which they interact with the system). Note that each use question can be answered using many possible units of analysis (e.g., answering *who* uses GenAI can call for a fine-grain description of an individual user, or a mention of their employer, job, or country, etc). Since we cannot enumerate in advance all the possible units in which such questions might be answered, our framework does not attempt to predefine these possibilities. Instead, our framework is designed to identify the major dimensions within which there will be such variation—and to provide a top-down structure within which units can be identified in a bottom-up manner, as demonstrated in Section 6.

4.1.2 Temporalities of use. Answers to these questions can also vary in their temporality: they might reflect the present situation, be it how users *currently* use GenAI (termed *actual* in the framework) or how they *could* use it *now*, e.g., if they chose to access a GenAI system (termed *potential*). Alternatively, documents can also report on use projected to be possible in the future (termed *projected*), based on how GenAI technologies are envisioned to improve over time.

4.2 Study design

4.2.1 Object of analysis (model/human-model interaction). From our coding exercise, we identified two potential objects of analysis

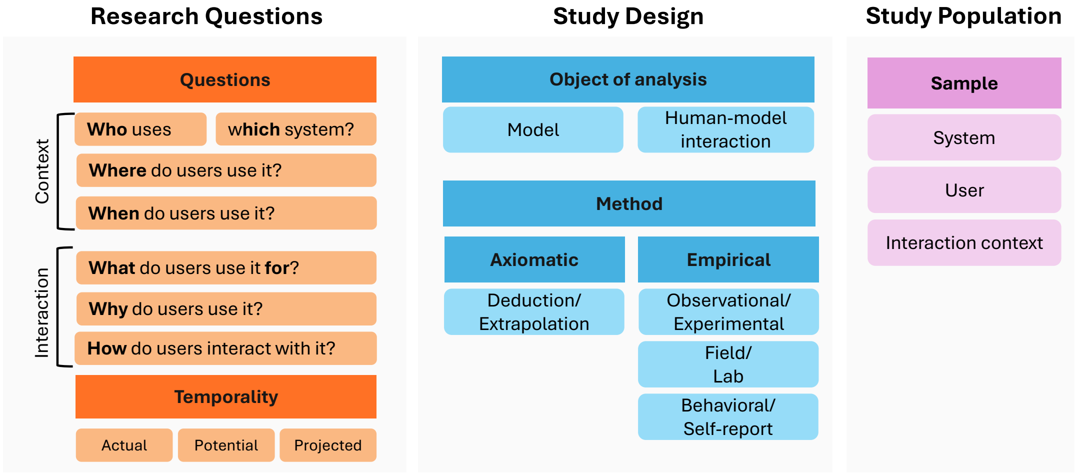


Figure 2: GenAI use framework.

Table 1: Summary of use questions in the framework.

Category	Question	Description	Example units
Use context and adoption	**Who uses GenAI?**	People who use GenAI and their characteristics	Socio-demographic attributes (age, gender, ethnicity, education, country); occupational role; language; attitudes toward the technology
	Which GenAI is used?	Systems in use and key properties	System/provider identifier; model family or version; capability categories; restrictions; implementation details; cards/datasheets
	Where do users use GenAI?	Settings where GenAI is used	Physical setting; geographic region; organizational context (sector, supportiveness toward new technologies); digital environment (developer tool, consumer platform, integrated application)
	When do users use GenAI?	Temporal patterns of GenAI use	Time or occasion of use (e.g., weekday vs weekend); frequency (once, weekly, daily); duration of session (minutes vs hours)
Interaction and appropriation	**What do users use GenAI for?**	Activities carried out with GenAI	Types of activities or tasks in which users engage (work-related, study-related, personal)
	Why do users use GenAI?	Motivations and goals for using GenAI	Objectives such as increasing quality, speed, or productivity; ability to perform activities that would otherwise be difficult or impossible; anticipated second-order effects
	How do users interact with GenAI?	Interaction strategies and practices	Prompting strategy (single vs iterative prompts); degree of automation in prompting; language choice in prompts; behavioral adaptations to the system's affordances

in reporting on GenAI use. Because we are studying the *use* of GenAI systems by end users, the most common object of analysis is the user and their *interaction* with the system. For instance, the results of a user study often describe how end users have made use of a system (e.g., what kinds of prompting strategies they have employed). However, reporting on GenAI use can also sometimes take *GenAI models or systems*, in isolation, as their objects of analysis, such as when the capabilities of a model are tested using a benchmark evaluation (e.g., [60]). Such reporting is often presented as useful evidence of potential use because it is assumed that users will employ models to perform tasks enabled by such capabilities.

4.2.2 Method. Broadly, we identified two categories of methods commonly employed to answer questions about GenAI use. Either *empirical* methods are used, such as user studies, workshops, log analyses, or evaluation benchmarks; or *axiomatic* reflections are used, where authors reason from principles, assumptions, or trends without collecting new data. The natures of empirical methods often differ depending on the object of analysis (e.g., benchmarks, red-teaming, capability evaluations for models, usage logs, surveys, workshops with people using or anticipating using GenAI systems for interactions). Axiomatic reflections can consist of logical projections that envision future trends (e.g., observing an increase in adoption of GenAI systems at time t and inferring an even steeper increase at time $t + 1$). They can also take the form of deductive inferences, such as when a model is shown to exhibit a given capability and the author reasons syllogistically that it could be used for specific activities or in specific domains where such a capability is relevant. In many documents, however, the underlying method is *unstated* or *ambiguous*.

4.2.3 Empirical methods. Empirical methods vary in substantial ways. Some draw on logs of real end users interacting with deployed systems, while others rely on controlled experiments or one-off user studies in lab-like settings. Because these differences shape how the resulting evidence can be interpreted (e.g., whether it speaks to *actual* or *potential* use), we characterized each empirical method along three cross-cutting dimensions.

- **Observational versus experimental study:** We label a method as *observational* when it passively records or analyzes behavior or outputs without introducing a systematic intervention (e.g., analysis of usage logs from a deployed chatbot, descriptive reports of how a tool is used in practice, or ratings collected on organically occurring model outputs). In contrast, we label it *experimental* when the authors deliberately manipulate conditions or inputs in order to test effects or capabilities—for example, randomized controlled trials (RCTs) with users, A/B tests on alternative system designs, scripted red-teaming of models, or standardized benchmark evaluations where prompts and scoring criteria are fixed by the researchers.

- **Field versus laboratory study:** We use *field* for methods that analyze data from in-the-wild deployments, such as telemetry from production systems, log data from real end users, or studies embedded in existing organizational workflows. We use *laboratory* for more controlled or contrived settings, including online experiments, internal pilot deployments, or user studies where participants are recruited to perform specific tasks under researcher-defined conditions. For example, user-based studies such as randomized controlled trials that recruit participants and assign them to model and control conditions in a synthetic task environment are treated as lab studies, whereas analyses of how workers actually use a GenAI assistant in their daily work are treated as field studies.
- **Behavioral versus self-report**: We label evidence as *behavioral* when it is based on what people or models do: e.g., interaction logs, click-through traces, acceptance or rejection of suggestions, code edits, or other observable actions. We label it as *self-report* when the evidence consists of what people say about their use, experiences, or expectations: e.g., survey responses, interview transcripts, diary entries, or human ratings of model outputs (such as safety or quality judgments).

4.3 Study population

4.3.1 Sample frame. GenAI use refers to the use of a GenAI *system* by an *individual user* in some mode or context of *interaction*. Each of these—system, user, interaction context—can be their own basis for developing a sample population for reporting on GenAI use.

- **System**: Information about a single specific system might be collected, or information about a group of systems that share certain characteristics, e.g., the same modality or the same type of access point.
- **User**: The sampled population of users included can be determined according to all sorts of characteristics, including their socio-demographic features, their employment in a specific industrial sector, or their specific occupational role.
- **Interaction context**: The data can be collected within a specific timeline (i.e., in-between two specific dates) and using various contextual exclusion criteria, e.g., only interactions that rely on specific modalities (e.g., only text) or languages (e.g., often solely English interactions are studied), no user interactions relying on certain modes of access (e.g., via desktop). The degree of interaction might even be used to set the sampling frame for a study: only those users who make use of a GenAI system on a regular, rather than one-off basis, will be included in the study.

4.3.2 Generalizing from the sample. Notably, the scope of the sampled population can differ from the scope of the claims made in reports on GenAI use. Even worse, the sampling frame might not be made explicit in the reporting, rendering it impossible to appropriately interpret the generalizability of the finding.

5 Method: Framework Application

To exploit our framework and investigate current reporting practices around GenAI use, we applied it to industry documents that report on use, and analyzed the trends highlighted by our framework. Details about our method (Figure 3) are provided in Appendix A.

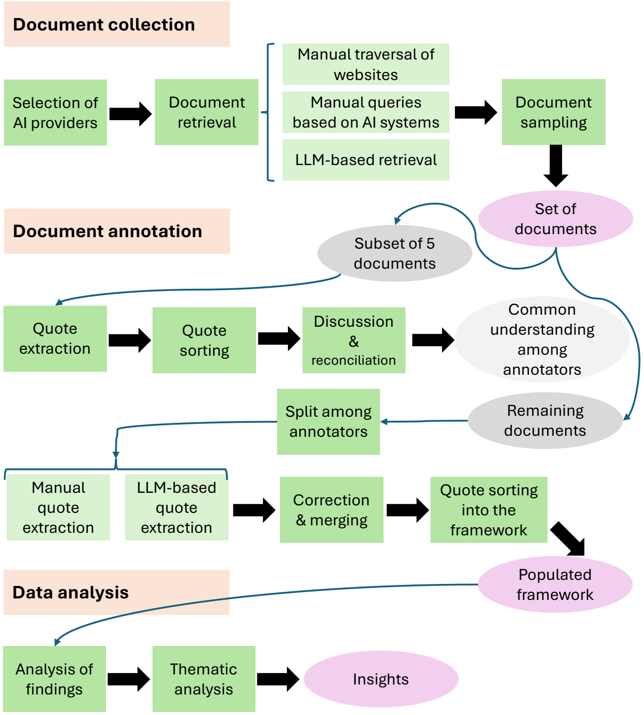


Figure 3: Method for exploiting our framework.

Document collection. We collected a corpus of documents produced by industry GenAI providers as they are arguably the actors in the AI supply chain who are best placed to collect and release information about the use of their GenAI systems. We selected six influential model providers—**AI2, Anthropic, Google, Meta, Microsoft, OpenAI**—who are known to have large user bases and who report use-related information. For each provider, we compiled a corpus that reflects the kinds of information they typically release about AI-in-use, with a total number of 111 documents (Table 2).

Document annotation. To apply the framework to the corpus of industry documents, we turned the framework into a codebook with which we annotated the corpus—i.e., we extracted quotes from each document and matched them to each entry in the framework. For instance, the quote *"Millions of Americans are using ChatGPT in their daily lives and work."* [80] corresponds to *actual* use information, and represents the *which* entry (ChatGPT), the *who* entry (Americans), and the *when* entry (daily life, work).

Table 2: Composition of our corpus of industry documents.

	AI provider					
	OpenAI	Anthropic	Google	Meta	Microsoft	AI2
Targeted models	GTP-4, GPT-4o	Claude 3, Claude 4	Bard, Gemini	Llama 2, Llama 3	Bing Chat, Copilot	OLMO
Total	16	21	17	21	18	18
Marketing	6	7	6	7	9	6
Policies	0	1	2	1	1	2
Guides	1	2	0	10	4	0
Documentation	6	3	8	3	1	3
Reports	3	8	1	0	3	7

Data analysis. Finally, we analyzed the results of our annotation exercise. To answer RQ2, we conducted an exploratory analysis on top of our dataset of annotated documents, and paid particular attention to identifying informational and methodological trends in terms of information and methods that might be reported and employed. Aided by our knowledge of studies about technology use (Appendix A.1), we further investigated the presence of any notable gaps in the reporting of the companies included in the corpus. To validate our answer to RQ1 (i.e., our framework), we looked for interesting patterns across the dimensions of our framework that demonstrated its utility. We specifically looked for limitations in current reporting practices thrown into greater relief by the different components of our framework. Ultimately, we conducted a thematic analysis over our findings, the results of which we present next.

6 Findings: Insights from Applying the Framework

We first discuss what we can learn by analyzing the study designs employed by the documents and the research questions they address, including notable trends and information gaps. We then discuss a number of limitations with the *reporting* style of the documents, that we identified by analyzing the formulation of use statements, the composition of the study samples, and the description, if any, of methods. Appendix C contains an extensive list of figures and tables that supplement the major findings discussed below. Appendix A.4 contains links to the 111 documents collected in our corpus.[3]

6.1 Study designs and questions span a surprisingly broad space

6.1.1 Documents draw on a surprisingly diverse set of study designs. Applying our framework shows the diversity of methods employed across the corpus. Out of the 178 methods that are mentioned, 41 are axiomatic (26 based on deductive inference and the rest on projections), 71 pertain to empirical research about human-model interactions, and 65 pertain to empirical research about the GenAI systems. Table 3 and Table 4 show which methods are employed to study the interactions and the systems themselves, respectively, with at least eight different methods for interactions and six different ones for systems. Interestingly, around half of the documents in the corpus employ more than one method (see Figure 4): we especially find combinations of empirical methods focused on models and on interactions, as well as combinations of empirical model-focused methods with axiomatic reflections, and to a lesser extent combinations of various types of empirical user studies (e.g., behavioral and self-reports). This diversity of methods employed partially explains the richness of the information reported, which we describe next.

Table 3: Number of times interaction-focused methods are employed in the documents of our corpus.

Observational / experimental	Field / laboratory	Behavioral / self-report	Count
observational	field	behavioral	15
observational	field	self-report	15
observational	field	unstated	10
experimental	lab	behavioral	7
unstated	unstated	unstated	5
experimental	field	behavioral	3
experimental	lab	self-report	3
unstated	field	unstated	3
observational	lab	behavioral	2
experimental	field	self-report	1
experimental	lab	unstated	1
observational	unstated	behavioral	1
experimental	unstated	unstated	1
unstated	lab	unstated	1
observational	lab	self-report	1
unstated	unstated	behavioral	1
observational	unstated	self-report	1
Total			71

Table 4: Number of times model-focused methods are employed in the documents of our corpus.

Observational / experimental	Field / laboratory	Behavioral / self-report	Count
experimental	lab	behavioral	41
experimental	lab	self-report	5
observational	lab	behavioral	5
observational	field	behavioral	4
observational	lab	self-report	2
observational	field	self-report	2
experimental	unstated	unstated	2
experimental	unstated	behavioral	2
experimental	lab	unstated	1
observational	unstated	behavioral	1
Total			65

6.1.2 The corpus covers most research questions. The top-down application of the framework to our collection of documents reveals the types of information that are typically reported and surfaces the richness of the information contained in our corpus. We find that all use questions and all temporalities are covered in our corpus. Moreover, almost all combinations of use questions and temporalities are covered, with the exceptions of the *how projected* and *when projected* (see Figure 5).[4]

[3] In the remainder of the section, we indicate the provenance of the illustrative quotes we use by adding in parentheses the identifier of the document from which they are extracted, as specified in Appendix A.4.

[4] Table 15 in the Appendix offers examples of the kind of information provided for each use question and temporality.

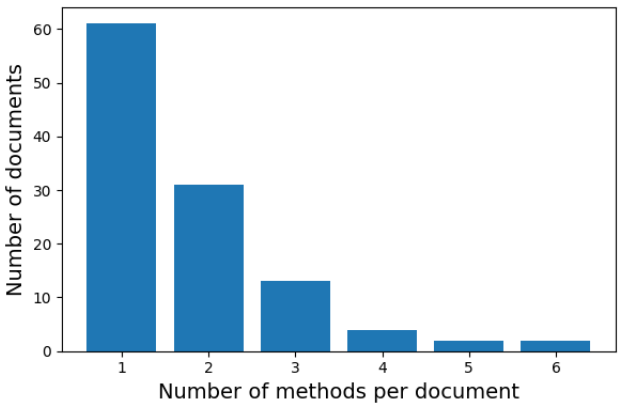


Figure 4: Distribution of the number of methods employed per document.

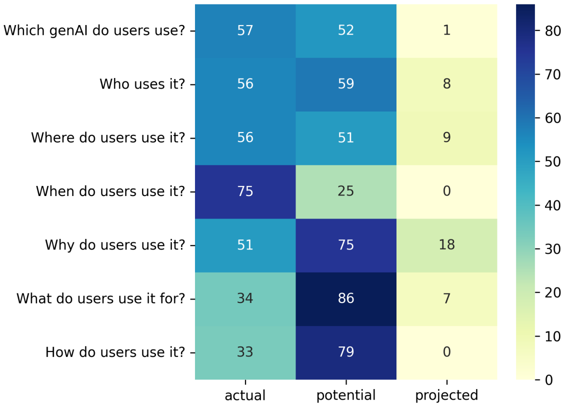


Figure 5: Percentage of documents that provide information about use questions in different temporalities.

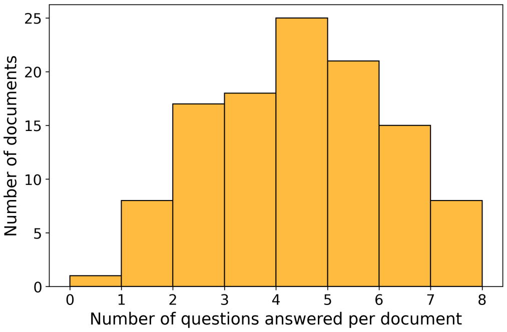


Figure 6: Distribution of the number of use questions answered per document in our corpus.

Because users' use of GenAI systems might differ depending on the company that provides a system and the types of users it targets, we explore patterns in how different companies tackle the research questions. Most companies do report information across all use questions. Furthermore, individual documents tend to report a wide number of types of information, with 36% of the documents providing between four and seven distinct types of information (Figure 6). The diversity of types of information reported may be due to the diversity of methods employed across the corpus. For instance, while model-centered methods may provide extensive information about *what* GenAI can be used *for* due to their focus on benchmarking model capabilities, interaction-focused methods tend to also cover *who* the users are and *why* they use GenAI (see Figure 7).

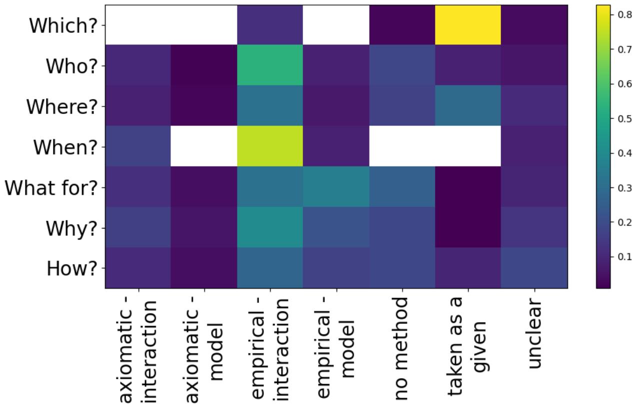


Figure 7: Methods used to answer each use question.

6.1.3 *Many documents provide detailed, multi-level use information.*

The bottom-up analysis of each information category reveals the diversity of units of analysis employed in the documents to answer each use question (see Figure 8).[5] For example, analyzing the *who* across documents reveals that users of the systems are not only described with the expected socio-demographic indicators and professional expertise, but also with a wide range of industry-related detail, such as which professional role the users occupy, in which organization, department, or sector they work, and in what kind of company. Interestingly, within these categories of information, we find further variation. For instance, beyond naming professional roles, some documents describe in more detail what kinds of users can best leverage their systems, as the following quote from AI2 illustrates: *"Given the corpus, ScholarQA will be most useful for researchers in fields with most papers available on arXiv."* (AI2_12) Similarly, exploring the units answered in the *what for* and *why* reveals a wide diversity of information, including actual examples of user interactions; lists of topics discussed; descriptions of the activities that are conducted when interacting with the systems, such as learning, solving a problem, decision making, and information seeking; and the anticipated or observed effects of using GenAI.

It is especially interesting to note the multiple dimensions across which such use questions are answered. For instance, *why* a user uses a system might be justified based on how it transforms the

[5] Table 16 in the Appendix offers a detailed explanation of each unit and accompanying examples.

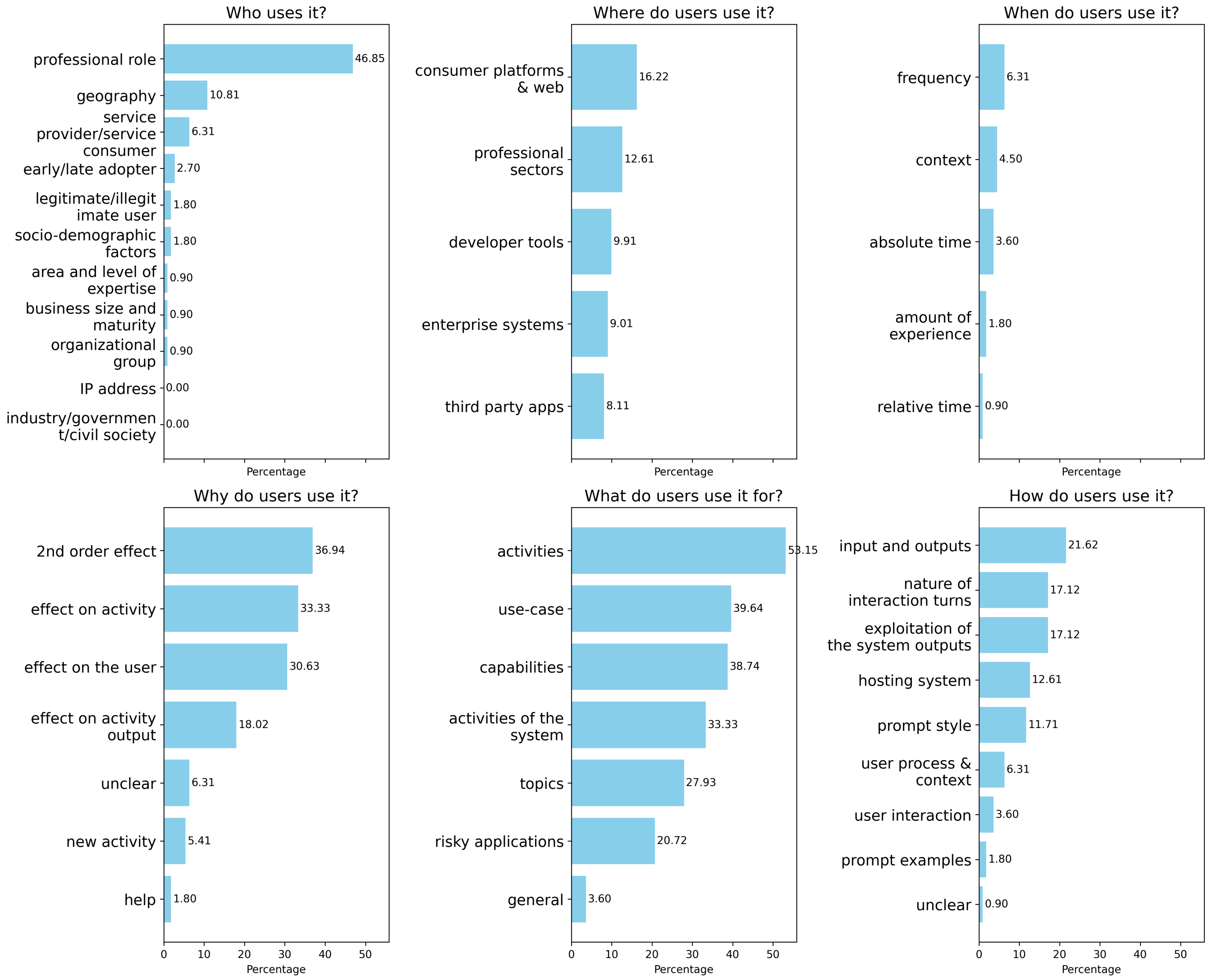


Figure 8: Percentage of documents in our corpus that answer each question using the identified units of analysis. We omit the units related to *which* as they are not commonly reported and did not exhibit much diversity.

activity they conduct (e.g., improving the activity's speed), the activity's output (e.g., increasing the accuracy of the outputs), themselves (e.g., augmenting their skill, saving time), or even the broader business environment (e.g., increasing the productivity of an organization) and third parties (e.g., reallocating time toward customer needs). Notably, while a large amount of documents (39%) answer *what* GenAI can be used *for* by relying on abstract capabilities (e.g., reasoning)—perhaps unsurprising given that such documents include well-known model documentation published by GenAI providers—we find that 33% of documents (typically marketing materials and reports) describe concrete activities that users conduct with the GenAI systems (e.g., summarization) and 53% describe the broader tasks that users aim at achieving with the help of GenAI. Comparatively fewer documents report on the topics present in system inputs and outputs.

