EVALUATING LLM METRICS THROUGH REAL-WORLD CAPABILITIES

Justin K. Miller
School of Physics
University of Sydney
Camperdown, NSW 2006
justin.k.miller@sydney.edu.au

Wenjia Tang
School of Art, Communication and English
University of Sydney
Camperdown, NSW 2006
wenjia.tang@sydney.edu.au

June 13, 2025

ABSTRACT

As generative AI becomes increasingly embedded in everyday workflows, it is important to evaluate its performance in ways that reflect real-world usage rather than abstract notions of intelligence. Unlike many existing benchmarks that assess general intelligence, our approach focuses on real-world utility, evaluating how well models support users in everyday tasks. While current benchmarks emphasise code generation or factual recall, users rely on AI for a much broader range of activities—from writing assistance and summarization to citation formatting and stylistic feedback. In this paper, we analyse large-scale survey data and usage logs to identify six core capabilities that represent how people commonly use Large Language Models (LLMs): Summarization, Technical Assistance, Reviewing Work, Data Structuring, Generation, and Information Retrieval. We then assess the extent to which existing benchmarks cover these capabilities, revealing significant gaps in coverage, efficiency measurement, and interpretability. Drawing on this analysis, we use human-centred criteria to identify gaps in how well current benchmarks reflect common usage that is grounded in five practical criteria: coherence, accuracy, clarity, relevance, and efficiency. For four of the six capabilities, we identify the benchmarks that best align with real-world tasks and use them to compare leading models. We find that Google Gemini outperforms other models—including OpenAI's GPT, xAI's Grok, Meta's LLaMA, Anthropic's Claude, DeepSeek, and Qwen from Alibaba—on these utility-focused metrics.

1 Introduction

DeepSeek-R1, released in January 2025, was widely promoted as the leading AI model [1, 2, 3, 4, 5]. Yet, such claims often relied on vague assertions and selective quotes rather than clear evidence. Subsequent AI models have similarly been declared superior, supported mainly by benchmarks centred around tasks like coding proficiency [6], multilingual abilities [7], and ancient script translation [8]. Concerns have arisen regarding these benchmarks' practical relevance [9], statistical robustness [10], susceptibility to adversarial inputs [11], and reliance on superficial fluency over factual accuracy [12]. Closed-source evaluations further exacerbate transparency issues [13, 14], and automated metrics may penalise useful outputs for deviating from rigid reference answers [15, 16].

This paper investigates how generative text-based AI is actually utilised in practice, identifying core capabilities that are critical in real-world interactions. While models are increasingly trained to align with human preferences: emphasizing helpfulness, harmlessness, and truthfulness via methods like reinforcement learning from human feedback (RLHF) and Constitutional AI [17, 18]. However, they continue to be evaluated on benchmarks such as MMLU (a suite of academic multiple-choice tasks across diverse subjects) and AIME (a benchmark based on American math competition problems), which prioritise abstract problem-solving and cognitive performance [19, 20]. This evaluation paradigm risks conflating intelligence with alignment, despite longstanding philosophical recognition that the ability to achieve goals does not guarantee ethical or value-aligned behaviour [21].

We evaluate whether existing benchmarks effectively capture these real-world interactions and propose a capability-aligned framework grounded in human-centred evaluation criteria. This approach aims to provide researchers, developers, and policymakers a clearer basis for assessing and comparing AI models according to practical effectiveness.

Generative AI produces content based on learned distributions from a vast multitude of different data sources [22, 23, 24, 25]. Text-to-text models, especially Transformer-based architectures like GPT-4 [26], LLaMA 3 [27], and DeepSeek V3 [28], dominate current applications. These models power widely adopted tools such as ChatGPT and Gemini, supporting tasks like writing and summarization [29, 30]. Their conversational capabilities arise from exposure to extensive linguistic data, enabling contextual adaptability [31, 32]. Given their widespread use and established evaluation ecosystem, this paper focuses specifically on these text-to-text models.

1.1 Generative AI Limitations

Generative AI faces several notable limitations. Stability remains a significant issue, particularly in tasks such as automated assessments, where subtle changes can cause inconsistencies in results [30, 33]. Ethical concerns also persist, including biases, inappropriate content Generation, privacy risks, and misuse [34]. Tools like ChatGPT and Midjourney have highlighted AI's struggles with sensitive content and biases [35, 36]. Accuracy further complicates AI deployment, as Large Language Models (LLMs) can fabricate plausible but incorrect information, notably citations [37, 38]. Such persistent issues underscore the necessity for careful evaluation aligned with realistic user expectations and contexts.