6.2 Key research questions remain under-reported and imbalanced over time

6.2.1 "Where" and "when" remain under-reported. Looking at the individual types of information reported, we identify notable categories of missing information—for example, the lack of contextual information about *where* and *when* users use GenAI. While a few units of analysis are employed (with low frequency) in the corpus, other units are entirely absent. For instance, no document reports the type of location where users use GenAI (e.g., whether it is at their home, in an office, outside in the street, etc.) despite the possibility for such contexts to shape interactions. As for the *why*, we find that most of the information presented refers to post-use rationales for why users adopt GenAI, instead of reporting on the reasons that initially led users to try out GenAI. Understanding both users' goals and outcomes might offer a more nuanced understanding of their journeys with GenAI, including what they might have found (dis)satisfying. Prior literature also suggests ways to answer the *how* that were absent from our corpus, such as whether a user uses the system alone or as part of a group [18].

We also find that reported information always pertains to the interaction between a user (or group of users) and one single GenAI system. Yet in practice it might be that a user uses multiple systems for different purposes. Reporting such information could offer insights for researchers in exploring what system properties are actually valuable to users in which contexts. While the absence of such information is perhaps unsurprising given our focus on documents produced by individual companies, it nevertheless means that reporting on one model may be misleadingly presented as reporting that generalizes across all models, and may not be able to tell us about differences in use across models.

6.2.2 Reporting skews toward potential use over actual use. Looking more specifically at the extent to which each temporality is reported also reveals notable gaps (Figure 5). In particular, we find that *potential* information is on average reported more often (61%) than *actual* (51%) or *projected* (6%) information. This difference is even more striking if we compare the amounts of information pertaining to interactions with GenAI (actual: 39%; potential: 80%; projected: 8%). In comparison, more *actual* contextual information is reported than *potential* contextual information. Arguably, it may be important to have more *actual* information as it is what provides the most accurate picture of the current state of GenAI use. Such disparities might be explained by the different methods employed in the documents: only 39% of methods reported are focused on interactions with the GenAI systems, and among these methods, only half of them are observations from the field that provide *actual* information about use, instead of lab observations or experiments (Table 3).

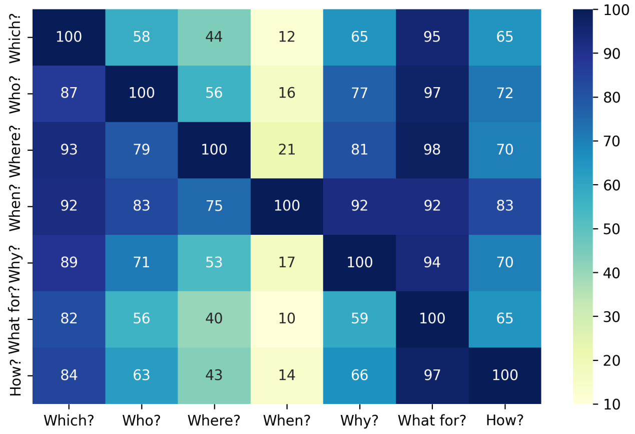


Figure 9: The extent to which use questions co-occur within the same document. The values in each cell represent the percentage of times the two corresponding questions co-occur when the question in the row is answered.

6.2.3 Key use details are rarely reported together in one place. Further investigating which types of questions or units of analysis co-occur within a single document reveals many additional reporting gaps (see Figure 9). For instance, beyond the *when* and *where* which are in any case rarely reported, we find that the interaction information is rarely reported with contextual information (e.g., only 55% of documents that report about the *what for* also report about the *who*). Such information might yield insights about the kinds of activities that different groups of people have identified as useful to conduct with GenAI. Interestingly, the reverse is not the case: when *who* is discussed, in 97% of documents the *what for* is also reported. Delving deeper into the types of documents that report on such information provides an explanation: model documentations that rely on evidence about the model (and not the interaction with the user) of course discuss the *what for* (e.g., by evaluating model capabilities), yet do not refer to any specific users as the purpose of the documents is only to present the model. Similarly, a majority of documents that employ empirical methods focused on interactions solely investigate behavioral data (e.g., with log analysis) without complementing such data with any direct engagement with the user that is observed. Consequently, such documents can only report on interactions without broader contextual information. In a similar spirit, few documents report on the *why* or *how* when they report on the *what for*, which again is an important limitation preventing us from surfacing the most appropriate ways in which users have learned to use GenAI for specific goals. Again, analyzing the methods employed explains this observation: benchmarks only provide *what for* information, and log analyses can provide the *how* but not the *why*. By contrast, methods focused on interactions with self-reports (e.g., interviews, user studies, surveys) typically prompt the users to discuss the *what for* and *why*, but do not leave much space for the users to further describe their fine-grained interactions (*how*).

Investigating the co-occurrence of the units of analysis also reveals nuanced reporting gaps about the same use question. For instance, although holistically understanding the reasons for using a system (*why*) would require knowing the effects this system has on the activity conducted and the user, as well as the second-order effects of interacting with the system, many documents report only on one of the three. For instance, a statement like the following seems to focus on the effect on users without concretely explaining the advantages of using GenAI: *"At its core, AI is helping people scale their ability to think, learn, create, and build. It's scaling human ingenuity itself."* (O_13)—i.e., when compared to "thinking," "learning," and "creating" without such a system, does the system increase the speed of these activities, or their quality? Similar observations can be made about the *who*; few documents (less than 20%) that report about socio-demographic factors also report the organization or industry in which users work. Similarly, for the *what for*, only 28% of documents that report the activities conducted by the system also report on the broader activity the user aims at conducting.

6.3 Reporting patterns vary more by document type than by company

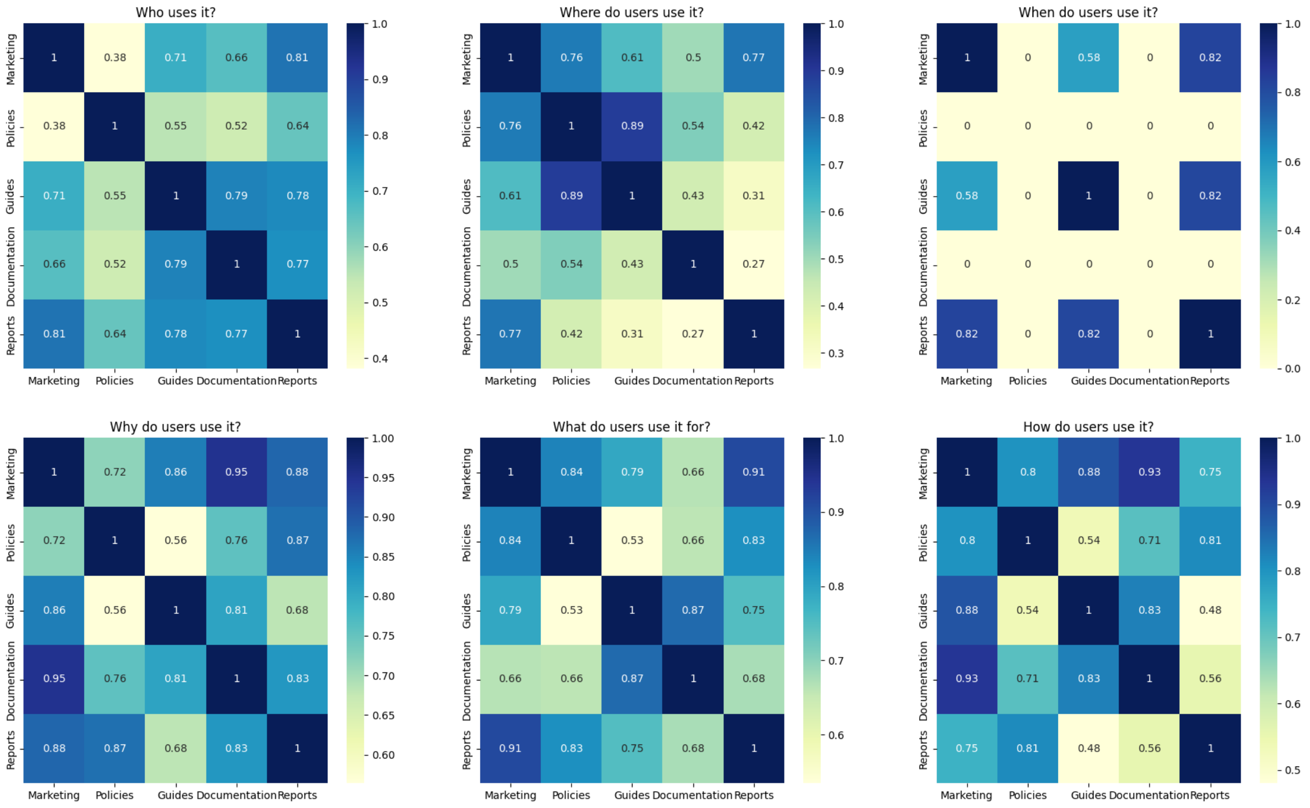


Figure 10: Measures of similarity between the different types of documents in our corpus, computed based on the distributions of the units of analysis in which they report.

6.3.1 *Reporting patterns differ systematically across document types.*

Investigating different document types reveals some expected similarities and less expected differences (see Figure 10). In particular, marketing materials and reports tend to provide information with a similar distribution of information in terms of the units of analysis, while developer guides and model documentation are similar to each other. This makes sense because the target audiences for these two groups of documents tend to be close to each other. Examining document types more closely, we however also observe differences. Overall, marketing materials and reports cover more types of information than the other types of documents, which might be explained by the larger breadth of methods these documents rely on. In addition, the *who* and *why*, the overall *actual* information, and in particular the *actual what for* exhibit significant differences across types of documents (see Figure 11), with marketing materials and reports broadly reporting this information more often than the others. Moreover, such document types are also those that provide greater coverage of questions within the same document. This is particularly interesting because researchers tend to overlook marketing materials to instead focus on longer reports, yet such materials may ultimately be richer than anticipated. A caveat however is that half of the marketing materials present an unclear method (25% ambiguously report on the method, and 36% do not state any method at all), contrary to only 20% of reports and 25% of model documentation, which makes it challenging to assess the generalizability and validity of the information they report. Finally, at the level of the units of analysis, we can also observe meaningful differences. For instance, for the *who*, while the marketing materials focus on professional roles and industry sectors, policies instead describe individual users with their socio-demographic data. This difference is aligned with the intended audiences of these documents: organizations likely to adopt GenAI or individual users who have to abide by the usage policies, respectively.

6.3.2 *Company-level differences exist, but they are not dominant.*

We conduct a similar fine-grained analysis to compare the reporting done by different companies. Overall, we do not identify any striking differences. Still, certain companies (e.g., Microsoft) tend to share more information than others (e.g, Meta, AI2), with disparities especially revolving around the *why*. Moreover, they typically do not actually report on the same types of units of analysis, as illustrated by the high standard deviations in Figure 12. For instance, for the *why*, while OpenAI tends to report extensive information about the first and second-order effects of GenAI, AI2 does not do so but emphasizes cases when a new type of activity is enabled by GenAI. Similarly, for the *how*, while Anthropic and OpenAI tend to describe how a user can make use of a system's outputs, the other

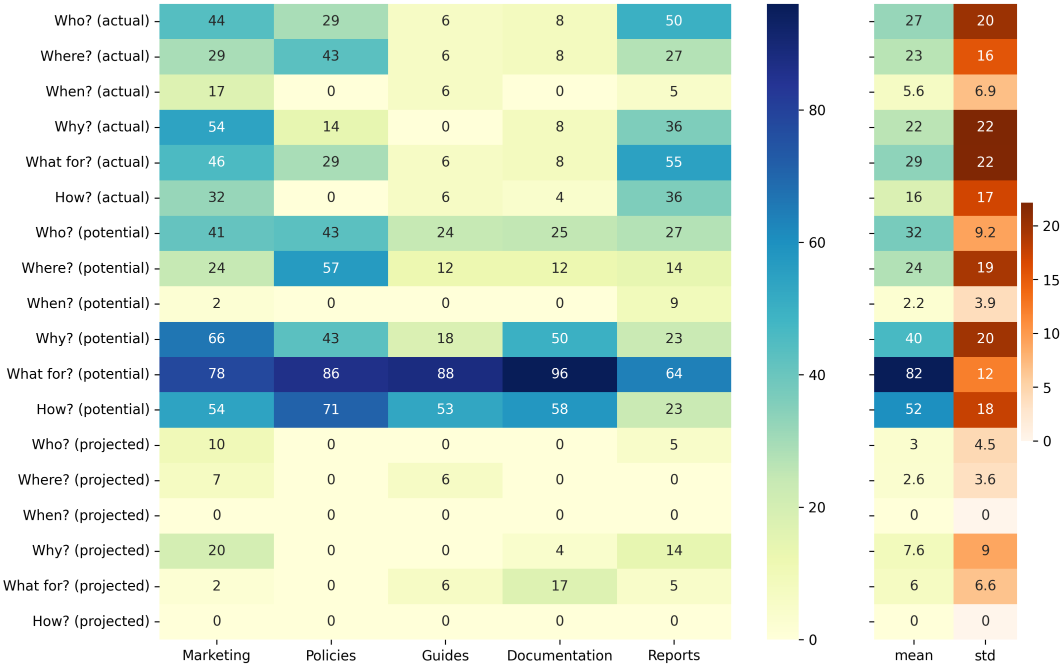


Figure 11: Percentage of times the research questions are answered in different temporalities in each document type in our corpus, including the average and standard deviation of this percentage across all document types.

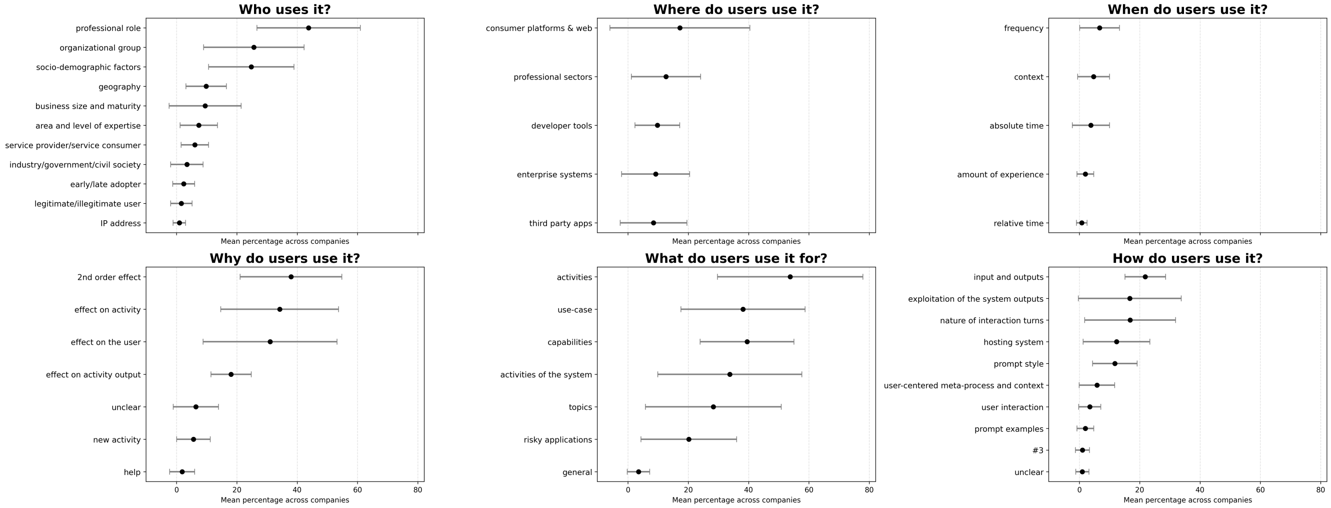


Figure 12: Means and standard deviations computed over the six companies in our corpus, based on the percentage of times each unit of analysis is used in the documents of each company.

companies do so only rarely; instead Microsoft spends considerable space describing the interaction turns between the user and a GenAI system. Delving deeper into the methods employed by each company suggests explanations for the differences we identified. For instance, we see that OpenAI and Microsoft make use of a wide diversity of methods (11 or more methods) which allows them to collect a wide diversity of information types, while Meta only explicitly describes the use of five methods.

6.4 Study populations are often unclear, narrow, or mis-scoped

We now describe the findings revealed by applying the third component of our framework: the study population.

6.4.1 Some claims overgeneralize beyond the underlying sample. By analyzing the units of reporting (Table 5) in comparison to the collected sample of evidence, we identify ways in which statements about use in the documents might actually be inaccurate or incomplete.[6] The *object* of the report might not correspond to the system covered by the evidence. For instance, some documents make statements about the use of one foundation model, although the evidence refers to one specific system built on top of the foundation model, and the other systems built on top of the same foundation model could present different use patterns for various reasons (e.g., the mode of access and the subscription required might be barriers to certain types of users). Some documents make general statements about any GenAI system despite the importance of differentiating GenAI systems when studying use, as this quote illustrates: *"Our analysis reveals clear specialization in how these models are used. Relative to* Sonnet, Opus *sees higher usage for creative and educational work [*what for*] (e.g., 'Produce and perform in film, TV, theater, and music')".* (P_06) In a similar fashion, we find that the *subject* of certain use statements might not be aligned with the evidence underlying these statements. Information pieces might be framed as representative of any user, while the collected data focuses on specific (skewed) distributions of users. For instance, when a document reports the results of an interview study about a provider's own employees working within a provider organization—who might have different habits from lay users—a reader might believe these findings apply to any user. Some documents also claim to report on interactions by specific types of users (e.g., students), yet the ways in which they collect the relevant data cannot ensure they indeed analyze such users; for instance, one document, reporting about students, describes collecting logs and filtering them in if the interaction appeared to refer to a learning situation.

6.4.2 Many documents leave population scope and prevalence unclear. We also find that the scope of the sampled population is often unclear, which makes it impossible to understand the generalizability of the use-related information. Only a few documents quantify the information they provide (absolute number of users or percentages of users within the sampled population) or discuss how likely it is to be applicable to a large user population (e.g., surveys might be more or less representative of a user group); instead, most provide anecdotal information about a few individual users. This is certainly an artifact of the methods employed in the corpus: only 11 documents conduct large-scale log analysis, while a majority conduct interviews and user studies with small sample sizes. In addition, when documents explicitly refer to a group of users, they often do not specify any information about the size of this group and about the individual users. For instance, many marketing documents present case studies about one industrial sector, one organization, or one team within an organization using GenAI, yet do not specify who actually uses GenAI within the sector, organization, or team. The following statement illustrates this lack of information: *"Around 40% of small businesses report using AI today."* (O_11) The units of analysis answering the *who* question further illustrate this point as they primarily cover information aggregated above the individual level.

6.4.3 Reporting undersamples important users, contexts, and time horizons. Investigating the sampled populations more deeply reveals notable gaps in the users and GenAI systems subject to analysis, including the types of interactions between users and systems. For instance, many documents focus on specific types of users (e.g., students, developers), and reveal common purposes of their interaction (*why*, e.g., getting access to learning resources at any time) associated with the activities they use the system for (*what for*, e.g., learning and information seeking), and their various modes of interaction (*how*, e.g., direct prompting for the answers versus iterative prompting). Other types of users and interactions remain under-studied (e.g., the categories of users who might not have extensively adopted GenAI yet, users working in specific sectors such as the military). For instance, while several documents describe the potential (negative) misuses of the systems by adversarial actors (as informed by uplift experiments), none of the documents report on potentially harmful uses by regular users (i.e., unintentional misuses). Similarly, we find that the time span of the user interactions that are studied using direct observations is typically short: the documents report on data they collected over a single user session, or over a short timespan (e.g., a few days). This prevents us from learning about the evolution of adoption or the evolution of the interaction style over time, as one document also emphasizes: *"Due to privacy considerations, we only analyze Claude.ai usage within a single 18-day retention window. Students' usage likely differs across the year as their educational commitments fluctuate."* (P_05)

6.5 Many use statements are hard to validate, easy to misread, or too vague

Finally, we describe limitations arising from how documents are written, which we identify by asking, for each statement, which question is it answering and what methods is it drawing on.

6.5.1 Missing or mismatched methods make many claims hard to validate. We find that 34% of documents do not report any method (the method can be inferred in 14% of the documents), and 18% reported their method ambiguously. We also discovered during the annotation process that it is often difficult to map with certainty a document's described methods to the described use-related information (some use-related questions do not seem to be answerable by the methods stated). Such gaps in the reporting make it impossible for a reader to judge the validity of many statements contained in the documents. Furthermore, when the documents do state their methods, we found a multitude of cases where the evidence collected with these methods does not accurately support the information described in the documents. In particular, we found that 23% of documents rely on axiomatic reflections, which might not be factually correct. For instance, while some collected data (e.g., logs) enabled the authors of a document to identify that a user interacts with the system *at* work or that a user has a particular occupation (*where* or *who*), we found that these documents also misleadingly report that the user interacts with the system *for* work (*what for*)—a fact that is simply assumed based on these other details, not empirically confirmed.

Similarly, we find that users' intentions (*why, what for*) tend to be over-interpreted from information obtained from interaction logs (*how*), as illustrated in the following quote that describes findings

[6] Note that some documents do provide footnotes and appendices that specify these details, which might be missed by a quick reader.

Table 5: Examples of study populations, as reported in the documents of our corpus. We provide examples where the report is or is not aligned with the sample of the data collected.

	Unit of analysis	Example
System	One specific system	“Today, over 98% of advisor teams actively use AI @ Morgan Stanley Assistant" [O_02] [aligned]
	One foundation model	"Overall, we found that developers commonly use Claude for building user interfaces and interactive elements for websites and mobile applications.” [P_08] [misaligned: the scope was solely the first-party API, yet Claude is available via other APIs, too.]
	Generative AI technology	"We identified four patterns by which students interact with AI, each of which were present in our data at approximately equal rates (each 23-29% of conversations): Direct Problem Solving, Direct Output Creation, Collaborative Problem Solving, and Collaborative Output Creation." [P_05] [misaligned: the scope was a specific system]; “Generative AI is developing rapidly and is being driven by research, open collaboration, and product releases that are putting this technology in the hands of people globally.” [F_03] [general statement]
User	Individual user	“Geographically, the majority of data originates from users based in the United States, Russia, and China” [A_16] [factual description of the sampled data]
	Organization	“Startups are the main early adopters of Claude Code, while enterprises lag behind. In a preliminary analysis, we estimated that 33% of conversations on Claude Code served startup-related work, compared to only 13% identified as enterprise-relevant applications.” [P_08] [misaligned: the data does not contain any information about the company sizes of the individual users.]
	Team	“The Product Design *team* uses Claude Code to write comprehensive tests for new features.” [P_13] [aligned]
	Occupational role	"Our analysis reveals highest use for tasks in software engineering roles (e.g., software engineers, data scientists, bioinformatics technicians), professions requiring substantial writing capabilities (e.g., technical writers, copywriters, archivists), and analytical roles (e.g., data scientists)." [P_06] [misaligned: the scope of the study was at the interaction level without providing information about the user occupation.]
	Industrial sector	“We analyze exposure by industry and discover that information processing industries exhibit high exposure, while manufacturing, agriculture, and mining demonstrate lower exposure.” [O_12] [aligned]

from a log analysis: *“The first axis was ‘mode of interaction’. This could involve: (1) Direct conversations* [how]*, where the user is looking to resolve their query as quickly as possible* [why]*, and (2) Collaborative conversations* [how]*, where the user actively seeks to engage in dialogue with the model to achieve their goals* [why]*.”* (P_05) In practice, each *how* may not necessarily imply the associated *why*; for example, a user may not iterate over their prompt even if they are not looking for quick answers from a GenAI system. Finally, we also notice that certain documents describe GenAI adoption by occupation, by type of organization (e.g., start ups versus SMEs), or by country, although the data (logs) does not contain any metadata about the users’ occupation or organization or their nationality. Instead the authors of the document use the topics of the prompts or the geographic location of the user as proxies for occupation, organization size, and citizenship—which might be incorrect.

6.5.2 Quick reads can misinterpret what counts as “use” or “users”. We find a multitude of statements that can be misinterpreted when read quickly (and that were often misinterpreted by the LLM classifier we used—see Appendix A.3.3). Such potentially misleading statements sometimes involve the inclusion criteria for deciding that someone is a user. Many documents state a number of users of a system, without specifying this criteria. For instance, we do not know how many times users have used GenAI, how long ago they started using it, or how frequently they use it, and to what extent the answers to these questions shape who the document counts as a user: *“Chatbots such as GPT-4 and ChatGPT are now serving millions of users”* (AI2_19); *“Today, 28% of employed US adults report using ChatGPT at work, compared to just 8% in 2023.”* (O_11) The lack of answers to the *when* question further reveals that such temporal information is neither specified in the description of the methods nor in the core information contained in the documents.

Potentially misleading statements might also pertain to whether a document reports on users of a GenAI application (i.e., users targeted in a study) or on users of a GenAI model who are themselves developers of a GenAI application using this model. For instance, a quick read of the following quote might give the impression that it describes application users, yet additional context about the company that wrote this document reveals that the document describes deployers of GenAI applications built on top of GenAI models: *“From top tech companies to universities, people and organizations all over the world are using Llama to innovate, drive scientific advances, and unlock new economic opportunities.”* (F_18) Similarly, when a document reports about the extent of adoption based on the number of times the system has been downloaded, or the number of organizations where it has been deployed, or the extent of areas (e.g., countries, products) where it has been released, an inattentive reader might believe the sentence to be about actual use although it is about potential use. For instance, the following quote describes potential users (200 countries), and we do not know whether these potential users actually use GenAI: *“NotebookLM was made available in over 200 countries and territories.”* (G_11) Such statements are particularly hard to interpret when the documents do not state their methods.

Even when methods are stated, statements might still be misleading. For example, 45 documents present experimental data from benchmarking and red-teaming exercises, which were in turn used to qualify and quantify system capabilities, and in some cases to infer potential use activities. The reported information can easily be mistaken with actual *what for*, especially for red-teaming exercises which report on the ways in which the red-teamers used the system.

6.5.3 Vague wording blurs key distinctions. Finally, by sorting the documents’ quotes across the use questions and running into challenges doing so, we find that many statements lack terminological specificity, preventing us from interpreting them with certainty. It can be difficult to pinpoint the use question answered by a document. For instance, we observe that many quotes are stated with clear value judgments about the enormous potential and widespread adoption of the systems: *“There are limitless ways that users can engage with Gemini, and equally limitless ways Gemini can respond.”* (G_13) *“Large multimodal models are used* ubiquitously *today.”* (AI2_4) While this might not be incorrect, these statements are remarkably general and are not accompanied by any evidence, making it exceedingly difficult to annotate which questions they are really answering. We also find a lack of terminological specificity which limits our ability to classify whether a statement is about a system or users’ interactions with a system. For example, the following sentence obviously refers to the *what for* question, yet it does not clarify whether it reports on a user or system activity: *"Both expert engineers and 'vibe coders' lean on GPT models to generate boilerplate code, refactor legacy code, and debug algorithms."* (O_11) The user could prompt the system to conduct the activity for them (the activity would be automated), or decompose this activity in multiple sub-activities and the system would independently perform some of them (e.g., identifying errors in the code, correcting the errors, providing suggestions). Finally, when the type of evidence itself is unknown, we do not always know what temporality is described: *"Teens in most countries around the world will be able to use Bard to easily find inspiration and learn new skills."* (G_17) This quote could describe a potential use or a projected use depending on whether the authors consider the capabilities of the system to be good enough at the present to support teens.

7 Discussion & Future Work

In the previous section, we described applying our framework to a corpus of documents that report on GenAI use. Doing so validated that the framework helps to capture a wide variety of important information contained in our corpus—information that is crucial to appropriately interpret and effectively assess the claims based on this information. It also highlighted that the different components of our framework, the relationship between components, and variation within each component each have important implications for the claims made in reporting, validating the utility of disentangling them. Moreover, it demonstrated that the framework can reveal important informational and methodological trends, make visible categories of information that are under-reported, and draw our attention to information that may be inappropriately or ambiguously reported. These findings have various implications for different audiences.

7.1 Implications for HCI researchers

The findings highlight research gaps that HCI researchers are best positioned to tackle. In particular, while documents employ various units of analysis when reporting on GenAI use, our framework also allows us to pinpoint missing units of analysis that have a long tradition of study in HCI. For instance, there are extensive answers to *how* questions in our corpus, ranging from descriptions of users' prompt formats to nuances about the prompt content that might affect the quality of systems' output, to the sequence of prompts written during the interaction, to the ways users exploit system outputs outside the system to conduct an activity. Nevertheless, we cannot find any information about the challenges users face when interacting with the systems, nor about the extent to which they are satisfied with the interaction, despite this being one of the primary topics of HCI research [10, 48]. Similarly, although user experience researchers have demonstrated and taxonomized the complexities of technologies' contexts of use (e.g., the physical and social context of a user [58]), the contextual information (*where* and *when*) described in our corpus does not cover such types of context. Our findings suggest that the types of research that HCI excels at are often missing from industry reporting, and that reporting would strongly benefit from efforts to include more of such work in industry reports, for example by re-introducing additional methods (e.g., diary studies, focus groups) that might be especially appropriate for uncovering certain categories of use information [10].

These research gaps might in part be explained by the lack of understanding of the needs of these documents' potential audiences. To the best of our knowledge, no work has yet to rigorously investigate who might need to understand how GenAI systems are used, and what kind of information would be useful. Hence, we also recommend that HCI researchers conduct need-finding studies aided by our framework to further characterize what kind of use information is required and to map these needs to corresponding stakeholder groups (e.g., the public, policymakers, system designers).

7.2 Implications for practitioners & policymakers

The findings prompt us to reflect critically about the narratives that are constructed by the collection of documents in our corpus. Reading the corpus yields an impression that GenAI is already widely adopted, by a wide diversity of users, for a wide diversity of purposes, and for a breadth of activities—potentially encouraging uncritical adoption and discouraging critical research or regulation. Nevertheless, our analysis suggests this may be an incomplete picture. For example, our findings also illustrate that statements about use sometimes draw broader conclusions than the evidence they rely on should permit (e.g., based on interviews within a company's users rather than the general public) or describe potential rather than actual use. More generally, even relatively measured descriptions may not offer the context necessary to appropriately nuance (and thus meaningfully understand) them (e.g., criteria for who counts as a system user may be unstated).

Why are documents constructed this way, and what work do they do? These choices might be in response to the providers' obligations (e.g., sensitivity of the use cases, privacy of user interactions) [28, 38, 53] or to practical constraints, as certain data collection methods can be more costly and more challenging to put into practice than others (e.g., in-depth qualitative studies require a more complex set-up than computational analysis of logs that AI providers are uniquely well-positioned to explore) [21]. They might also be strategic for system providers; for instance, marketing materials' primary focus on a few actual, detailed, but non-quantified *who*, *why*, and *what for* may be intended to inspire potential users to adopt the systems. By contrast, research-oriented reports extensively describe the potential *which* and *what for*, which might inform researchers and developers about future development opportunities. And policy-oriented reports' projections of the *who*—increasing numbers of users (and categories of users)—may suggest the inevitability of future GenAI adoption, to which policymakers must respond (perhaps in cooperation with system providers [46, 80]). Future research could investigate whether documents produced by GenAI providers from other regions of the world, GenAI deployers, or external observers (e.g., independent research and policy institutes) exhibit similar reporting patterns and why.

These interpretations bear implications for policy [104]: we believe our systematic exploration of reporting possibilities can help policymakers and responsible AI practitioners make more deliberate and informed choices in crafting transparency obligations and documentation approaches, respectively, particularly in light of the dimensions we find are often ambiguous or unstated altogether. Finally, we encourage GenAI system providers to use our framework as a guiding tool to structure their analyses and reporting of use, in order to produce documents that contribute as meaningfully and productively as possible to public understandings of current and potential GenAI system use.

8 Conclusion

In this paper, we proposed a framework that systematically characterizes how generative AI use might be reported. The framework captures the many different kinds of questions that such reporting might answer, the diverse set of methods that can be employed when seeking to answer these questions, and the varied populations that can serve as the basis for such studies. In so doing, we provide a way to understand the different choices that go into reporting on GenAI, including alternative choices to those commonly made in current reporting. It also helps to highlight when there are mismatches between the claims made in such reporting and the method upon which such claims are based. To demonstrate our framework's analytical utility, we applied it to a corpus of over 100 documents released by large providers of GenAI systems. We found that while these documents present a rich landscape of use information, there remain categories of use information that are systematically under-reported, mismatches between claims and accompanying evidence, as well as of misleading and under-specified statements. Together, these patterns of reporting contribute to a narrative of accelerating and ubiquitous GenAI adoption while obscuring important limitations in the available evidence about actual use. We hope that our framework, annotated corpus, and the results of our own analysis of current reporting practices will foster greater discussion about how GenAI use should be documented, more deliberate and effective reporting, and further inquiry into how narratives about GenAI are produced.

References

[1] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International journal of educational technology in higher education* 21, 1 (2024), 10.

[2] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318* (2025).

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Dennis A Adams, R Ryan Nelson, and Peter A Todd. 1992. Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS quarterly* (1992), 227–247.

[5] Ahmad Al-Dahle. 2024. With 10x growth since 2023, llama is the leading engine of AI Innovation. https://ai.meta.com/blog/llama-usage-doubled-may-through-july-2024/

[6] Elske Ammenwerth, Carola Iller, and Cornelia Mahler. 2006. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. *BMC medical informatics and decision making* 6, 1 (2006), 3.

[7] Anthropic. 2025. *System Card: Claude Opus 4 and Claude Sonnet 4*. Technical Report. Anthropic. https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf

[8] Ruth Appel, Peter McCrory, Alex Tamkin, Miles McCain, Tyler Neylon, and Michael Stern. 2025. Anthropic economic index report: Uneven geographic and enterprise AI adoption. *arXiv preprint arXiv:2511.15080* (2025).

[9] Muneera Bano, Didar Zowghi, Jon Whittle, Liming Zhu, Andrew Reeson, Rob Martin, and Jen Parsons. 2025. A Qualitative Study of User Perception of M365 AI Copilot. *arXiv preprint arXiv:2503.17661* (2025).

[10] Javier A Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2689–2698.

[11] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. 2025. *The International Scientific Report on the Safety of Advanced AI*. Technical Report. Technical Report. Department for Science, Innovation & Technology.

[12] Timothy Besley and Anne Case. 1993. Modeling technology adoption in developing countries. *The American economic review* 83, 2 (1993), 396–402.

[13] Alexander Bick, Adam Blandin, and David J Deming. 2024. *The rapid adoption of generative AI*. Technical Report. National Bureau of Economic Research.

[14] Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, et al. 2025. Advancing science-and evidence-based AI policy. *Science* 389, 6759 (2025), 459–461.

[15] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941* (2023).

[16] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. 2024. Foundation model transparency reports. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 181–195.

[17] Peter Börjesson, Wolmet Barendregt, Eva Eriksson, Olof Torgersson, Liza Arvidsson, and Linda Persson. 2018. The Merits of Situated Evaluation as an Alternative UX Evaluation Method to understand Appropriation. *Interaction Design and Architecture (s) Journal-IxD&A* 37 (2018), 78–98.

[18] Bertram C Bruce, Andee Rubin, et al. 2013. *Electronic quills: A situated evaluation of using computers for writing in classrooms*. Routledge.

[19] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics* 140, 2 (2025), 889–942.

[20] Leonardo Bursztyn, Alex Imas, Rafael Jiménez-Durán, Aaron Leonard, and Christopher Roth. 2025. *Social Dynamics of AI Adoption*. Technical Report. National Bureau of Economic Research.

[21] Paul Cairns and Anna L Cox. 2008. *Research methods for human-computer interaction*. Vol. 10. Cambridge University Press Cambridge.

[22] Aylin Caliskan and Kristian Lum. 2024. Effective AI regulation requires understanding general-purpose AI. (2024).

[23] Flavio Calvino, Daniel Haerle, and Sarah Liu. 2025. *Is generative AI a General Purpose Technology?: Implications for productivity and policy*. Technical Report. OECD Publishing.

[24] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[25] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How people use chatgpt*. Technical Report. National Bureau of Economic Research.

[26] Jingwen Cheng, Kshitish Ghate, Wenyue Hua, William Yang Wang, Hong Shen, and Fei Fang. 2025. Realm: A dataset of real-world LLM use cases. *arXiv preprint arXiv:2503.18792* (2025).

[27] Kasia Chmielinski, Sarah Newman, Chris N Kranzinger, Michael Hind, Jennifer Wortman Vaughan, Margaret Mitchell, Julia Stoyanovich, Angelina McMillan-Major, Emily McReynolds, Kathleen Esfahany, et al. 2024. The CLeAR Documentation Framework for AI Transparency. *Harvard Kennedy School Shorenstein Center Discussion Paper* (2024).

[28] Jennifer Chubb, Darren Reed, and Peter Cowling. 2024. Expert views about missing AI narratives: Is there an AI story crisis? *AI & society* 39, 3 (2024), 1107–1126.

[29] European Commission. 2024. Second Draft General-Purpose AI Code of Practice. https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts

[30] Jessica Dai, Inioluwa Deborah Raji, Benjamin Recht, and Irene Y. Chen. 2025. Aggregated Individual Reporting for Post-Deployment Evaluation. arXiv:2506.18133 [cs.CY] https://arxiv.org/abs/2506.18133

[31] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[32] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 91–104.

[33] Marnik G Dekimpe, Philip M Parker, and Miklos Sarvary. 2000. "Globalization"': Modeling technology adoption timing across countries. *Technological Forecasting*

and Social Change 63, 1 (2000), 25–42.

[34] discussion forum. [n. d.]. R/GPT3 on reddit: What are your most creative uses for Generative AI? https://www.reddit.com/r/GPT3/comments/13jbedw/what_are_your_most_creative_uses_for_generative_ai/

[35] Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2025. The ShareLM Collection and Plugin: Contributing Human-Model Chats for the Benefit of the Community. arXiv:2408.08291 [cs.CL] https://arxiv.org/abs/2408.08291

[36] Pablo Dorta-González, Alexis Jorge López-Puig, María Isabel Dorta-González, and Sara M González-Betancor. 2024. Generative artificial intelligence usage by researchers at work: Effects of gender, career stage, type of workplace, and perceived barriers. *Telematics and Informatics* 94 (2024), 102187.

[37] William Easterly, Robert G King, Ross Levine, and Sergio Rebelo. 1994. Policy, technology adoption, and growth.

[38] Abdallah El Ali, Karthikeya Puttur Venkatraj, Sophie Morosoli, Laurens Naudts, Natali Helberger, and Pablo Cesar. 2024. Transparent AI disclosure obligations: Who, what, when, where, why, how. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[39] Jaran Faret and Stig Hagum. 2024. *Generative AI utilization: How developers utilize Generative AI*. Master's thesis. University of Agder. Available at https://uia.brage.unit.no/uia-xmlui/bitstream/handle/11250/3144275/no.uia:inspera:240853236:50984919.pdf?sequence=5.

[40] Andrew D Foster and Mark R Rosenzweig. 2010. Microeconomics of technology adoption. *Annu. Rev. Econ.* 2, 1 (2010), 395–424.

[41] Kevin Frazier and Andrew W. Reddie. 2025. Why Context, Not Compute, is the Key to AI Governance. *Tech Policy Press* (2025). https://www.techpolicy.press/why-context-not-compute-is-the-key-to-ai-governance/

[42] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[43] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838* (2024).

[44] Bronwyn H Hall and Beethika Khan. 2003. Adoption of new technology.

[45] Kunal Handa, Drew Bent, Alex Tamkin, Miles McCain, Esin Durmus, Michael Stern, Mike Schiraldi, Saffron Huang, Stuart Ritchie, Steven Syverud, Kamya Jagadish, Margaret Vo, Matt Bell, and Deep Ganguli. 2025. *Anthropic Education Report: How University Students Use Claude.* https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude

[46] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. 2025. Which economic tasks are performed with AI? evidence from millions of Claude conversations. *Anthropic* (2025).

[47] Jonathan Hartley, Filip Jolevski, Vitor Melo, and Brendan Moore. 2024. The labor market effects of generative artificial intelligence. *Available at SSRN* (2024).

[48] Marc Hassenzahl and Noam Tractinsky. 2006. User experience-a research agenda. *Behaviour & information technology* 25, 2 (2006), 91–97.

[49] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data protection and privacy* 12, 12 (2020), 1.

[50] Kasper Hornbæk and Morten Hertzum. 2017. Technology acceptance and user experience: A review of the experiential component in HCI. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–30.

[51] Anders Humlum and Emilie Vestergaard. 2024. The Adoption of ChatGPT. *University of Chicago, Becker Friedman Institute for Economics Working Paper* 2024-50 (2024).

[52] Anders Humlum and Emilie Vestergaard. 2025. The unequal adoption of ChatGPT exacerbates existing inequalities among workers. *Proceedings of the National Academy of Sciences* 122, 1 (2025), e2414972121.

[53] Marco Iansiti and Karim R Lakhani. 2020. *Competing in the age of AI: Strategy and leadership when algorithms and networks run the world.* Harvard Business Press.

[54] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

[55] Hwiyeol Jo, Taiwoo Park, Hyunwoo Lee, Nayoung Choi, Changbong Kim, Ohjoon Kwon, Donghyeon Jeon, Eui-Hyeon Lee, Kyoungho Shin, Sun Suk Lim, et al. 2024. Taxonomy and Analysis of Sensitive User Queries in Generative AI Search. *arXiv preprint arXiv:2404.08672* (2024).

[56] Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. Eagle: Ethical dataset given from real interactions. *arXiv preprint arXiv:2402.14258* (2024).

[57] Sahil Koul and Ali Eydgahi. 2017. A systematic review of technology adoption frameworks and their applications. *Journal of technology management & innovation* 12, 4 (2017), 106–113.

[58] Carine Lallemand and Vincent Koenig. 2020. Measuring the contextual dimension of user experience: development of the user experience context scale (UXCS). In *Proceedings of the 11th nordic conference on human-computer interaction: shaping experiences, shaping society*. 1–13.

[59] Lingyao Li, Renkai Ma, Zhaoqian Xue, and Junjie Xiong. 2025. Towards Trustworthy AI: Characterizing User-Reported Risks across LLMs" In the Wild". *arXiv preprint arXiv:2509.08912* (2025).

[60] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic Evaluation of Language Models. *Trans. Mach. Learn. Res.* (2023).

[61] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI transparency in the age of LLMs: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[62] Yi-Chi Liao and Christian Holz. 2025. Redefining Affordance via Computational Rationality. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 1188–1202.

[63] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. *arXiv preprint arXiv:2406.04770* (2024).

[64] Yan Liu and He Wang. 2024. *Who on Earth is using generative AI?*

[65] Mo Adam Mahmood, Laura Hall, and Daniel Leonard Swanberg. 2001. Factors affecting information technology usage: A meta-analysis of the empirical literature. *Journal of organizational computing and electronic commerce* 11, 2 (2001), 107–130.

[66] Miles McCain, Ryn Linthicum, Chloe Lubinski, Alex Tamkin, Saffron Huang, Michael Stern, Kunal Handa, Esin Durmus, Tyler Neylon, Stuart Ritchie, et al. 2025. How people use Claude for support, advice, and companionship.

[67] Mary Meeker, Jay Simons, Daegwon Chae, and Alexander Krey. 2025. Trends – Artificial Intelligence. https://www.bondcap.com/reports/tai

[68] Courtney Miller, Rudrajit Choudhuri, Mara Ulloa, Sankeerti Haniyur, Robert DeLine, Margaret-Anne Storey, Emerson Murphy-Hill, Christian Bird, and Jenna L Butler. 2025. " Maybe We Need Some More Examples:" Individual and Team Drivers of Developer GenAI Tool Use. *arXiv preprint arXiv:2507.21280* (2025).

[69] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438* (2024).

[70] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[71] Jimin Mun, Liwei Jiang, Jenny Liang, Inyoung Chung, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. 2025. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 997–1010.

[72] Jimin Mun, Wei Bin Au Yeong, Wesley Hanwen Deng, Jana Schaich Borg, and Maarten Sap. 2025. Why (not) use AI? Analyzing People's Reasoning and Conditions for AI Acceptability. *arXiv preprint arXiv:2502.07287* (2025).

[73] Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. Prototypical Human-AI Collaboration Behaviors from LLM-Assisted Writing in the Wild. *arXiv preprint arXiv:2505.16023* (2025).