1.2 Patterns of AI Use and Implications for Benchmarking

AI adoption has significantly expanded, with over 75% of organizations deploying AI primarily to enhance operational efficiency and mitigate immediate risks [39]. 88% of AI users are non-technical employees, using generative AI tools predominantly for productivity-focused tasks such as writing, summarization, and idea Generation [40]. This widespread adoption by non-specialists emphasizes that benchmarks for AI must prioritize ease of use, reliability, and relevance to common productivity workflows, rather than purely technical performance metrics.

Sector-specific variations further inform benchmark design. For instance, AI adoption is significant in legal (26%), retail (40%), and healthcare industries, the latter forecasting substantial growth in diagnostics and efficiency improvements [41]. These distinct usage contexts suggest that effective benchmarks must be versatile enough to accommodate industry-specific applications and performance standards.

Individual engagement with generative AI, such as ChatGPT (used by 23% of U.S. adults), primarily involves text Generation for professional, educational, and entertainment purposes, a trend particularly strong among younger demographics [42]. This aligns closely with industry trends, highlighting that benchmarks should effectively evaluate AI's capabilities in realistic, text-oriented tasks reflective of actual usage.

The education sector provides a valuable example of the ethical complexities surrounding AI adoption, with significant numbers of students using generative tools despite acknowledging issues of academic integrity [43, 44]. Thus, benchmarks should not only measure technical proficiency but also consider transparency, explainability, and ethical guidelines as critical evaluation criteria.

In technical fields such as software development, AI integration is driven by tangible efficiency gains, with adoption rates reaching 62% among developers who primarily value task acceleration and productivity improvements [45]. Benchmarks aimed at technical sectors should therefore explicitly measure productivity enhancement, integration ease, and reliability in code Generation and debugging tasks.

Finally, professional task alignment strongly influences AI adoption, with higher integration observed in fields like marketing and journalism, and cautious approaches in finance, healthcare, and education due to heightened concerns around accuracy, privacy, and ethical responsibility [46]. This pattern suggests benchmarks must account for domain-specific tolerances for error and incorporate assessments that reflect real-world consequences of AI inaccuracies or misuse.

Collectively, these adoption trends highlight the necessity for human-centric AI benchmarks that not only capture technical capability but also meaningfully assess the practical value, ethical considerations, and contextual relevance of AI across diverse sectors and user groups.

1.3 AI use as Conversation

Text-to-text generative AI inherently adopts a conversational structure, characterized by user prompts and corresponding AI-generated responses [47]. This has frequently led to their framing as artificial intelligence-based chatbots, for example, in educational [48], medical [49], and commercial contexts such as Google's Gemini [50], Microsoft's Copilot [51], and META's LLAMA [12]. Such portrayals underscore dialogue's integral role in generative AI's functioning—interpreting instructions, engaging users, facilitating iterative task completion, and aligning closely with user intent [34, 37].

Despite this conversational framing, existing benchmarks inadequately capture the complexities of genuine conversational interaction. Historically, evaluative frameworks like Grice's Cooperative Principle—with maxims emphasizing clear, sufficient, relevant, and contextually appropriate communication—have guided conceptions of effective human dialogue [52, 53]. Such principles reflect essential conversational dimensions, including mutual understanding, cooperation, and grounding to prevent misunderstandings [54, 55]. However, this standard cannot be directly adopted into human-to-machine communication with different contexts and scenarios. Current AI evaluation standards predominantly focus on discrete task performance, neglecting critical conversational qualities such as iterative clarification, contextual relevance, and implicit meaning interpretation [56, 57, 58].

Moreover, human-computer dialogues exclusively rely on textual communication without supplementary non-verbal cues, which are natural qualities of conversation between humans [59]. Effective communication with AI thus demands explicit attention to context, subtext, and implied meanings—elements often absent from some prevailing metrics.

To bridge these gaps, this study proposes a capability-aligned, human-centred evaluation framework explicitly designed around conversational dynamics. This framework integrates key conversational dimensions—coherence, accuracy, clarity, relevance, and efficiency—and emphasizes iterative adaptation and learning [60]. By aligning benchmarks with authentic conversational interactions, this approach aims to enhance AI's practical effectiveness in real-world applications.

1.4 What are "Good" AI Conversations?

Building upon historical research on conversational quality and contemporary expectations for generative AI, we draw from classic theories of human-to-human conversation and incorporate characteristics of AI usage contexts to propose criteria for systematically evaluating interactions between LLMs and human users. These criteria are divided into two clear categories: *objective criteria*, which are directly observable and measurable within the conversation itself, and *subjective criteria*, which reflect users' personal evaluations and perceptions regarding conversational effectiveness and task performance.