[74] Arvind Narayanan and Sayash Kapoor. 2023. Generative AI companies must publish transparency reports. *Algorithmic Amplification and Society Blog* (2023).

[75] IBM Newsroom. 2024. Data suggests growth in enterprise adoption of AI is due to widespread deployment by early adopters, but barriers keep 40% in the exploration and experimentation phases. *IBM, January* 10 (2024).

[76] Gabriel Nicholas. 2024. *Grounding AI policy: Towards researcher access to ai usage data.* Technical Report. Center for Open Science.

[77] Abiodun Afolayan Ogunyemi, David Lamas, Marta Kristin Lárusdóttir, and Fernando Loizides. 2019. A systematic mapping study of HCI practice research. *International Journal of Human–Computer Interaction* 35, 16 (2019), 1461–1486.

[78] Paulina Oliva, B Kelsey Jack, Samuel Bell, Elizabeth Mettetal, and Christopher Severen. 2020. Technology adoption under uncertainty: Take-up and subsequent investment in Zambia. *Review of Economics and Statistics* 102, 3 (2020), 617–632.

[79] OpenAI. [n. d.]. GPT-4 system card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[80] OpenAI. 2025. Unlocking Economic Opportunity: A First Look at ChatGPT-Powered Productivity. https://cdn.openai.com/global-affairs/be0fe9e0-eb97-43d1-9614-99f2bd948bcc/OpenAI_Productivity-Note_Jul-2025.pdf

[81] Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2375–2393. doi:10.18653/v1/2023.emnlp-main.146

[82] Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker,

Laurance Fauconnet, et al. 2025. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374* (2025).

[83] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).

[84] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608* (2024).

[85] Reva Schwartz, Rumman Chowdhury, Akash Kundu, Heather Frase, Marzieh Fadaee, Tom David, Gabriella Waters, Afaf Taik, Morgan Briggs, Patrick Hall, et al. 2025. Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects. *arXiv preprint arXiv:2505.18893* (2025).

[86] Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, et al. 2025. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Transactions on the Web* 19, 3 (2025), 1–29.

[87] Aijaz A Shaikh and Heikki Karjaluoto. 2015. Making the most of information technology & systems usage: A literature review, framework and future research agenda. *Computers in Human Behavior* 49 (2015), 541–566.

[88] Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-llm grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975* (2025).

[89] Renee Shelby, Fernando Diaz, and Vinodkumar Prabhakaran. 2025. Taxonomy of User Needs and Actions. *arXiv preprint arXiv:2510.06124* (2025).

[90] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[91] Yuya Shibuya, Andrea Hamm, and Teresa Cerratto Pargman. 2022. Mapping HCI research methods for studying social media interaction: A systematic literature review. *Computers in Human Behavior* 129 (2022), 107131.

[92] Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. 2024. Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday* (2024).

[93] Torsten Sløk. 2025. AI Adoption Rate Trending Down for Large Companies. https://www.apolloacademy.com/ai-adoption-rate-trending-down-for-large-companies/. Accessed: 2025-12-04.

[94] Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of business research* 104 (2019), 333–339.

[95] WIRED Staff. 2025. How software engineers actually use AI. https://www.wired.com/story/how-software-engineers-coders-actually-use-ai/

[96] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication.* Cambridge university press.

[97] Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. 2024. Clio: Privacy-preserving insights into real-world AI use. *arXiv preprint arXiv:2412.13678* (2024).

[98] Mary Frances Theofanos, Yee-Yin Choong, and Theodore Jensen. 2024. AI use taxonomy: A human-centered approach. (2024).

[99] Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2703–2707. doi:10.1145/3626772.3657914

[100] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[101] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly* (2012), 157–178.

[102] Angelina Wang, Daniel E Ho, and Sanmi Koyejo. 2025. The inadequacy of offline LLM evaluations: A need to account for personalization in model behavior. *arXiv preprint arXiv:2509.19364* (2025).

[103] Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. 2024. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161* (2024).

[104] Qian Yang, Richmond Y Wong, Steven Jackson, Sabine Junginger, Margaret D Hagan, Thomas Gilbert, and John Zimmerman. 2024. The future of HCI-policy collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[105] Rebecca Yu, Valerie Chen, Ameet Talwalkar, and Hoda Heidari. 2025. Why Do Decision Makers (Not) Use AI? A Cross-Domain Analysis of Factors Impacting AI Adoption. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 8. 2772–2784.

[106] Marc Zao-Sanders. 2025. How People Are Really Using GenAI in 2025. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

[107] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *Proceedings of the International Conference on Learning Representations*.

[108] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=BOfDKxfwt0

A Methodological Details

A.1 Background: technology adoption & use

The broad context in which this work occurs is a wide range of literatures spanning fields such as economics, information systems, and human-computer interaction (HCI) that have examined technology adoption and use; we draw on the lenses and frameworks proposed by these literatures in the development of our framework and in its application.

Economists who study technology use typically center their research around identifying patterns—e.g., country, sectoral, or organizational—in the *adoption* of a technology [e.g., 12, 33, 37, 40, 44, 65, 78], allowing them to study current and projected economic impacts of these technologies. Information systems researchers also study technology adoption, but typically do so at a more granular level, often with a focus on specific individual or organizational users. Researchers have proposed several *technology adoption frameworks* to explain and predict why individuals or organizations adopt technologies [e.g., 4, 6, 31, 50, 57, 87, 100, 101]. These frameworks emphasize that technology adoption is affected by the extent to which the technology fits with the user and their environment, including their internal characteristics (e.g., beliefs, motivations, and goals) as well as their perceptions of the technology's usefulness and ease of use. Other research communities, such as HCI, are interested in the fine-grained interactions of users with technologies [77, 91]. Researchers have shown that technology use is emergent, contingent, and inseparable from its environment [96]; that users' perceptions of a technology's functionalities are not fixed but emerge as the result of a dynamic process of interaction with the technology [62]; and that rather than following pre-determined modes or paths of interaction, users co-create their goals and ways of using technology during the process of use [17]. Together, these insights help us recognize some of the many choices available for reporting on adoption and use—specifically, that reporting can span a wide range of levels of granularity and units—and inform our strong intuition that users' goals, desires, and behaviors may diverge from developers' intentions, thus motivating the various distinct dimensions our framework accounts for.

A.2 Framework development

To scope the documents we collected, we first defined use, then established a list of sources of documents, and then established a list of inclusion and exclusion criteria to refine what we mean by use and what kinds of documents were worth collecting.

Definition of use. We define use of GenAI as observed, self-reported, or anticipated interactions between an end user defined broadly (i.e., an individual user or any grouping of individual users (e.g., within an organization)) and a GenAI system. We include information about actual use (measured empirically), potential use (given system capabilities), and projected use (given trends in development and use over time). We include information about actual (empirically observed), potential and projected (planned/future) use.

Source of use-related documents. Table 6 lists the sources of use-related documents we investigated.

Inclusion criteria.

- Documents in English that provide use-relevant information about GenAI (LLMs or multimodal foundation models).
- Direct evidence (e.g., usage logs/telemetry, field deployments, in-context evaluations).
- User-level studies (surveys, interviews, ethnographies) describing tasks, workflows, or outcomes.
- Organizational, industry, civil-society, or policy reports that quantify or characterize adoption/deployment contexts.
- System/model cards and technical reports that enumerate use cases or evaluation targets linked to capabilities.
- Grey literature (blogs, forum posts, provider materials) when they include substantive descriptions of use or adoption.

Exclusion criteria.

- Non-English documents or items without accessible full text.
- Documents not focused on GenAI (e.g., predictive AI) or focused primarily on autonomous AI agents (an even more recent technology that would have further expanded the scope of our framework).
- Duplicates or superseded versions.

A.3 Framework application

A.3.1 Document collection.

Selecting providers. We selected AI providers using four principles:

1. **Documentation readiness.** We prioritized developers with a track record of publishing system documentation (e.g., model/system cards, safety reports), evidenced by coverage in transparency indices and participation in information-sharing initiatives; this ensured sufficient material to apply our framework [15, 16].
2. **User reach.** We favored providers with large user bases to maximize relevance for studying real-world interactions [64].
3. **Modality and application breadth.** We sought variation in marketed applications and modalities (e.g., text, code, image, multimodal) to capture diverse use-related information.
4. **Organizational heterogeneity.** We included companies that differ in size and release strategies (e.g., open vs. closed) to reflect how organizational practices shape what gets documented and shared.

On this basis, we selected six influential model providers—**AI2, Anthropic, Google, Meta, Microsoft, and OpenAI**—balancing influence, diversity, and organizational stability in line with the FMTI's selection logic [15]. All six are assessed by the May 2024 FMTI, which evaluates model providers, providing a common transparency baseline. Five—Meta, OpenAI, Anthropic, Google, and Microsoft—are part of the Frontier Model Forum[7], signed the White House voluntary AI safety commitments[8] (July 2023; later expanded to additional firms), indicating public governance commitments germane to our analysis. This set also spans release strategies and modalities: API-access, closed-weight services (OpenAI, Anthropic, Google); open-weight releases from large platforms (Meta's Llama

[7] https://www.frontiermodelforum.org/

[8] https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

Table 6: Sources for documents about the use of GenAI systems, and associated types of documents we collected.

Type of actors	Sub-type	Type of documents they produce	Examples
Industry	System provider (e.g., OpenAI), infrastructure provider (e.g., Amazon), service provider (e.g., pecan.ai), deployer (e.g., PwC), consultant (e.g., Gartner)	Technical report and system card, usage policies, research publication, educational resource, developer resource, product resource, blog post	[5, 46, 75, 79]
Academia	Machine learning and Natural Language Processing communities, Human-Computer Interaction community, AI ethics community, others	Research publications about system evaluations, usage datasets or a usage analysis methods, user surveys, reviews of the literature about GenAI	[1, 36, 39, 51, 52, 60, 81, 108]
Policy stakeholders and civil society	Governments, international panels, policy institutes and think tanks, consulting firms, industry bodies (e.g., frontier AI forum), mixed bodies (e.g., Partnership on AI)	Regulations, obligations, standards, bills; policy recommendations; synthesis of prior research; report about ongoing research	[11, 13, 29, 64, 67, 92, 98]
Others	Journalists, users (in specific industries or the broader public)	Journalistic investigations published in (online) newspapers, discussion forums, books, other brief analysis (e.g., Linkedin posts)	[34, 95]

3; Microsoft's Phi-2); and a research nonprofit producing fully open models (AI2's OLMo), enabling diverse comparative analysis across openness and product channels.

Sourcing documents from providers. To collect the documents, we adopted two strategies.

(1) **Model-anchored discovery.** For each provider, we selected one or two (when available) flagship GenAI models released near *Spring 2023* and *Spring 2024*, then ran targeted searches on provider sites and a general web search engine to gather model-specific use information. This resulted in the following models: **AI2 (OLMo), Anthropic (Claude 3, Claude 4), Google (Bard, Gemini), Meta (Llama 2, Llama 3), Microsoft (Bing Chat, Copilot), and OpenAI (GPT-4, GPT-4o)**. To select the two timeframes , we first listed the release timelines of general-purpose GenAI models from these major providers and observed a bimodal distribution: an initial wave of model releases and major updates around Spring 2023, followed by a second wave in Spring 2024. To ensure that our comparisons remained compatible across providers and did not conflate substantially different generations of systems, we focused on documents describing models from these two release clusters.
(2) **Website-centered discovery.** We manually traversed the websites for each provider—main site/blog, research and product pages, developer docs and forums, API portals, linked PDFs (e.g., system cards), and GitHub repos—to capture first-party materials about GenAI use that might concern other models or provide other types of non-model-centered information.

Additionally, we queried an Internet-enabled LLM-based search with "Deep Research" mode to verify that we had not missed any provider-released items. Once we had collected this initial list of documents, we filtered out documents which—with a deeper read—were revealed not to provide any use-related information or repetitive information.

Document filtering. We filtered out documents that were repetitive (i.e., that provided similar kinds of use-related information). For instance, OpenAI and Microsoft present lists of use cases per customer company, and while it is interesting to study a few of these lists, the types of information covered are quickly saturated; Meta has many developer-related pages which serve as demos for different AI use cases and these pages share similar format and types of information. This filtering step was essential in order to collect a tractable number of documents.

A.3.2 Description of our corpus of documents. Ultimately, our corpus contains 111 documents. Each document that we ended up retaining in our corpus falls in one of five categories: a) marketing material such as model announcements or case studies, b) policies (these are typically documents that specify usage policies for the system at hand), c) guides (these are typically documents that provide guidelines to developers to build an effective GenAI system or documents to guide end users in adopting GenAI), d) model documentation such as system cards and model cards, e) reports such as peer-reviewed academic publications, extensive blog posts, or other accounts of investigations (e.g., about the future of work). We established these categories based on the *nature* of the documents we identified that have to do with GenAI use, which is often self-reported and observable based on where the document is published (e.g., as a GitHub page, on arXiv, on a company's website) or how it is titled (many documents specify their nature in their title, such as "technical report" or "system card"). We did not distinguish the documents based on their intended audience as it is often not explicitly stated, and we would have had to engage in more guesswork.

A.3.3 Document annotation. Concretely, we proceeded in four steps to annotate the documents in our corpus.

(1) **Calibrating the codebook.** All authors independently annotated three documents, then met to reconcile differences and refine definitions/examples for consistent application. Following a reflexive, interpretive approach, we did not compute formal inter-rater reliability coefficients; instead, we used these joint-annotation sessions to surface disagreements and converge on a shared understanding of how to apply the codebook, resolving differences by discussion and consensus.
(2) **Extracting quotes manually and annotating them.** Two authors split the corpus and manually extracted, verbatim, use-relevant quotes. They sorted the quotes based on which questions they answered.
(3) **Extracting quotes with an LLM.** We prompted an LLM (with no access to the internet) with the finalized codebook to extract additional candidate quotes from each document. Concretely, we used OpenAI's `gpt-4o` model (accessed in August 2025). We

treated the LLM as an assistive tool rather than as an independent coder: its role was to surface additional candidate passages that might have been missed by manual reading. We therefore did not compute inter-rater reliability between human coders and the LLM, and did not treat LLM outputs as ground truth.

(4) **Merging quotes and verifying validity.** We manually merged human and LLM outputs, removed duplicates and out-of-scope items, resolved conflicts, and performed spot checks for fidelity to the codebook.

A.3.4 *Additional details about the coding scheme for methods.*

For each empirical method discussed in a document, we attempted to annotate all dimensions from our framework wherever possible, marking individual dimensions as *unstated* when the document did not provide sufficient detail to determine that aspect of the research design. Note that a document might report on multiple methods, in which case we annotated all of them. Instead, if a document contained no clearly described method of data generation (empirical or axiomatic), we labeled it simply as: `unstated`.

Specifically, we annotated (Model vs. Interaction), which captures whether the data concern:

- **Model**: model behavior, performance, or properties (e.g., benchmarks, red-teaming, evaluations of model outputs).
- **Interaction**: how people use or interact with models or systems (e.g., user studies, usage logs, surveys, workshops).

Empirical Methods. For empirical methods, each code has the following structure: `Empirical – [Model/Interaction] [Observational/Experimental/Unstated] – [Field/Lab/Unstated] – [Behavioral/Self-Report/Unstated]`

The last three slots capture methodological dimensions wherever the document provided sufficient detail:

- **Observational vs. Experimental**: whether the method passively observes existing behavior or involves an intervention.
- **Field vs. Lab**: whether the data arise from in-the-wild deployments (field) or from more controlled/contrived settings (lab), including online experiments.
- **Behavioral vs. Self-Report**: whether the data are about what people do (behavioral traces, logs, actions) or what people say (surveys, interviews, diary entries, ratings).

When the document did not provide enough detail to confidently assign a value for a given dimension, we coded that dimension as `Unstated` (e.g., `Empirical – Interaction – Observational – Field – Ambiguous`). We required at least one of the three dimensions to be specified; the others were set to `Unstated` if unclear.

We applied a few consistent mappings for common study types:

- **Benchmark Evaluations**: `Empirical – Model – Experimental – Lab – Behavioral`
- **Logs of real-world use of models**: `Empirical – Interaction – Observational – Field – Behavioral`

Axiomatic Methods. For axiomatic methods, each code has the following structure: `Axiomatic – [Deductive Inference / Extrapolation or Projection] – [Model/Interaction]`

Deductive Inference was used when authors derived conclusions from stated assumptions or frameworks without collecting new data. **Extrapolation or Projection** was used when authors speculated about future model performance, application scenarios, or impacts. As with empirical methods, we used **Model** when the reasoning primarily concerned capabilities or properties of models, and **Interaction** when it concerned human–AI interactions.

Annotation Procedure. For each document, we:

(1) Read the full document to identify any descriptions of how data were collected, generated, or analyzed, or how theoretical claims were derived.
(2) Identified up to six distinct methods of data generation per document. Each method was assigned a single code using the scheme above. We did not repeat the same exact method label more than once within a document; instead, multiple passages supporting the same method were grouped under that code.
(3) For each coded method, we recorded one or more verbatim excerpts from the document that justified the code.

Documents for which we could not reliably characterize any method after discussion with team members were coded as `Ambiguous`.

A.3.5 *Limitations of our methodology.*

- **Sampling.** We may have missed relevant documents. Provider sites change, and some items are not public. To reduce this risk, we used multiple sources (official pages, targeted web search, and snowballing from references) and also cross-verified with a search-based LLM with Deep Research. Two authors screened with fixed inclusion rules and deduplicated results. We stopped when new searches yielded no new types.
- **Annotation.** Human coding reflects our backgrounds and can contain errors, especially at this scale. We mitigated this with a calibration round on shared examples and a refined codebook. Two authors then split the corpus for single coding. An LLM produced a second pass. The authors reconciled human and LLM outputs, logged decisions, and ran spot checks.

A.4 Corpus of Industry Documents for Framework Exploitation

Doc ID	Title	Company	Link
O_01	ChatGPT for finance	OpenAI	https://openai.com/solutions/ai-for-finance/
O_02	Morgan Stanley uses AI evals to shape the future of financial services	OpenAI	https://openai.com/index/morgan-stanley/
O_03	Hebbia's deep research automates 90% of finance and legal work, powered by OpenAI	OpenAI	https://openai.com/index/hebbia/
O_04	Get answers. Find inspiration. Be more productive.	OpenAI	https://openai.com/chatgpt/overview/
O_05	Introducing 4o Image Generation	OpenAI	https://openai.com/index/introducing-4o-image-generation/
O_06	AI-based Clinical Decision Support for Primary Care: A Real-World Study	OpenAI	https://cdn.openai.com/pdf/a794887b-5a77-4207-bb62-e52c900463f1/penda_paper.pdf
O_07	GPT-4 System Card	OpenAI	https://cdn.openai.com/papers/gpt-4-system-card.pdf
O_08	GPT-4 Technical Report	OpenAI	https://arxiv.org/pdf/2303.08774
O_09	GPT-4o System Card	OpenAI	https://openai.com/index/gpt-4o-system-card/
O_10	Addendum to GPT-4o System Card: Native image generation	OpenAI	https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf
O_11	Unlocking Economic Opportunity: A First Look at ChatGPT-Powered Productivity	OpenAI	https://cdn.openai.com/global-affairs/be0fe9e0-eb97-43d1-9614-99f2bd948bcc/OpenAI_Productivity-Note_Jul-2025.pdf
O_12	GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models	OpenAI	https://arxiv.org/pdf/2303.10130
O_13	OpenAI's new economic analysis	OpenAI	https://openai.com/global-affairs/new-economic-analysis/
O_14	GPT-4 vs GPT-4o? Which is the better?	OpenAI	https://community.openai.com/t/gpt-4-vs-gpt-4o-which-is-the-better/746991/33
O_15	ChatGPT-4o GPT-4o model used in ChatGPT	OpenAI	https://platform.openai.com/docs/models/chatgpt-4o-latest
O_16	GPT-4 An older high-intelligence GPT model	OpenAI	https://platform.openai.com/docs/models/gpt-4

Table 7: Industry documents analyzed in this paper (OpenAI)

Doc ID	Title	Company	Link
P_01	Introducing Claude 4	Anthropic	https://www.anthropic.com/news/claude-4
P_02	Build with Claude Prompting best practices	Anthropic	https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/claude-4-best-practices
P_03	System Card: Claude Opus 4 & Claude Sonnet 4	Anthropic	https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf
P_04	Anthropic Economic Index: Insights from Claude 3.7 Sonnet	Anthropic	https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7
P_05	Anthropic Education Report: How university students use Claude	Anthropic	https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude
P_06	Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations	Anthropic	https://assets.anthropic.com/m/2e23255f1e84ca97/original/Economic_Tasks_AI_Paper.pdf
P_07	The Anthropic Economic Index	Anthropic	https://www.anthropic.com/news/the-anthropic-economic-index
P_08	Anthropic Economic Index: AI's impact on software development	Anthropic	https://www.anthropic.com/research/impact-software-development
P_09	The Claude 3 Model Family: Opus, Sonnet, Haiku	Anthropic	https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
P_10	Claude 3.7 Sonnet System Card	Anthropic	https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf
P_11	Usage Policy	Anthropic	https://www.anthropic.com/legal/aup
P_12	Claude Opus 4.1	Anthropic	https://www.anthropic.com/news/claude-opus-4-1
P_13	How Anthropic teams use Claude Code	Anthropic	https://www.anthropic.com/news/how-anthropic-teams-use-claude-code
P_14	How people use Claude for support, advice, and companionship	Anthropic	https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship
P_15	Appendix to "How People Use Claude for Support, Advice, and Companionship"	Anthropic	https://www-cdn.anthropic.com/bd374a9430babc8f165af95c0db9799bdaf64900.pdf
P_16	Detecting and countering malicious uses of Claude: March 2025	Anthropic	https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025
P_17	Our approach to understanding and addressing AI harms	Anthropic	https://www.anthropic.com/news/our-approach-to-understanding-and-addressing-ai-harms
P_18	Lyft to bring Claude to more than 40 million riders and over 1 million drivers	Anthropic	https://www.anthropic.com/news/lyft-announcement
P_19	Improve your prompts in the developer console	Anthropic	https://www.anthropic.com/news/prompt-improver
P_20	Salesforce teams up with Anthropic to enhance Einstein capabilities with Claude	Anthropic	https://www.anthropic.com/news/salesforce-partnership
P_21	Claude can now use tools	Anthropic	https://www.anthropic.com/news/tool-use-ga

Table 8: Industry documents analyzed in this paper (Anthropic)

Doc ID	Title	Company	Link
A_01	OLMo : Accelerating the Science of Language Models	AI2	https://arxiv.org/pdf/2402.00838
A_02	2 OLMo 2 Furious	AI2	https://arxiv.org/pdf/2501.00656
A_03	Dolma : an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research	AI2	https://arxiv.org/pdf/2402.00159
A_04	Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models	AI2	https://openaccess.thecvf.com/content/CVPR2025/papers/Deitke_Molmo_and_PixMo_Open_Weights_and_Open_Data_for_State-of-the-Art_CVPR_2025_paper.pdf
A_05	Model Card for OLMo 7B	AI2	https://huggingface.co/allenai/OLMo-7B
A_06	Responsible use guidelines	AI2	https://allenai.org/responsible-use
A_07	Research principles	AI2	https://allenai.org/research-principles
A_08	Open research is the key to unlocking safer AI	AI2	https://allenai.org/blog/open-research-is-the-key-to-unlocking-safer-ai-15d1bac9085d
A_09	OLMo 2: The best fully open language model to date	AI2	https://allenai.org/blog/olmo2
A_10	Introducing Ai2 ScholarQA	AI2	https://allenai.org/blog/ai2-scholarqa
A_11	MolmoAct: An Action Reasoning Model that reasons in 3D space	AI2	https://allenai.org/blog/molmoact
A_12	Galileo: Learning Global & Local Features of Many Remote Sensing Modalities	AI2	https://arxiv.org/pdf/2502.09356
A_13	Contextualized Evaluations: Judging language model responses to underspecified queries	AI2	https://allenai.org/blog/contextualized-evaluations
A_14	MoNaCo: More natural questions for reasoning across dozens of documents	AI2	https://allenai.org/blog/monaco
A_15	Introducing Ai2 Paper Finder	AI2	https://allenai.org/blog/paper-finder
A_16	WILDCHAT: 1M ChatGPT Interaction Logs in the Wild	AI2	https://www.semanticscholar.org/reader/8ba5d42e303b429ad3f160e2eb035635a0b18dbe
A_17	WILDBENCH: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild	AI2	https://www.semanticscholar.org/reader/5d12dfd7278cb8da26f9fd1956cad3c15cea9863
A_18	Broadening the scope of noncompliance: when and how AI models should not comply with user requests	AI2	https://allenai.org/blog/broadening-the-scope-of-noncompliance-when-and-how-ai-models-should-not-comply-with-user-requests-18b028c5b538

Table 9: Industry documents analyzed in this paper (AI2)

Doc ID	Title	Company	Link
M_01	Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web	Microsoft	https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/
M_02	The new Bing & Edge – Learning from our first week	Microsoft	https://blogs.bing.com/search/february-2023/The-new-Bing-Edge-Learning-from-our-first-week
M_03	Building the New Bing	Microsoft	https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing
M_04	Prompts for communicators using the new AI-powered Bing	Microsoft	https://blogs.microsoft.com/blog/2023/03/16/prompts-for-communicators-using-the-new-ai-powered-bing/
M_05	Bing preview experience guide	Microsoft	https://news.microsoft.com/wp-content/uploads/prod/sites/652/2023/02/Experience-Guide.pdf
M_06	The new Bing: Our approach to Responsible AI	Microsoft	https://msblogs.thesourcemediaassets.com/sites/5/2023/04/RAI-for-the-new-Bing-April-2023.pdf
M_07	Sparks of Artificial General Intelligence: Early experiments with GPT-4	Microsoft	https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/
M_08	Introducing Copilot+ PCs	Microsoft	https://blogs.microsoft.com/blog/2024/05/20/introducing-copilot-pcs/
M_09	Introducing GPT-4o: OpenAI's new flagship multimodal model now in preview on Azure	Microsoft	https://azure.microsoft.com/en-us/blog/introducing-gpt-4o-openais-new-flagship-multimodal-model-now-in-preview-on-azure/
M_10	Microsoft 365 Copilot release notes	Microsoft	https://learn.microsoft.com/en-us/copilot/microsoft-365/release-notes?tabs=all#may-2024
M_11	AI at Work Is Here. Now Comes the Hard Part	Microsoft	https://assets-c4akfrf5b4d3f4b7.z01.azurefd.net/assets/2024/05/2024_Work_Trend_Index_Annual_Report_6_7_24_666b2e2fafceb.pdf
M_12	Explore Microsoft Foundry Models	Microsoft	https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/foundry-models-overview
M_13	LLMs Project Guide: Key Considerations	Microsoft	https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/getting-started/llmops-checklist
M_14	AI app templates	Microsoft	https://learn.microsoft.com/en-us/azure/developer/ai/intelligent-app-templates?pivots=dotnet
M_15	The future of banking in the era of AI	Microsoft	https://marketingassets.microsoft.com/gdc/gdcghlne4/original
M_16	IU's Kelley uses Microsoft 365 Copilot to prepare business students for the AI era	Microsoft	https://www.microsoft.com/en/customers/story/24613-indiana-universitys-kelley-school-of-business-microsoft-365-copilot
M_17	Quicker Contract Review (Microsoft 365 Copilot)	Microsoft	https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fadoption.microsoft.com%2Ffiles%2Fcopilot-scenario-library%2FMicrosoft-Copilot-scenario-for-Legal_Quicker-Contract-review-(Microsoft-365-Copilot).pptx&wdOrigin=BROWSELINK
M_18	From idea to impact: Real-world success stories of building intelligent apps with Azure	Microsoft	https://azure.microsoft.com/en-us/blog/from-idea-to-impact-real-world-success-stories-of-building-intelligent-apps-with-azure/

Table 10: Industry documents analyzed in this paper (Microsoft)

Doc ID	Title	Company	Link
G_01	Gemini: A Family of Highly Capable Multimodal Models	Google	https://arxiv.org/pdf/2312.11805
G_02	Gemini 2.5 Deep Think Model Card	Google	https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf
G_03	Gemini 2.5 Flash-Lite Model Card	Google	https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Lite-Model-Card.pdf
G_04	Gemini 2.5 Flash Model Card	Google	https://modelcards.withgoogle.com/assets/documents/gemini-2.5-flash.pdf
G_05	Gemini 2.5 Pro Model Card	Google	https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf
G_06	Gemini API	Google	https://ai.google.dev/gemini-api/docs
G_07	Introducing Gemini: our largest and most capable AI model	Google	https://blog.google/technology/ai/google-gemini-ai/
G_08	Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.	Google	https://arxiv.org/pdf/2507.06261
G_09	Gemini 3	Google	https://deepmind.google/models/gemini/
G_10	PaLM 2 Technical Report	Google	https://arxiv.org/pdf/2305.10403
G_11	Responsible AI Progress Report	Google	https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf
G_12	AI Principles Progress Update 2023	Google	https://ai.google/static/documents/ai-principles-2023-progress-update.pdf
G_13	Policy guidelines for the Gemini app	Google	https://gemini.google/policy-guidelines/
G_14	Try Bard and share your feedback	Google	https://blog.google/technology/ai/try-bard/
G_15	An important next step on our AI journey	Google	https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/
G_16	Bard can now connect to your Google apps and services	Google	https://blog.google/products/gemini/google-bard-new-features-update-sept-2023/
G_17	How we've created a helpful and responsible Bard experience for teens	Google	https://blog.google/products/gemini/google-bard-expansion-teens/

Table 11: Industry documents analyzed in this paper (Google)

Doc ID	Title	Company	Link
F_01	LLama 4 MODEL_CARD	Meta	https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md
F_02	LLama 3 MODEL_CARD	Meta	https://github.com/meta-llama/llama-models/blob/main/models/llama3/MODEL_CARD.md
F_03	Llama Developer Use Guide: AI Protections	Meta	https://www.llama.com/static-resource/developer-use-guide/
F_04	Model Cards & Prompt formats Llama 4	Meta	https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/
F_05	Model Cards & Prompt formats Other Models	Meta	https://www.llama.com/docs/model-cards-and-prompt-formats/other-models/#meta-llama-3
F_06	Build an Automatic Issues Triaging System with Llama	Meta	https://www.llama.com/resources/cookbook/build_a_github_triaging_agent_with_llama/
F_07	Building a Notebook Llama: A Step-by-Step Guide	Meta	https://www.llama.com/resources/cookbook/how-to-build-notebook-llama/
F_08	Build a Text2SQL Use Case with Llama 3	Meta	https://www.llama.com/resources/cookbook/text2sql_natural_language_to_sql_interface/
F_09	WhatsApp and Llama 4 APIs : Build your own multi-modal chatbot	Meta	https://github.com/meta-llama/llama-cookbook/tree/main/end-to-end-use-cases/whatsapp_llama_4_bot
F_10	Model Cards & Prompt formats Llama 4	Meta	https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/
F_11	Book Character Mind Map With Llama4 Maverick	Meta	https://github.com/meta-llama/llama-cookbook/tree/main/end-to-end-use-cases/book-character-mindmap
F_12	Model Cards & Prompt formats Other Models	Meta	https://www.llama.com/docs/model-cards-and-prompt-formats/other-models/#meta-llama-3
F_13	Our open source Llama models are helping to spur economic growth in the US	Meta	https://ai.meta.com/blog/built-with-llama-writesea-fynopsis-srimoyee-mukhopadhyay-united-states-economy/
F_14	How Open Source AI is Evolving Healthcare	Meta	https://about.fb.com/news/2025/01/how-open-source-ai-is-evolving-healthcare/
F_15	Meta at SXSW: How Llama and open source AI is leveling the playing field and enabling innovation	Meta	https://ai.meta.com/blog/sxsw-meta-llama-open-source-innovation/
F_16	How Organizations Are Using Llama to Solve Industry Challenges	Meta	https://about.fb.com/news/2025/01/organizations-using-llama-solve-industry-challenges/
F_17	Meta Llama 3 Acceptable Use Policy	Meta	https://www.llama.com/llama3/use-policy/
F_18	Celebrating 1 Billion Downloads of Llama	Meta	https://about.fb.com/news/2025/03/celebrating-1-billion-downloads-llama/
F_19	RAFT-Chatbot	Meta	https://github.com/meta-llama/llama-cookbook/tree/main/end-to-end-use-cases/RAFT-Chatbot
F_20	customerservice_chatbots	Meta	https://github.com/meta-llama/llama-cookbook/tree/main/end-to-end-use-cases/customerservice_chatbots
F_21	LangChain <> Llama3 Cookbooks	Meta	https://github.com/meta-llama/llama-cookbook/tree/main/3p-integrations/langchain

Table 12: Industry documents analyzed in this paper (Meta)

B Additional Details about our Framework

B.1 Taxonomy of use questions

Table 13 illustrates each of the use questions with examples from our initial corpus.

- Use context and adoption:
  - ***Who** uses GenAI?* This question describes the individuals who use GenAI systems. It can include their socio-demographic information (e.g., age, gender, ethnicity, education, country), occupation, or language. It can also include more granular information, such as their attitudes toward the technology, which is a known factor that drives adoption and appropriation of a technology [17, 87].
  - ***Which** GenAI is used?* This question describes the kinds of GenAI systems that are (or could be) used. It might include information about their providers and any technical specifications in the form of model cards or datasheets [29, 42, 70], such as their capabilities, restrictions, and implementation details (e.g., resources needed). While this question does not strictly characterize "use", it precedes use and constitutes the basis to identify which AI providers are dominating the market in terms of number of users, and which systems (and their functionalities) might be the most successful and useful to current users.
  - ***Where** do users use GenAI?* This question describes the environment in which the individual users interact with the GenAI systems. This might be their physical location (e.g., at home) or their geographic location (e.g., in a specific country). Extending the technology adoption and appropriation frameworks [17, 87], when individuals interact with the system in a professional context, this question can also include the organizational environment in which they work and additional details such as in which sector their organization is, and whether it is supportive of new technologies. It can also include descriptions of the digital space in which the interaction happens (e.g., within a developer tool or consumer platform where GenAI has been integrated).
  - ***When** do users use GenAI?* This question describes the temporal conditions under which users choose to use GenAI. This includes information about the particular moments when the users might employ a system (e.g., on specific days of the week versus on the weekend), as well as the frequency with which they use it (e.g., once a week versus every day versus only once in their life), and the duration (e.g., a few minutes or a few hours).
- Interaction and appropriation:
  - ***What** do users use GenAI **for**?* This question describes the concrete activities in which users engage when interacting with a GenAI system. For instance, they might want to craft a presentation, prepare for an exam, brainstorm about their life.
  - ***Why** do users use GenAI?* This question describes the motivations for users to interact with the system. Users' objectives typically revolve around changing the property of the activity that they typically conduct without GenAI. For instance, they might want to increase the quality of the presentation they craft or craft it faster than they could do by themselves and become more productive. The users might also want to conduct an activity that they could perhaps not have conducted without GenAI. Users might also adopt a technology because of its second-order *effects* on the environment (e.g., impact on the economy [1]). Note that only when these second-order effects are described as the motivation for using a system do they respond to the *why* question.[9]
  - ***How** do users interact with GenAI?* This question asks for descriptions of the ways in which users interact with the system to conduct these activities and achieve their goals. This includes the strategies employed by users to interact with the system, such as the ways in which they formulate prompts (e.g., whether they prompt it only once or refine their prompts along the interaction) and use these prompts (e.g., whether they automate the prompting interaction), the language they use in the prompts, as well as the ways in which they adapt their behavior to fit the technology's affordances in comparison to when they conduct the activity without the technology.

Note that for all of the use questions, the documents might not only answer the question itself, but also describe the factors that led to the type of use reported in the answer. For instance, for the *who* question, multiple documents point to the drivers and blockers that impact a user's decision to become a user and interact with a system. This might be human-centered factors (e.g., the attitude of a user toward new technology that impacts their likelihood to interact with GenAI), or other types of factors such as economic ones (e.g., the economic conditions of a country impact the likelihood of its citizens to use GenAI). This is aligned with the technology adoption frameworks and situated evaluation literature [6, 50, 87] that consider that the fit between specific properties of the users (*who*), systems (*which*), and context (*where, when*) is the primary factor that affects adoption.

B.2 Categories of methods

Table 14 provides more detail about the types of methods we identified all across our corpus.

[9] We do not integrate the ultimate, downstream effects of use explicitly in our taxonomy as a separate question because it has already received extensive coverage in comparison to the use questions (e.g., [90]) and it is arguably not directly telling us about *use*.

Table 13: of questions and units (from the initial corpus of documents).

Question	Unit	Examples
Which?	Application	“Our smart assistant is available across Instagram, WhatsApp, Messenger, and Facebook, as well as via the web.” [5] “To drive the virtual world of its first-of-its-kind AR game Peridot, Niantic integrated Llama, transforming its adorable creatures, called “Dots,” into responsive AR pets that now exhibit smart behaviors to simulate the unpredictable nature of physical animals.” [5]
	Provider	“ChatGPT holds a dominant position among all tools, commanding 82.5% of the total traffic.” [64]
	Functionalities	“The tools were categorized into three types: chatbot, image, and video. It is worth noting that chatbots increasingly include additional functions such as image and video creation, while some video generation tools can also generate images.” [64]
Who?	Socio-demographic	“Youth and male bias is most pronounced for video generation tools, while chatbot users are more highly educated than Google users.” [64]
	Occupation	“Our analysis reveals highest use for tasks in software engineering roles (e.g., software engineers, data scientists, bioinformatics technicians), professions requiring substantial writing capabilities (e.g., technical writers, copywriters, archivists).” [46]
	Attitude toward AI	“We asked how each coder identifies—and the responses were almost evenly split: AI agnostic (26%), AI optimist (35.6%), AI pessimist (38.4%)”. [95]
Where?	Organizations	“*Accenture* is using Llama 3.1 to build a custom LLM for ESG reporting.” [5]
	National user	“ChatGPT shows impressive penetration across continents.” [64]
	Sectoral user	“LLM capabilities can have dual-use potential, meaning that the models can be used for "both *commercial and military* [..] applications".” [79] “This versatility makes it useful, allowing applications in numerous fields such as healthcare, finance, and engineering.” [11]
Why?	Productivity	“Workers in the exposed occupations see a substantial *productivity potential* in ChatGPT, confirming expert predictions: the average worker estimates that ChatGPT can halve working times in about a third of his job tasks.” [51]
	Quality	“They expect to improve *productivity* by 70% and *quality* by 20 – 30%, compared with the company’s existing way of generating Accenture’s annual ESG report.” [5]
	Cost reduction	“Through fine-tuning Llama models, they’ve been able to *cost effectively improve* customer care by better understanding key trends, needs and opportunities to enhance the experience moving forward.” [5]
What for?	Augmented user’s activity	“The same system may be used to *provide medical advice, analyse computer code for vulnerabilities, and generate photos*.” [11] “We observe AI usage for tasks related to *collaborating with colleagues on teaching issues and planning course content*, though not for activities like writing grant proposals or maintaining student records.” [46] [high-level task] “Then you use it for presentations and *get some cool illustrations* for PowerPoint-presentations. It helps very much also in getting some life, color, and something fun.” [39] [specific activity] “GPAI systems can perform a range of tasks useful to biologists, such as *predicting protein folding and aiding protein design.* [..] These models can predict protein structures under various conditions (e.g. in protein–protein complexes), generate useful novel proteins, and perform a wide range of protein-related tasks relevant to drug discovery and design.” [11] [specific activity] “Assistance in article drafting and editing accounts [specific activities] up to 5.1% of the writing assistance [high-level task] requests.” [81]
	Exploited system functionality	“These tools are predominantly accessed via desktop computers during weekdays.” [64] “More than half of employers, per survey results, pay for AI services.” [95] [Mode of access] “Youth and male bias is most pronounced for video generation tools, while chatbot users are more highly educated than Google users.” [64] [Modality]
	Topic	“We conduct a topic distribution analysis on user prompts by applying a clustering algorithm. [..] The majority of questions are related to *coding and software.* [..] Additionally, there is a significant number of unsafe topics. [..] The remaining clusters represent other typical uses, such as general knowledge, *business inquiries, and writing assistance*.” [108]
How?	Interaction strategy	“The vast majority of the Feedback Loop conversations were related to coding and debugging, where the user repeatedly relayed the error they received back to the model.” [46] “57% of interactions show *augmentative patterns* (e.g., back-and-forth iteration on a task) while 43% demonstrate *automation-focused* usage (e.g., performing the task directly).” [46]
	Adaptations	“As users engage with LLMs, they change their behaviors by adopting domain-specific queries and question formats.” [108]
	Language	“Signatories commit to ensuring that model evaluations match the expected usage context of a model. For example, language-based evaluations of multilingual models may focus not only on English but will take into account major European languages (or other languages the model is claimed to support).” [29]
When	Frequency	“3 in 4 coders have tried AI. Of them, the vast majority use it at least once a week. 17% of coders use it all the time.” [95]
	Moment in time	“Generative AI tools are primarily used as productivity tools, as these tools are predominantly accessed via desktop computers during weekdays.” [64]
	Absolute time	“ As shown in the figure, monthly visits surged from zero in November 2022 to 2.5 billion in May 2023.” [64]
	Duration	“Between 1 and 5 percent of all work hours are currently assisted by generative AI.” [13]

Table 14: Examples of methods employed to collect use information (from the initial corpus of documents).

Method	Evidence	Example
Observations	Interaction logs	“The dataset is curated from a larger set of LLM-user interaction data we collected by hosting a free, online LLM service.” [108]
	System meta-data	“By combining website traffic data with Google Trends data, this study also sheds light on early impacts of generative AI tools on individuals’ online activities.” [64]
Self-reports	Survey study	“Using a large-scale survey experiment linked to comprehensive register data in Denmark. Surveying 100,000 workers from 11 exposed occupations, we document [how] ChatGPT is pervasive.” [51]
	Interview study	“We have conducted semi-structured interviews with employees in consultant companies with technical knowledge, like developers, because they have a more natural curiosity towards new technology and can in turn provide better insights that other non-technical professions can provide.” [39]
	Discussion forum	“All Signatories (including those who are providers of open-source models) are encouraged to consider methods such as [..] monitoring public forums and social media for novel patterns of usage).” [29]
Experiments	Evaluation scores	“We would like to see work on more robust evaluations for the risk areas identified and more concrete measurements of the prevalence of such behaviors across different language models.” [79]
	User-studies	“Our survey includes an experiment, informing workers about expert assessments of ChatGPT in their job tasks, and a follow-up to see whether treatment effects persist.” [51]

C Details on the Framework Exploitation

C.1 Coarse-grain use information

C.1.1 What kinds of information do documents report? Table 15 provides examples of the information the documents report with regard to each use question and temporality.

C.1.2 Who answers which questions? Figure 13 reports on the companies that answer the use questions. Figure 14a reports on the types of documents that answer the use questions.

C.1.3 Which temporalities are reported? Figure 14b reports on the temporalities covered in the corpus.

C.2 Units of analysis

C.2.1 What kinds of units of analysis do documents use? Table 16 provides descriptions and concrete examples of the units of analysis identified in the corpus of industry documents.

C.2.2 In which temporalities are the units of analysis reported? Figure 15 and Figure 16 respectively present to what extent each temporality is covered with the units of analysis, and the variance across the three temporalities.

C.2.3 Which companies report on the units of analysis? Figure 18 and Figure 17 respectively present to what extent each company reports on the units of analysis, and the similarities across companies.

C.2.4 Which types of documents use which units of analysis? Figure 19 and Figure 20 respectively present to what extent each type of document reports on the units of analysis, and the variance across types of document.

C.3 Concurrences of use information

C.3.1 To what extent are multiple questions answered together? Figures 21a and 21b respectively show how many use questions are answered within the same document across types of documents and companies.

C.3.2 Which questions are answered together? Figures 22a and 22b respectively show to what extent use questions co-occur across types of documents and companies.

C.3.3 Which units of analysis co-occur? Figure 23 shows to what extent the units of analysis co-occur in each document of the corpus. Table 17 provides a ranked list of these top co-occurences.

C.4 Similarities across groups of documents

Figures 24 and 25 show the extent to which the distributions of reporting of units of analysis are similar respectively across companies and types of documents. Similarities are computed by applying the cosine similarity between two distributions reflecting the frequencies of reporting of the units of analysis (among the documents of a company or of a category of documents).

C.5 Methods employed

C.5.1 Which methods are employed? Figure 26 and Figure 27 respectively show which methods are employed by which company or category of documents. When we report "taken as a given", this means that no method is specified because the use information that is provided did not require any study to be obtained. This is, for instance, the case when a document describes for which types of users (e.g., children) (*who*) or platforms (e.g., Google Doc) (*where*) a GenAI system was built, without studying whether the intended users are indeed making use of the system via the targeted platform. We report "no method" when we could not identify any method associated to the answer to a research question, and we annotate "unclear" when a document reports on a method but it is impossible to tell whether a specific research question was indeed answered using this method.

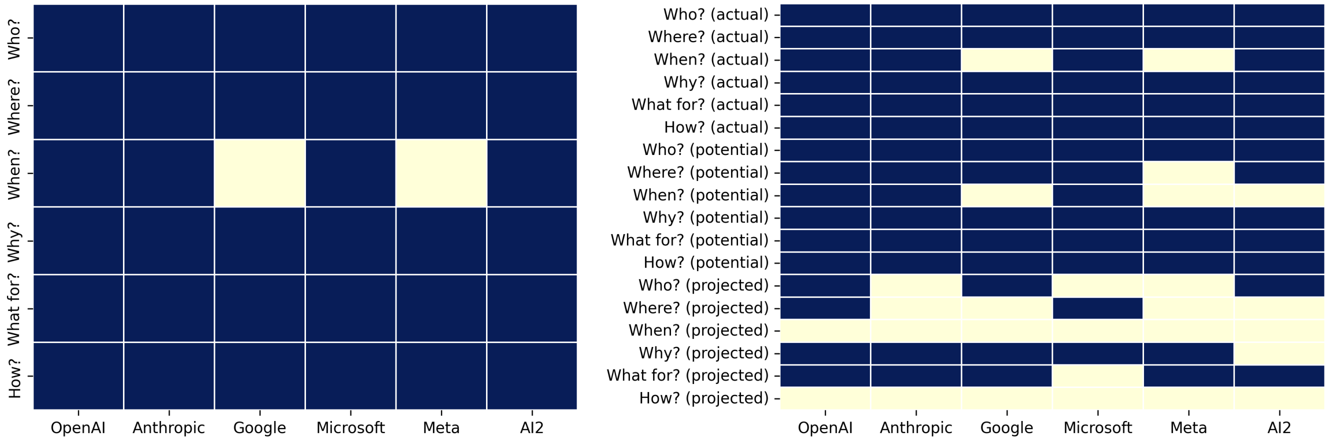
C.5.2 How many methods are employed per document? Figure 28 and Figure 29 breakdown how many methods are used within each document, based on respectively the company that produces the document or the category of documents. Figure 30 displays specifically which methods co-occur.

C.6 Use information in conjunction with the methods employed

Figure 31, Figure 32, and Figure 33 respectively show which methods are used to answer which combinations of use question and temporality, which use questions, which temporality.

Table 15: Examples of quotes from the documents to illustrate the coverage of use questions per temporal dimension. Cells remain empty when no quote could be found in our corpus. (We further specify with ◯ when the information is qualitative and with ∗ when it is quantified.)

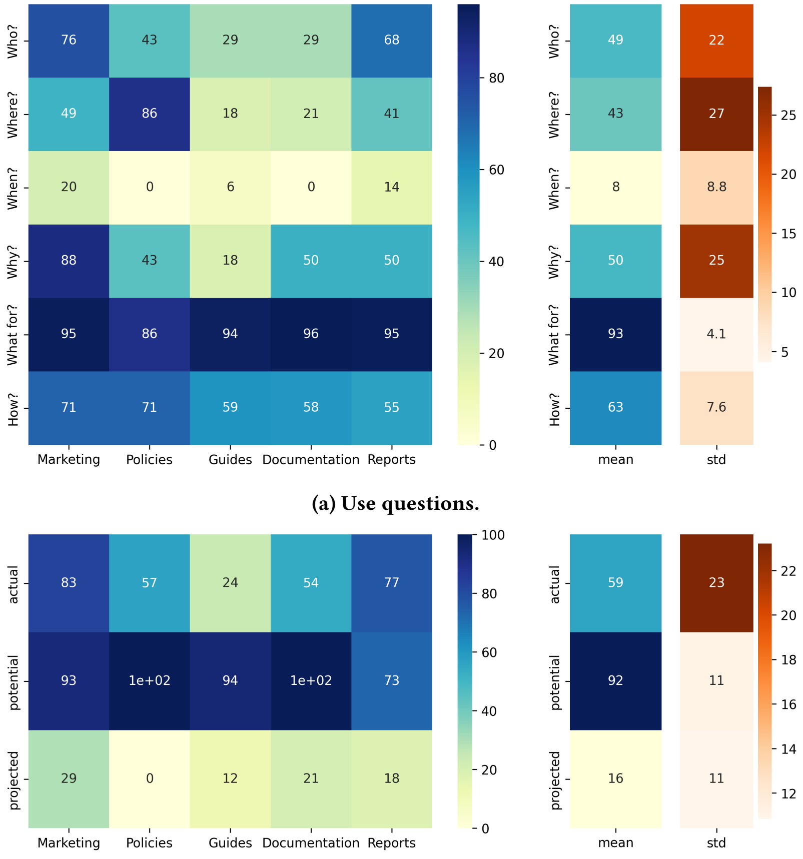
Use question	Temporality		
	Actual	**Potential**	**Projected**
Who?	◯ "Business, Health, and Humanities students show lower adoption rates relative to their enrollment numbers." [P_05] ∗ "Over half a billion people around the world actively use OpenAI’s AI tools." [O_13]	◯ "Students could use Gemini to learn new concepts interactively" [G_01] ∗ -	◯ “We’ll begin rolling out access in the U.S. and U.K. today and expanding over time to more countries and languages.” [G_14] ∗ “The increase in the performance, utility, and flexibility of these models will likely lead to their *ubiquity*, as the value they bring to some pre-existing use cases may outweigh operational costs of deploying the systems.” [F_05]
Where?	◯ "Teams across the company use Claude Code to help new hires and even long-time employees get up to speed on our codebases." [P_01] ∗ "Among OpenAI’s large enterprise customers: 20% are in finance and insurance, 9% are in manufacturing, 6% are in educational service.” [O_11]	◯ “Job Search Genius provides an essential resource for candidates who need help reaching better outcomes in a competitive job market.” [G_04] ∗ -	◯ - ∗ -
When?	◯ “We have been using it daily.” [A_17] ∗ "75% of knowledge workers use AI at work today” [M_11]	◯ "Nextpeer’s Nora platform is transforming mental health support by utilizing Llama’s advanced language capabilities to provide AI-driven companionship 24/7. Nora delivers personalized guidance, multimodal support, and secure, anonymous conversations." [F_04] ∗ -	◯ "One quarter (24%) of US users are between the ages of 18 and 24, and one third (32%) are between ages 25 and 34. This means that many students and workers in the early stages of their careers are becoming AI natives who will bring this expertise to their careers for years to come." [O_03] ∗ -
Why?	◯ "The team began with three targeted goals for their first AI use cases: Faster information retrieval to save advisors hours of document searching. Automation of repetitive tasks like summarizing research reports. Enhanced insights tailored to client needs." [O_02] ∗ -	◯ "We hope this study demonstrates the potential for LLM-based clinical decision support tools to reduce errors in real-world settings" [O_06] ∗ -	◯ "We believe AI can unlock incredible possibilities: more meaningful work, faster progress, and broadly shared prosperity." [O_11]; "As these systems improve, the economic benefits are expected to be significant." [O_11]; "Economists differ in their projections for how AI will impact productivity, but even at the lower end, AI will expand the economic pie." [O_11] ∗ -
What for?	◯ "GPT-4 was used in the following ways: to help us iterate on LaTeX formatting; for text summarization; and as a copyediting tool." [O_07] ∗ "We’re seeing the most use in support of learning and improved written communication: 20% and 18% of all US-based messages, respectively." [O_11]	◯ "we aim to maximize creative freedom by supporting valuable use cases like game development, historical exploration, and education" [O_10] ∗ -	◯ "we observed some uplift in human participant trials on proxy CBRN tasks." [P_03] ∗ "Our findings reveal that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted. We do not make predictions about the development or adoption timeline of such LLMs." [O_12]
How?	◯ "Power users are 68% more likely to frequently experiment with different ways of using AI" [M_11] ∗ “While most conversations have fewer than 10 turns, the distribution exhibits a long tail, with 3.7% of conversations extending beyond 10 turns." [A_16]	◯ "Please use your own judgment to review and validate generated outputs from the Ai2 Tools: no outputs should be accepted at face value and users are expected to evaluate all outputs critically." [A_06] ∗ -	◯ - ∗ -



(a) At the level of use questions.

(b) At the level of use questions combined with temporalities.

Figure 13: Heatmap indicating whether the documents provided by each company present information about each use question. Dark blue indicates the existence of at least one document answering the question.



(a) Use questions.

(b) Temporalities.

Figure 14: Percentages of information reported across types of documents. The plots also indicate the average value and standard deviation across the document types.

Table 16: Units of analysis for each use question.

Question	Unit	Description	Example
Who uses GenAI?	Professional role	Job, company department, or student.	"Clinicians with access to AI Consult" [O_06]; "Nearly all advisor teams now use AI tools like the Assistant daily." [O_02]
	Organizational group	Company team, department.	"The Product Design team uses Claude Code to write comprehensive tests for new features." [P_13]
	Socio-demographic factors	Age, ethnicity, gender.	"American users of ChatGPT also skew younger" [O_11]
	Business size and maturity	Characteristics of the organizations that use GenAI.	"Startups and enterprises around the world are growing with our AI tools." [G_07]
	Geography	Country, region, where the user is.	"Generative AI is developing rapidly and is being driven by research, open collaboration, and product releases that are putting this technology in the hands of people globally." [F_03]
	Area and level of expertise	Background training, years of experience.	"Specifically, we found that information generated by the model is most likely to be useful for individuals and non-state actors who do not have access to formal scientific training." [O_08]
	Service provider / consumer	Relations between GenAI provider and user.	"Investors, bankers, consultants, and lawyers spend countless hours combing through market research [...] Hebbia set out to change that." [O_03]
	Industry / government / civil society	What type of organization the user works in.	"Startups appear to be the primary early adopters of Claude." [P_08]
	Early / late adopter	User adopts the technology early or late after its deployment.	"STEM students are early adopters of AI tools like Claude." [P_05]
	Legitimate / illegitimate user	Adversarial users.	"We are committed to preventing misuse of our Claude models by adversarial actors while maintaining their utility for legitimate users." [P_16]
	IP address	Self-explanatory.	Self-explanatory
Where do users use GenAI?	Enterprise systems	Employer-owned, employee-only tools inside an organization.	"By embedding GPT-4 into their workflows, Morgan Stanley Wealth Management has enhanced how financial advisors access the firm's knowledge base and respond to client needs."[O_02]
	Third-party apps	General workplace software from external vendors.	Google Workspace (Gmail / Slides/ Meet) [G_01]; Microsoft 365 Copilot [M_16]; NotebookLM [G_11]
	Developer tools	Surfaces for coding and data work.	Azure OpenAI Studio [M_09]
	Consumer platforms & web	Public OS, browsers, search, and messaging used by anyone.	iOS/macOS features [M_10]; Bing.com [M_05]; Google Search [G_08]; WhatsApp [G_07].
	Professional sectors	Industry- or occupation-level adoption or usage statistics.	20% are in finance and insurance, 9% are in manufacturing, 6% are in educational services [O_11]
When do users use GenAI?	Frequency	Frequency per day, year, etc.	"Nearly all advisor teams now use AI tools like the Assistant daily." [O_02]
	Absolute time	Moment of the day, week, year, when a user uses GenAI.	first seven days of testing [M_02]
	Amount of experience	How long a user has been using GenAI	"I have been using it for the last two years." [O_14]
	Relative time	Indication that use differs at different moments.	"Students' usage likely differs across the year as their educational commitments fluctuate" [P_05]
	Context	Other fine-grained information about the spaces in which one uses GenAI.	At work [P_13], even when offline [M_08]
What do users use GenAI for?	Capability	Abstract capabilities that the GenAI models exhibit.	"natural language generation, translation, and reasoning." [G_10]
	Use cases	Domains of use, broad applications.	"We aim to maximize creative freedom by supporting valuable use cases like game development, historical exploration, and education." [O_05]
	Activities	Activities conducted by a user, supported by GenAI.	"Writes, brainstorms, edits, and explores ideas with you" [O_04]; "Analyze data and create charts." [O_04]
	Activities of the system	Activity specifically conducted by the system in the context of the broader activity the user conducts.	"Instantly summarize earnings reports, filings, or internal documents." [O_01]
	Topic	Themes of the interaction with GenAI.	"People seek Claude's help for practical, emotional, and existential concerns. Topics and concerns discussed with Claude range from career development and navigating relationships to managing persistent loneliness and exploring existence, consciousness, and meaning." [P_14]
	Risky application	Applications that can cause harm.	"Known risks associated with smaller language models are also present with GPT-4. GPT-4 can generate potentially harmful content, such as advice on planning attacks." [O_07]
	General	Statements that are so general that they do not inform on any specific activity.	"There are limitless ways that users can engage with Gemini, and equally limitless ways Gemini can respond. This is because LLMs are probabilistic, which means they are always producing new and different responses to user inputs." [G_13]
Why do users use GenAI?	Second-order effect	Effect that repeated interactions with GenAI can have beyond changes to the activity and the user.	"Omni models could facilitate both mundane scientific acceleration." [O_09]
	Effect on activity	Changes to the activity process (e.g., sped up).	"The result: an "AI associate" that can perform in seconds what used to take entire teams days or weeks, and deep research that can process any amount of offline data to automate 90% of finance and legal work." [O_03]
	Effect on the user	Effect that the interaction with GenAI has on the user.	"research shows the role that AI and generative models, including GPT-3 and GPT-3.5, can play in augmenting human workers, from upskilling in call centers." [O_08]
	Effect on activity output	Changes to the activity output (e.g., increased accuracy) resulting from using GenAI.	"The automatic issues-triaging system with Llama offers several benefits, including: Improved issue management efficiency. Enhanced accuracy in issue categorization and prioritization. Better insights into the state of the repository." [F_06]
	New activity	Activities that could not have been conducted without GenAI.	"From mobile manipulators to humanoids, MolmoAct enables coherent, grounded, and transparent behavior in robotics and other hardware, opening the door to robust generalization across diverse real-world settings." [A_11]
	Help	General and vague statements about the "helpfulness" of GenAI.	"[It provides] more helpful responses. [G_12]"
How do users interact with GenAI?	Input and outputs	Types of input and output data.	"Mendel's Hypercube is an AI platform that helps health and science organizations draw insights from patient data using a chat-like tool built on open source AI, including Llama." [F_14]
	Exploitation of the system outputs	Ways in which one makes use of the outputs of the GenAI system.	"You cannot and must not use any language model for this purpose in a generalised way, built into any application used by other people, unless each user is supervised by a health professional." [O_14]
	Nature of interaction turns (description of the content)	Types of prompts inputted at each turn within an interaction session.	"We also analyze how AI is being used for tasks, finding 57% of usage suggests augmentation of human capabilities (e.g., learning or iterating on an output) while 43% suggests automation (e.g., fulfilling a request with minimal human involvement)." [P_06]
	Hosting system, e.g., web application	Type of system on which GenAI can be interacted with.	"It contains a notebook which shows a complete example of how to build a Meta Llama 3 chatbot hosted on your browser." [F_07]
	Prompt style	Description of the type of prompt employed.	"We want to ensure that the questions are general enough to be used for web search engine queries and are related to Llama." [F_04]
	User-centered meta-process and context	Ways in which the user learns to interact with the system.	"Power users are 68% more likely to frequently experiment with different ways of using AI." [M_11]
	User interaction	Format of the user interaction.	"On average, each conversation includes 2.52 user-chatbot interaction rounds (turns)." [A_16]
	Sharing of actual prompts	Actual prompt a user has inputted into GenAI systems.	"Transform an input math problem into a step-by-step solutions." [P_02]

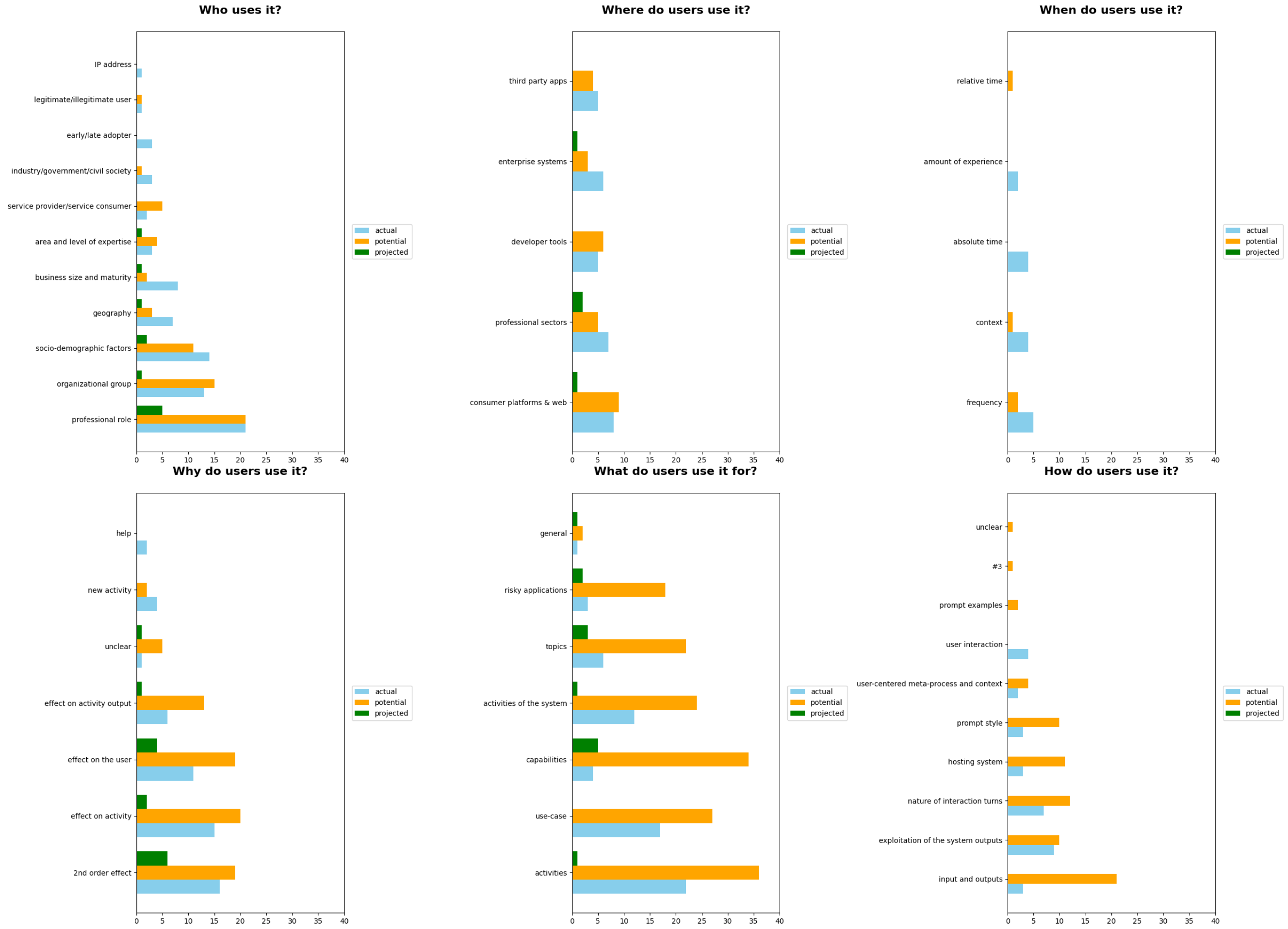


Figure 15: Percentage of documents of each document category that provides information about each unit of analysis we identified.

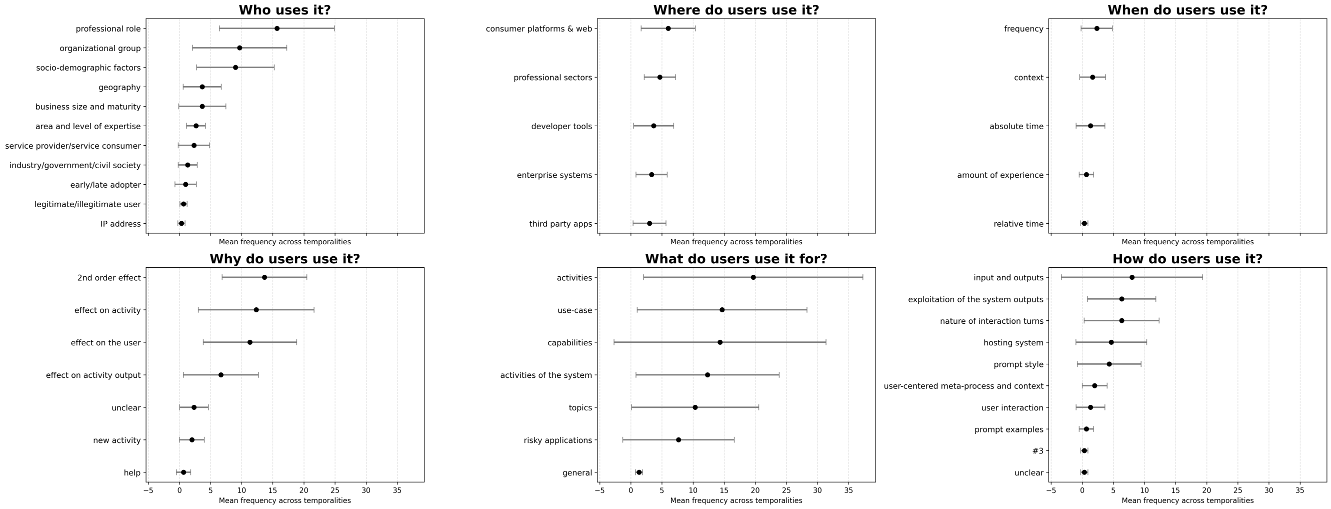


Figure 16: Mean and variance of the reporting of the units of analysis of each use question across temporalities.

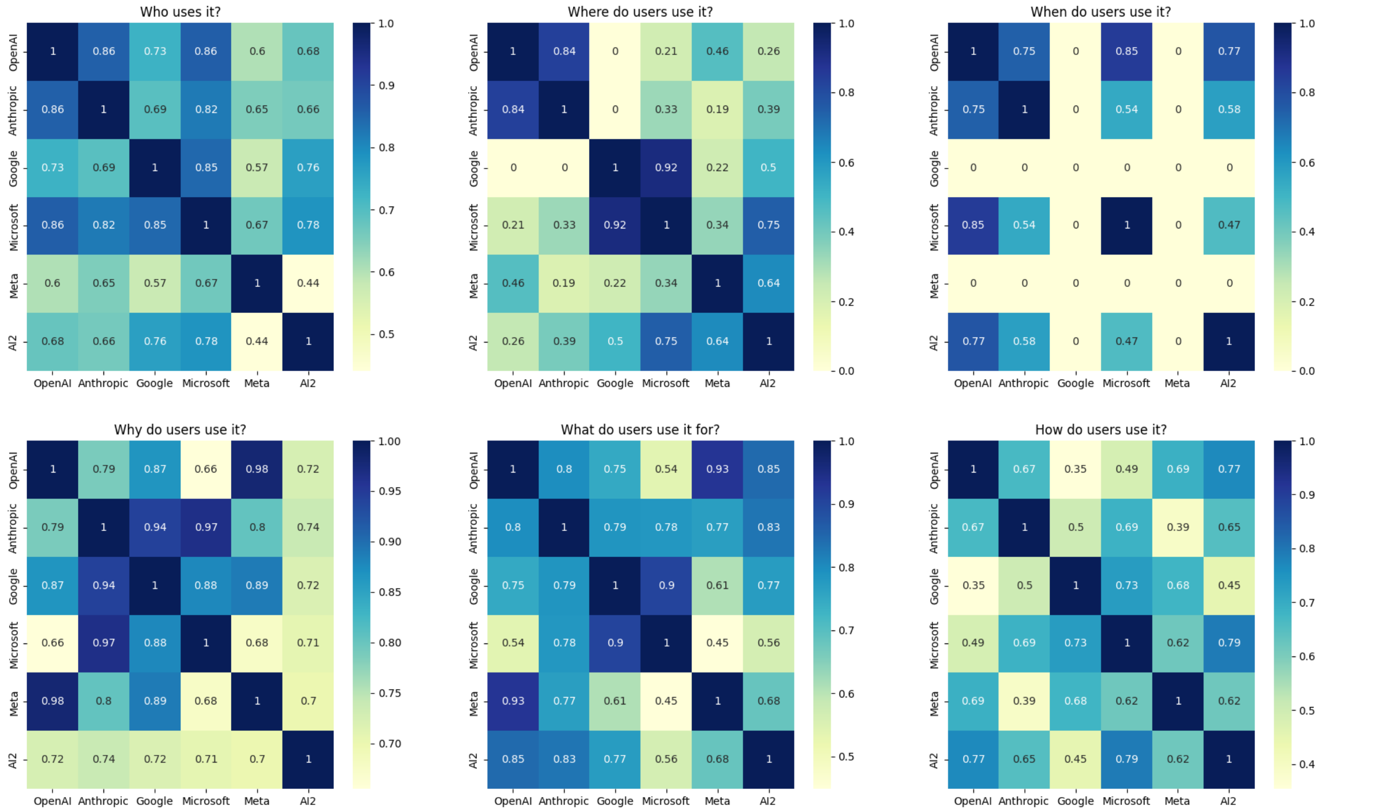


Figure 17: Matrices presenting the similarities of reporting of the units of analysis of each use question across companies.

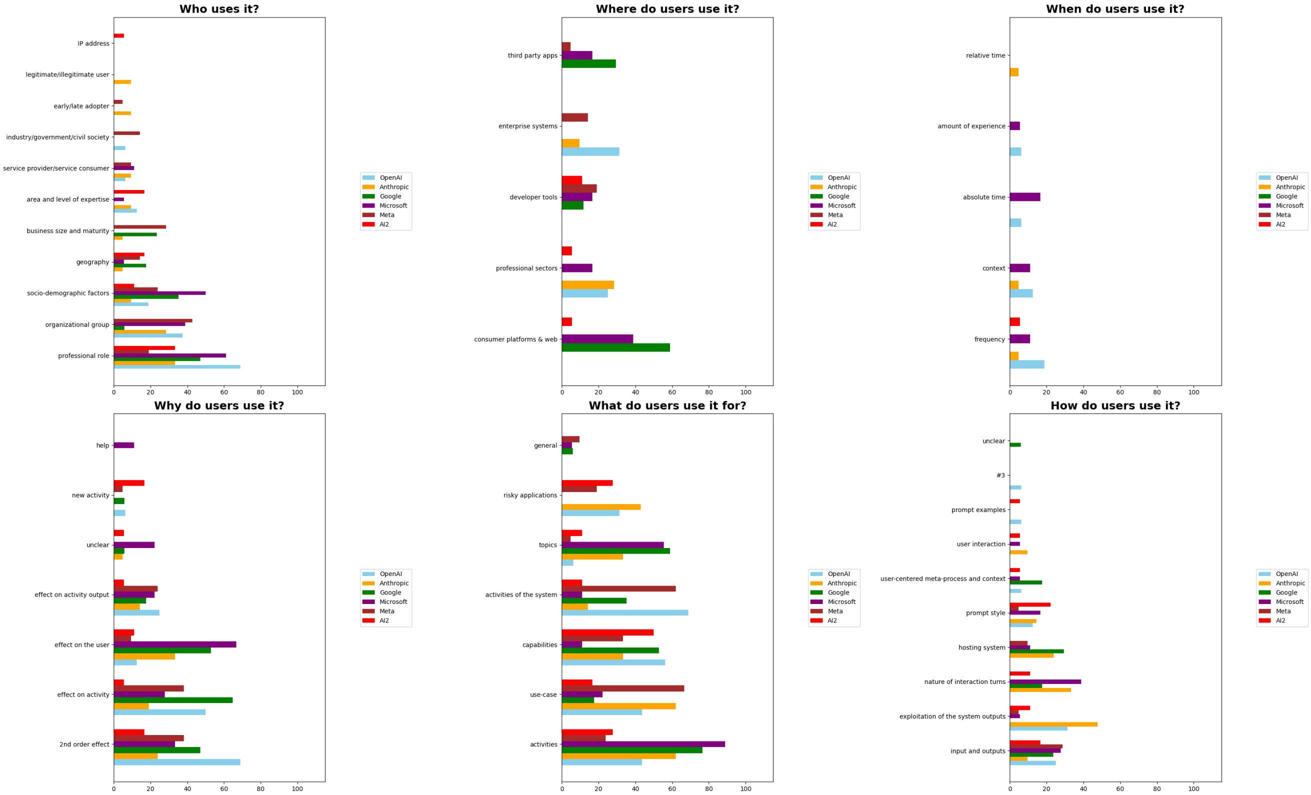


Figure 18: Percentage of documents of each company that provides information about each unit of analysis we identified.

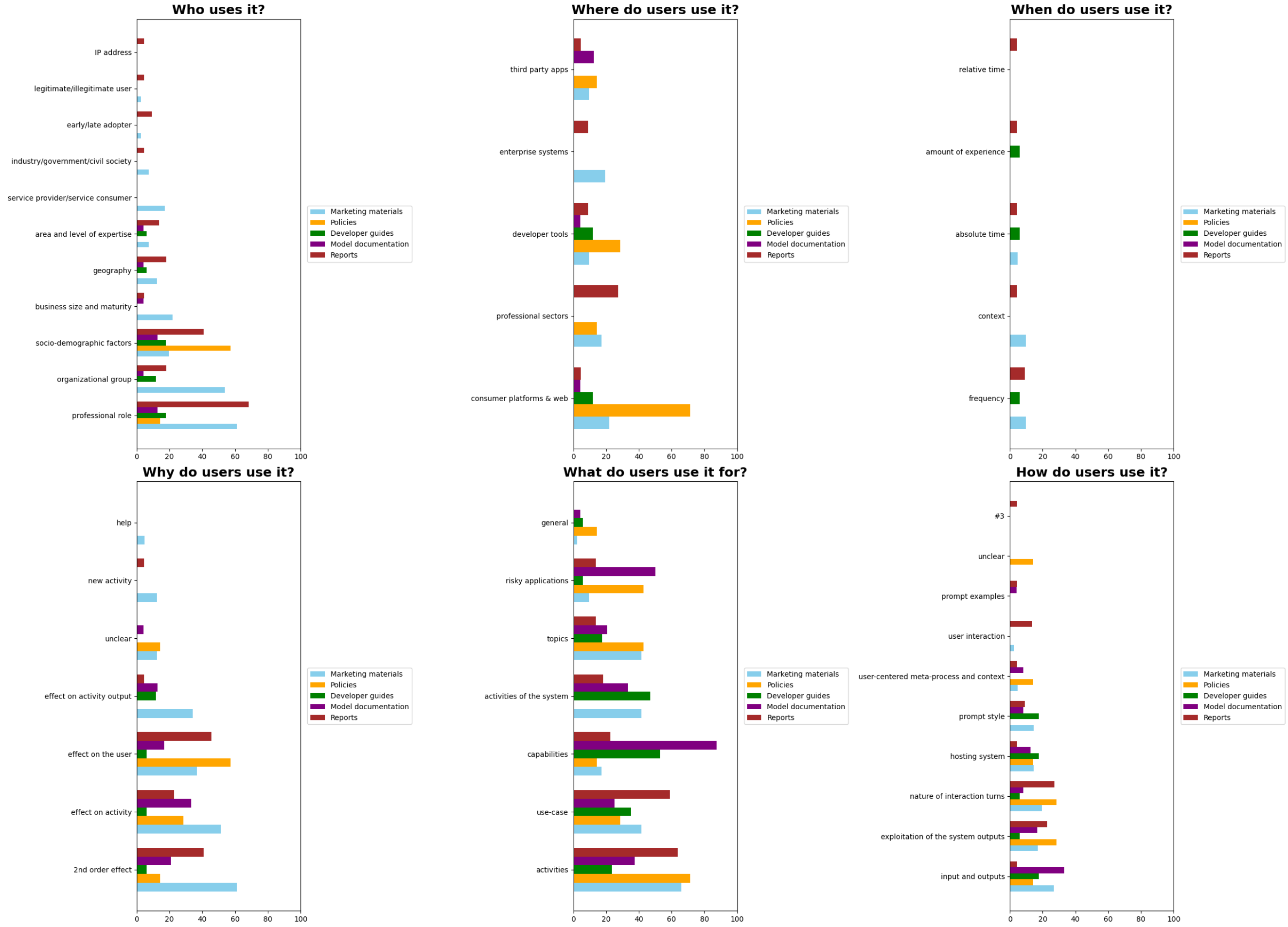


Figure 19: Percentage of documents of each document category that provide information about each unit of analysis we identified.

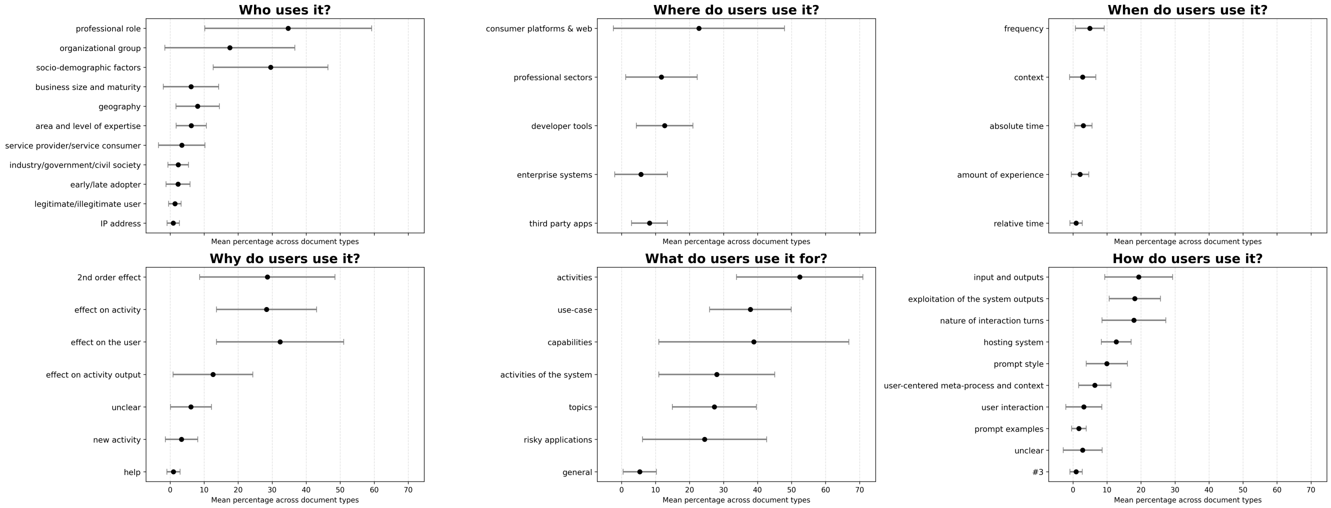
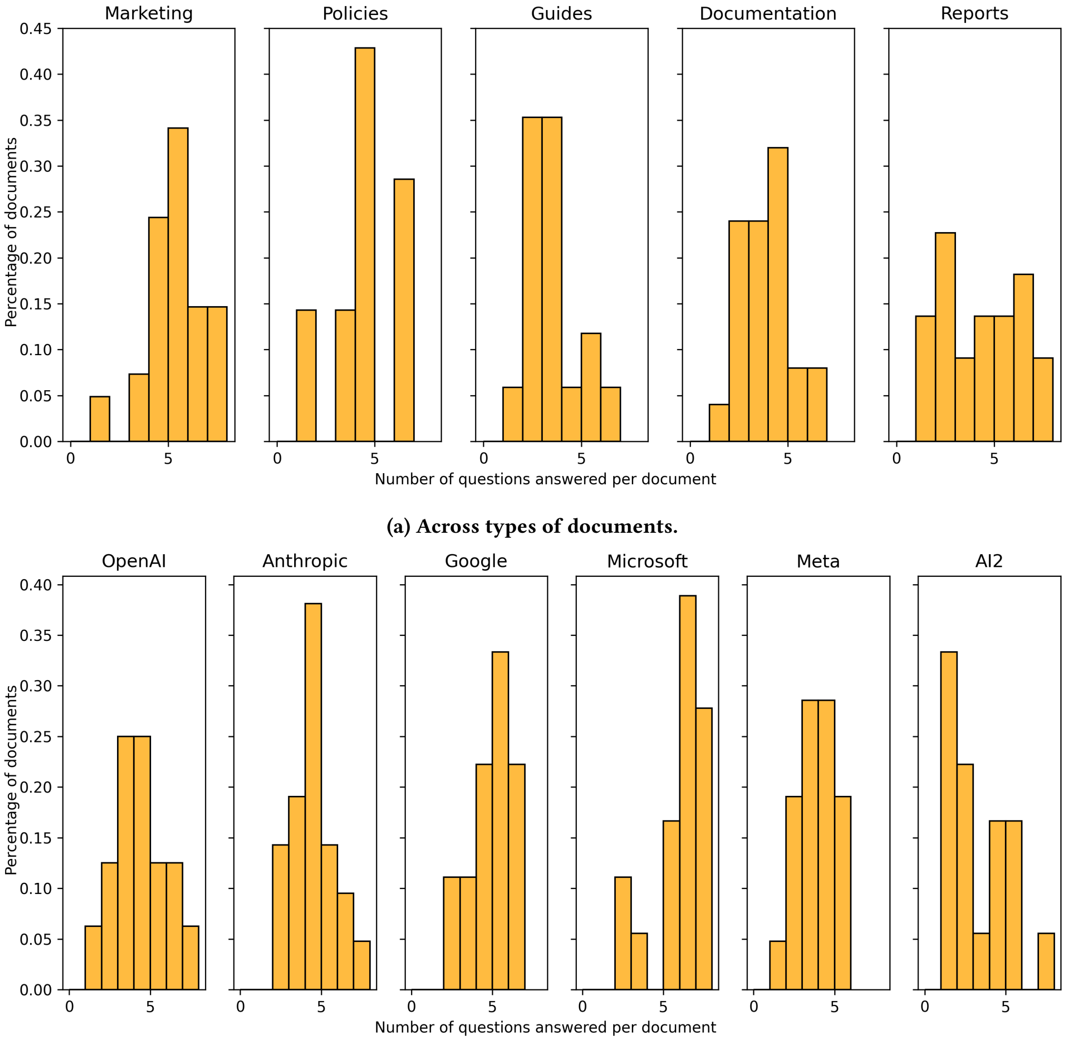


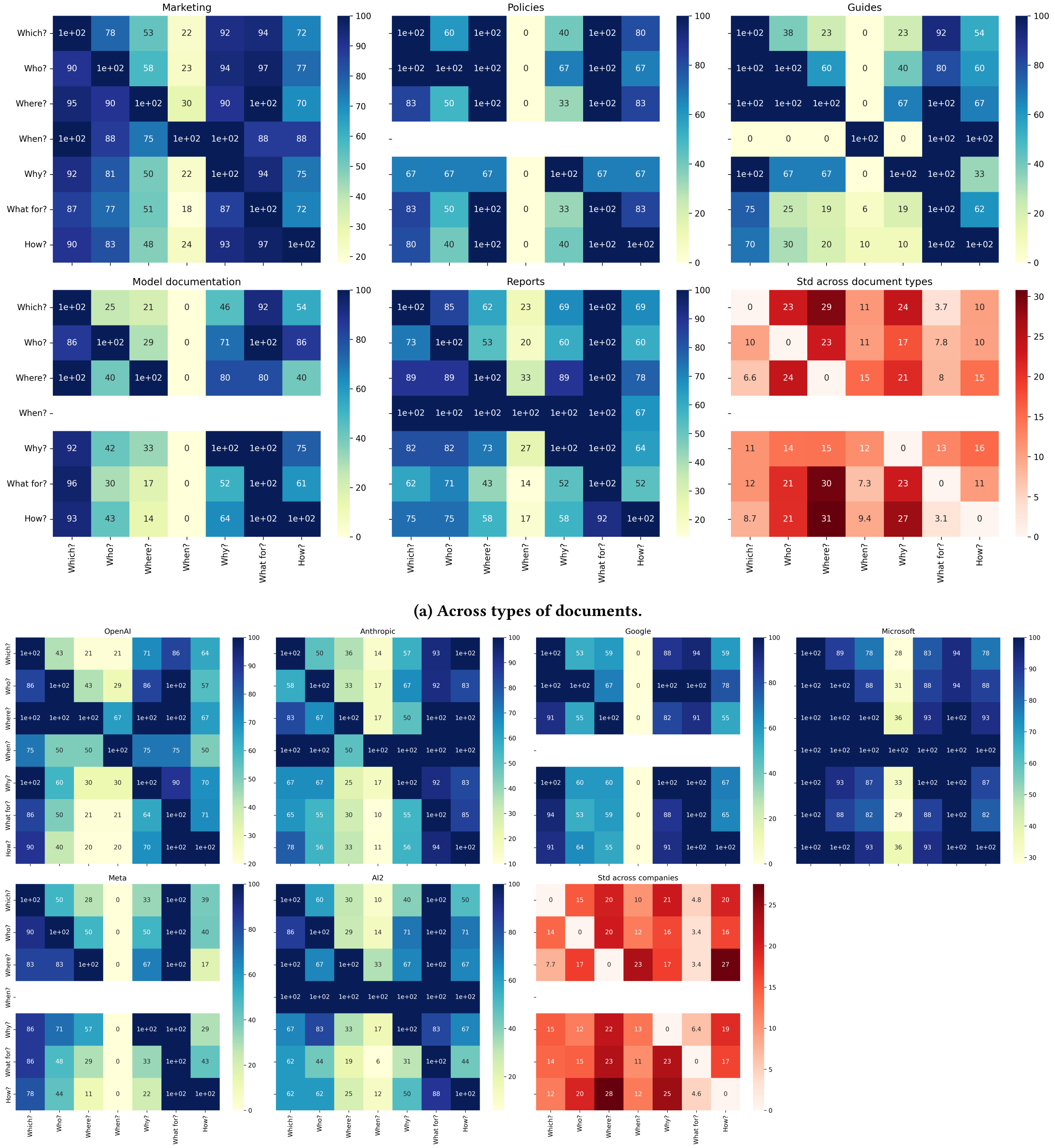
Figure 20: Mean and variance of the reporting of the units of analysis of each use question across document types.



(a) Across types of documents.

(b) Across companies.

Figure 21: Distribution of the number of questions (up to 7) answered within the same document.



(a) Across types of documents.

(b) Across companies.

Figure 22: Percentage of times a question (in the column) is answered together with the question in the row, among all the times the question in the row is answered.

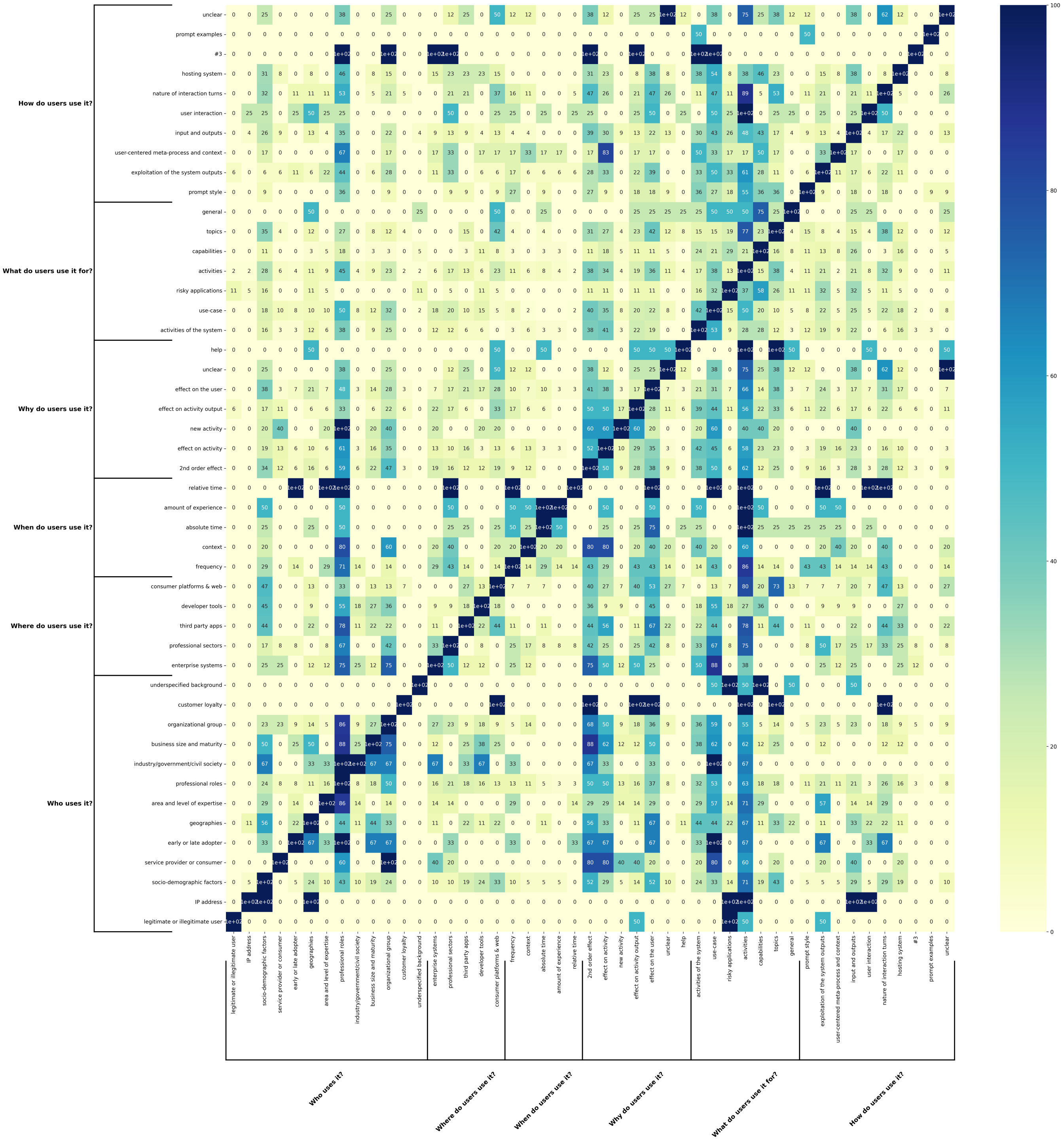


Figure 23: Matrix showing to what extent the units of analysis co-occur within the same document. The values in each cell represent the percentage of times the two corresponding units co-occur among the times that the unit in the row is answered.

Table 17: Co-occurrences ranked according to the fraction of documents in which they appear.

IP address	geographies	1.00
user interaction	activities	1.00
IP address	activities	1.00
early or late adopter	use-case	1.00
new activity	professional roles	1.00
help	activities	1.00
help	topics	1.00
IP address	socio-demographic factors	1.00
customer loyalty	2nd order effect	1.00
customer loyalty	effect on activity output	1.00
legitimate or illegitimate user	risky applications	1.00
customer loyalty	effect on the user	1.00
customer loyalty	activities	1.00
customer loyalty	topics	1.00
customer loyalty	nature of interaction turns	1.00
IP address	user interaction	1.00
underspecified background	risky applications	1.00
IP address	input and outputs	1.00
underspecified background	capabilities	1.00
IP address	risky applications	1.00
service provider or consumer	organizational group	1.00
industry/government/civil society	professional roles	1.00
industry/government/civil society	use-case	1.00
early or late adopter	professional roles	1.00
nature of interaction turns	activities	0.89
business size and maturity	professional roles	0.88
business size and maturity	2nd order effect	0.88
organizational group	professional roles	0.86
area and level of expertise	professional roles	0.86
user-centered meta-process and context	effect on activity	0.83
service provider or consumer	2nd order effect	0.80
service provider or consumer	use-case	0.80
service provider or consumer	effect on activity	0.80
topics	activities	0.77
general	capabilities	0.75
business size and maturity	organizational group	0.75
area and level of expertise	activities	0.71
socio-demographic factors	activities	0.71
organizational group	2nd order effect	0.68
user-centered meta-process and context	professional roles	0.67
early or late adopter	exploitation of the system outputs	0.67
early or late adopter	activities	0.67
early or late adopter	geographies	0.67
geographies	effect on the user	0.67
early or late adopter	2nd order effect	0.67
geographies	activities	0.67
early or late adopter	business size and maturity	0.67
early or late adopter	effect on activity	0.67
early or late adopter	organizational group	0.67
early or late adopter	nature of interaction turns	0.67
early or late adopter	effect on the user	0.67
industry/government/civil society	2nd order effect	0.67
industry/government/civil society	activities	0.67
industry/government/civil society	socio-demographic factors	0.67
industry/government/civil society	organizational group	0.67
industry/government/civil society	business size and maturity	0.67
effect on the user	activities	0.66
professional roles	activities	0.63
2nd order effect	activities	0.62
business size and maturity	activities	0.62
business size and maturity	use-case	0.62
business size and maturity	effect on activity	0.62
exploitation of the system outputs	activities	0.61
effect on activity	professional roles	0.61
new activity	effect on activity	0.60
new activity	effect on activity output	0.60

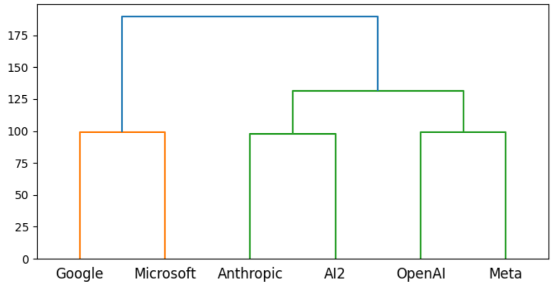


Figure 24: Dendogram summarizing the order of similarities across companies.

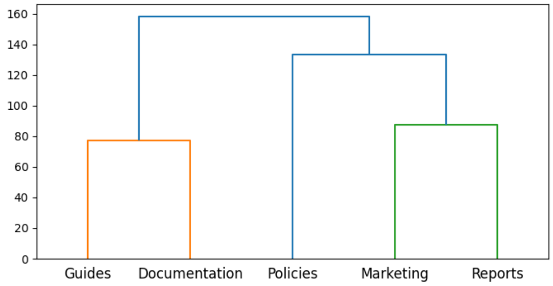


Figure 25: Dendogram summarizing the similarities between types of documents.

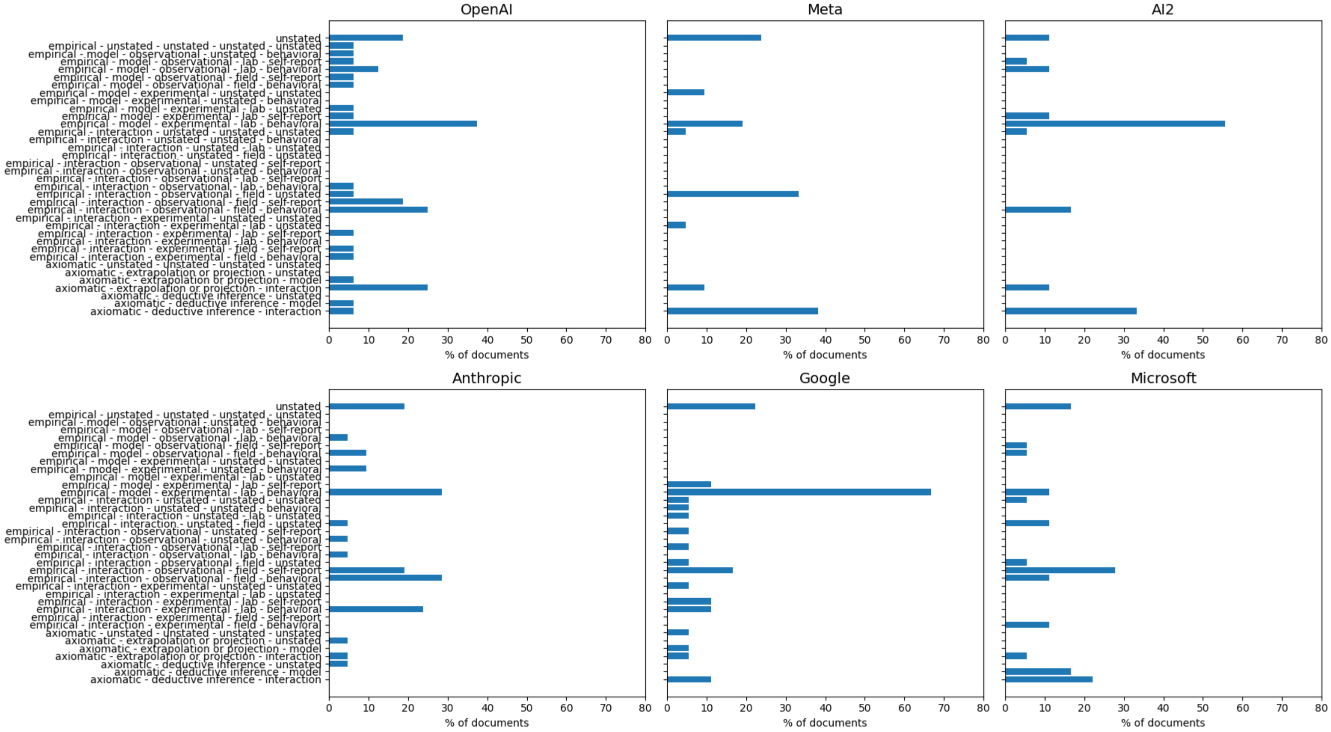


Figure 26: Percentage of times a method is employed within the documents of each company studied in our corpus.

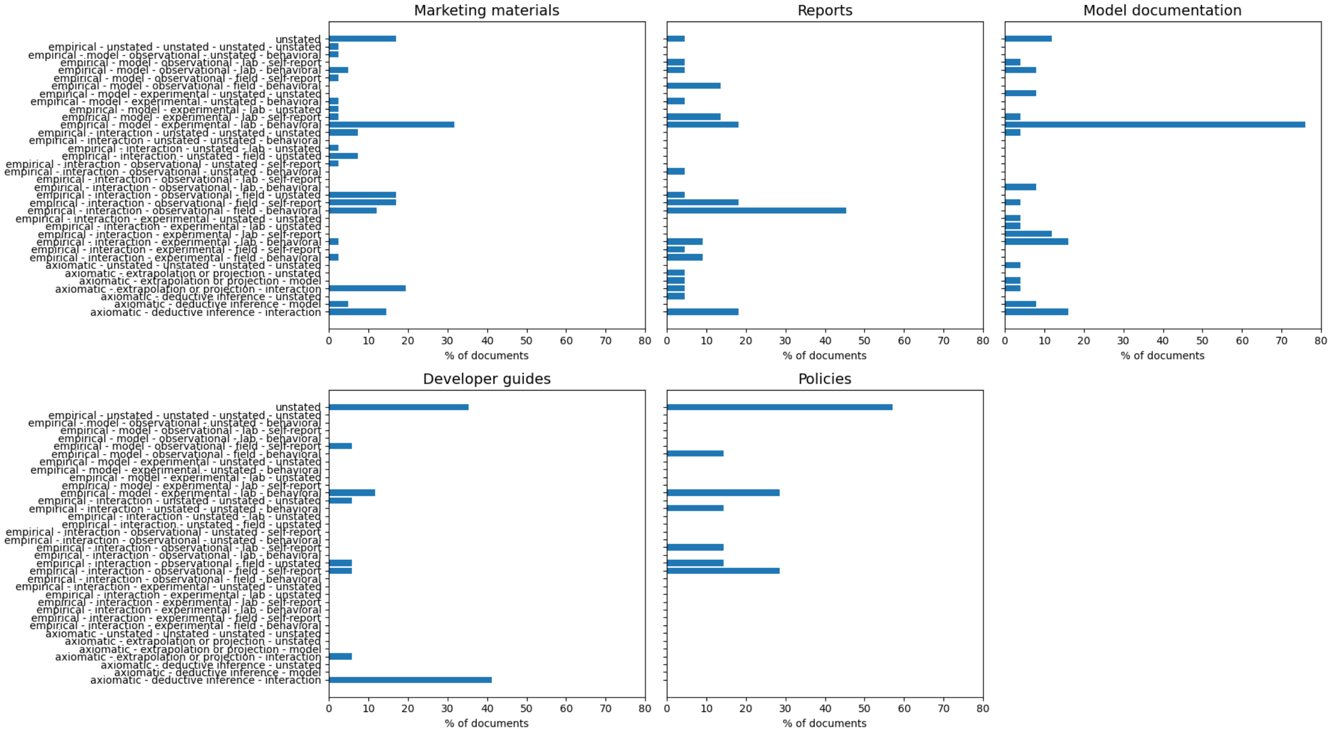


Figure 27: Percentage of times a method is employed within the documents of each document type we analyzed.

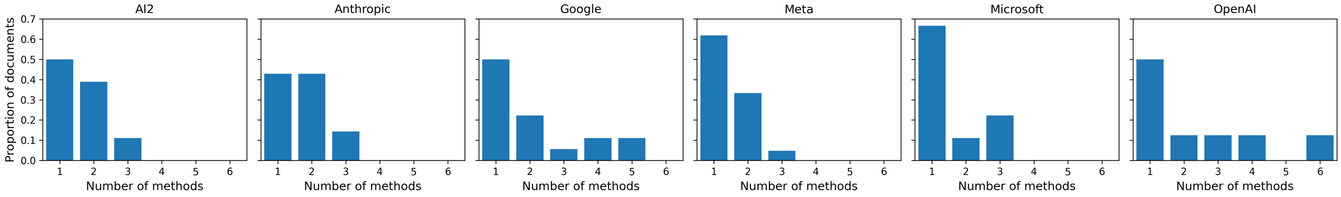


Figure 28: Distribution of the number of methods employed per document, normalized per company.

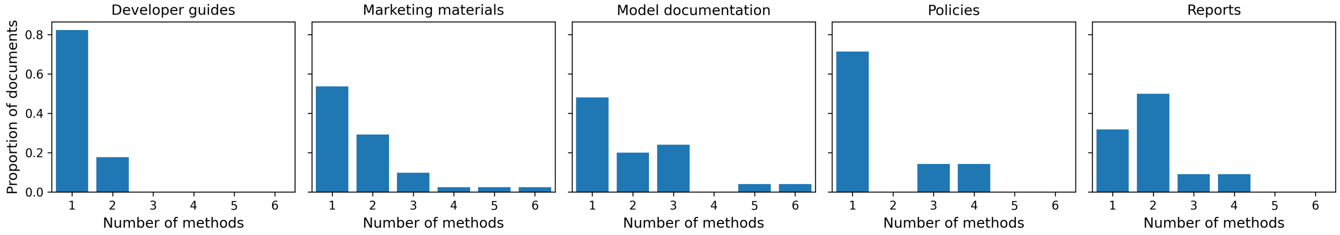


Figure 29: Distribution of the number of methods employed per document, normalized per document category.

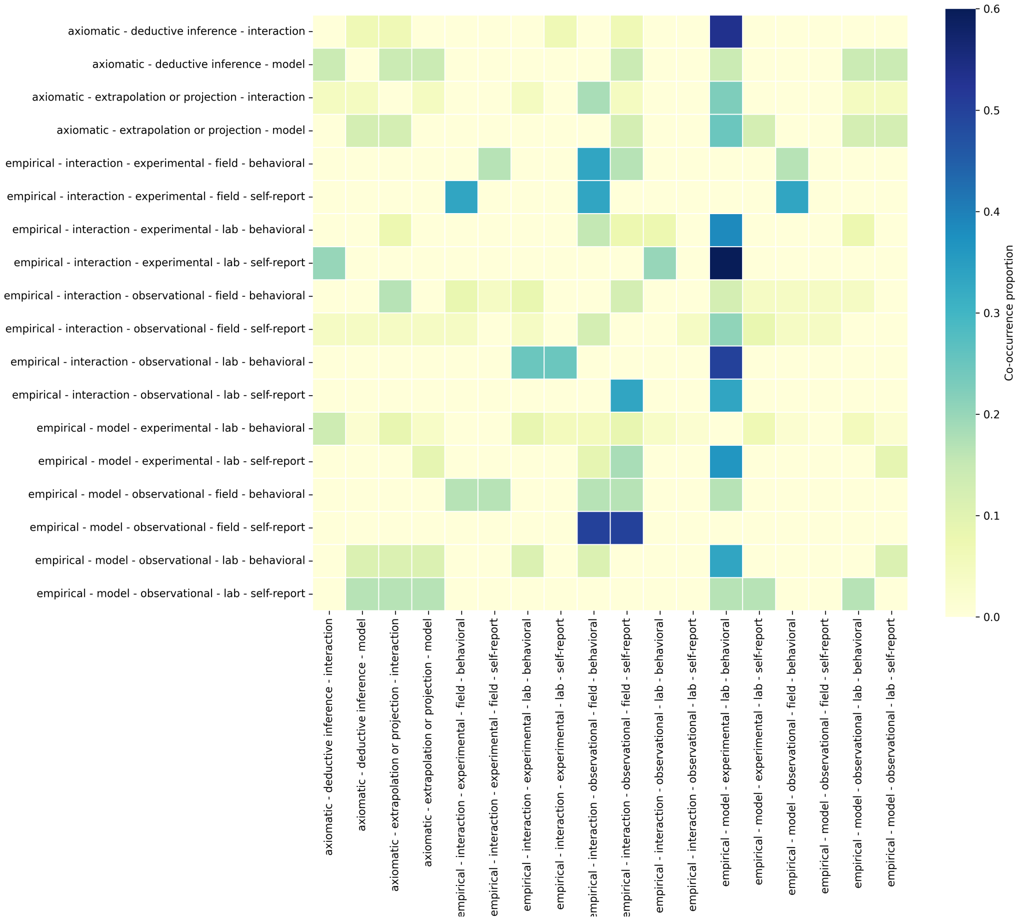


Figure 30: Percentage of times two methods co-occur in the documents of the corpus.

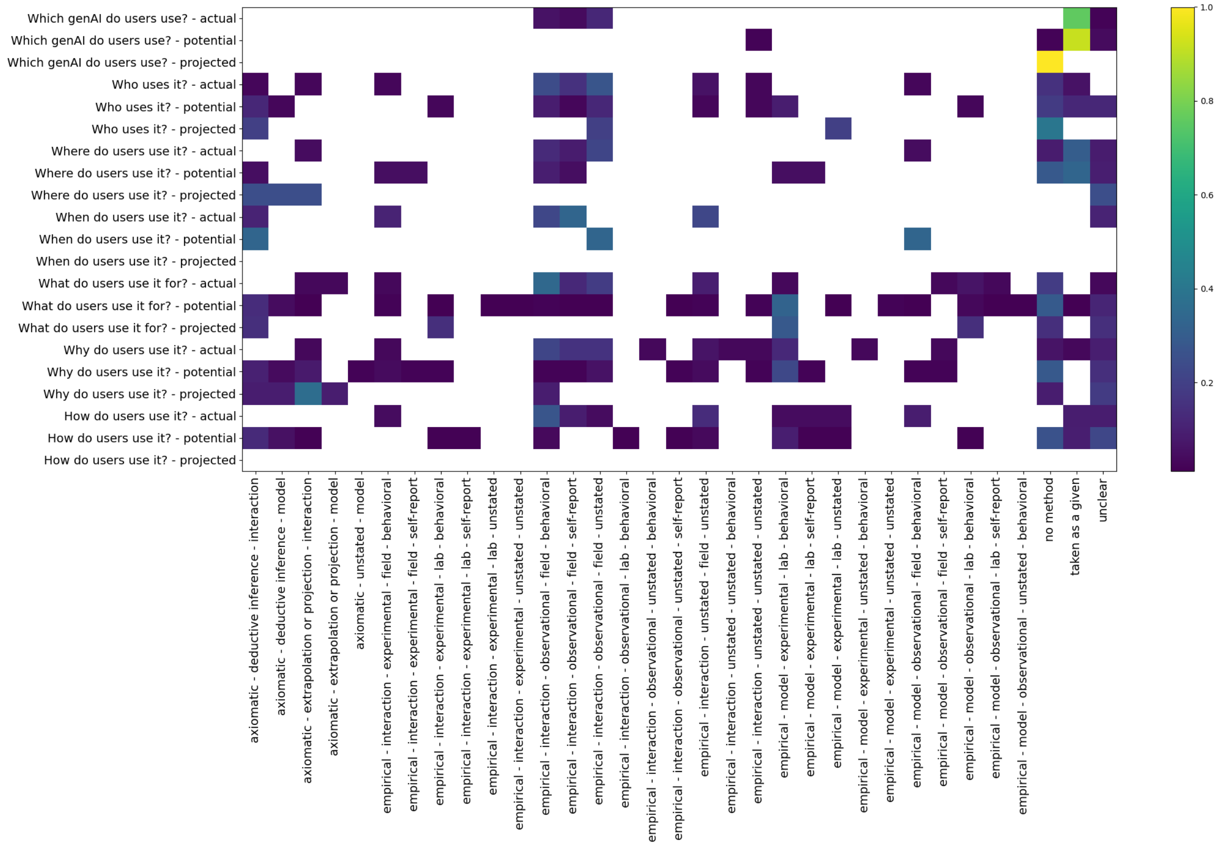


Figure 31: Percentage of times a method is employed to answer each use question and temporality (among all the documents that do anwser the use question at the associated temporality).

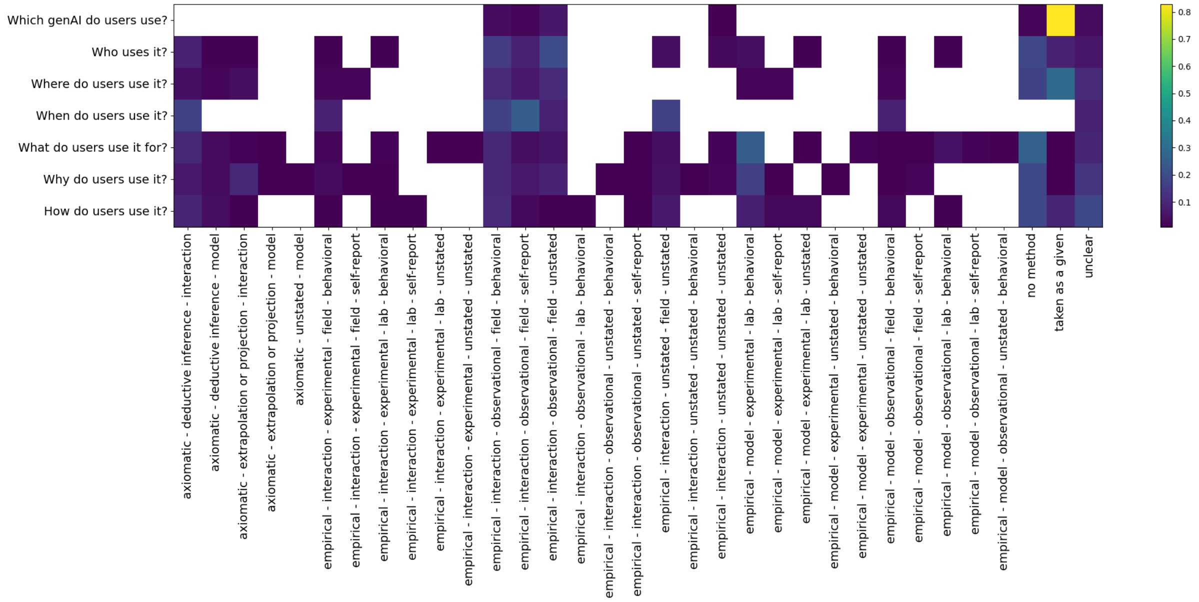


Figure 32: Percentage of times a method is employed to answer each use question (among all the documents that do anwser the use question).

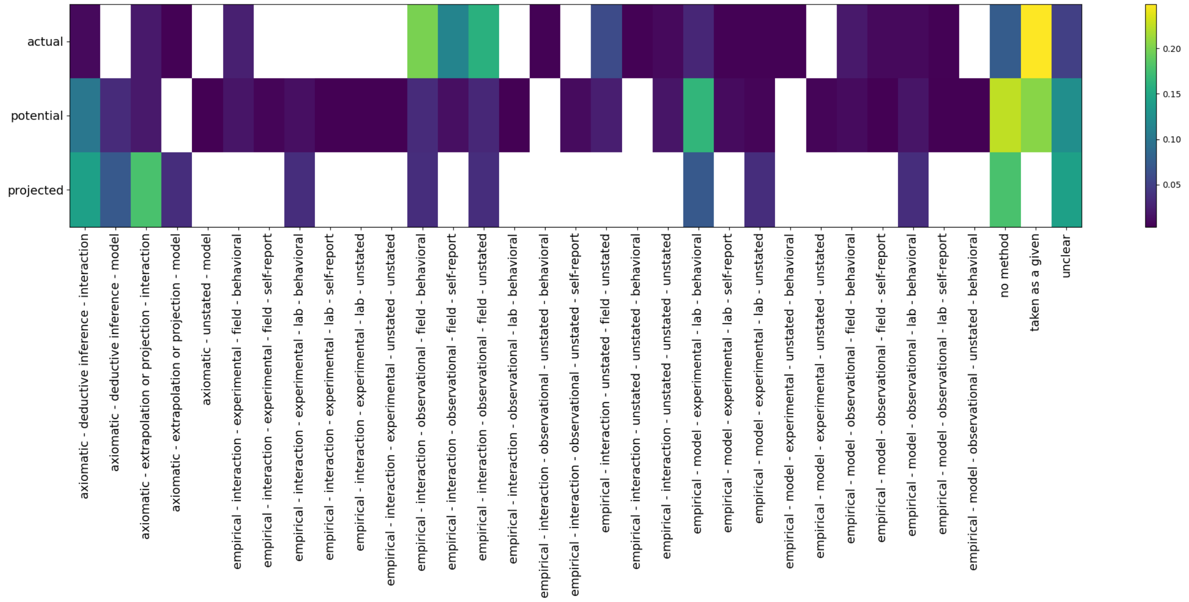


Figure 33: Percentage of times a method is employed to answer each temporality (among all the documents that do anwser the temporality).