Objective Criteria

Objective criteria encompass measurable conversational characteristics directly observable without requiring subjective user interpretation:

- **Coherence:** Maintaining logical consistency and structured dialogue is essential for effective communication [61]. Coherent conversations enable both human and AI participants to effectively integrate previous conversational context, clarify ambiguities, and address implicit meanings, thereby collaboratively progressing toward shared conversational objectives [53].
- **Accuracy:** Accuracy directly determines the reliability of information provided by AI, thus significantly impacting task effectiveness and user trust [62]. Generative AI often produces inaccuracies or entirely fabricated information—a phenomenon known as "*hallucination*"—due to inherent limitations in adapting to new or unseen data [63, 37]. Consequently, validating content accuracy and truthfulness is essential for practical scenarios where precise and trustworthy outcomes are necessary, such as medical, legal, or financial contexts [30].
- **Relevance:** Providing relevant information focused explicitly on the task significantly improves conversational effectiveness. Extraneous or redundant details can obscure intent, disrupt user focus, and diminish the quality and effectiveness of the interaction [52]. Therefore, AI responses should prioritize delivering content directly aligned with achieving user-defined objectives.
- **Clarity:** Effective conversations prioritize concise and clear expressions, thereby reducing cognitive demands on users. By avoiding unnecessary complexity, users can quickly interpret and utilize AI-generated content without additional cognitive overhead or extensive clarification [61].

- **Efficiency:** Conversational efficiency involves proactively providing and requesting only essential information required for task completion, reducing redundancy, and minimizing cognitive overload [64]. Efficient interactions optimize the conversational process, ensuring streamlined task accomplishment without unnecessary cognitive or temporal expenditures.

Subjective Criteria

Subjective criteria involve user perceptions and emotional responses, significantly influencing the perceived quality and task-oriented success of interactions:

- **Politeness:** Politeness significantly impacts users' perceived satisfaction and conversational quality. While users recognize AI as non-human entities, polite interactions foster social presence, enhance user comfort, and encourage user engagement [65]. Politeness thus indirectly supports effective task completion by increasing users' willingness to cooperate, provide thorough input, and positively engage with AI tools.
- **Conversation Structure:** The conversational structure—encompassing vocabulary selection, linguistic style, and adherence to conversational norms—directly influences perceived effectiveness and user satisfaction [66, 67, 68]. Adhering to established social and conversational norms enhances perceived interaction quality, indirectly improving task outcomes by fostering user comfort and engagement [69].
- **Satisfaction:** User satisfaction reflects the alignment between conversational outcomes and user expectations. Unlike purely transactional interactions, users typically engage AI tools seeking supportive, reliable assistance analogous to interactions with peers or friends [70]. Satisfaction directly impacts user willingness to repeatedly utilize AI tools and ultimately affects perceptions of conversational effectiveness and success in task-related contexts.
- **Motivation:** Addressing intrinsic motivational needs—including emotional fulfillment, personal security, and recognition—improves users' emotional investment and trust in conversational exchanges [55]. AI tools capable of recognizing and addressing these motivational dimensions enhance overall user engagement, indirectly improving task performance by sustaining user interest, cooperation, and trust.

These criteria, synthesized from prior literature and recent AI advancements, provide clear theoretical benchmarks for conversational quality. Understanding these expectations enables a structured evaluation of real-world AI performance, identifying demonstrated capabilities and highlighting areas for targeted improvements. The following sections explore how well current AI implementations align with these established conversational expectations, clarifying the existing gaps and opportunities for future enhancements.

2 Methods

To investigate how AI is being used in general tasks, we built on two key sources: a large-scale survey of Danish workers [46], and an empirical analysis of generative AI usage using Anthropic's Claude.ai platform [71].

2.1 Identifying AI-Capable Tasks

To identify the types of occupational tasks most amenable to AI assistance, we conducted a small-scale exploratory analysis using data from a study that surveyed over 18,000 workers across 11 occupations in Denmark and linked survey responses with administrative labor market records [46]. The authors used a refined version of OpenAI's "Direct Exposure (E1)" measure to assess whether GPT-3.5 could significantly reduce the time required to complete specific job tasks. A task was marked as providing "Large" time savings if access to ChatGPT could halve the time needed for an average worker to complete it, at equivalent quality.

These task ratings were generated through a hybrid process involving GPT-3.5 prompting and human validation. Specifically, the authors assessed Detailed Work Activities (DWAs) from the U.S. O*NET database, selecting six representative tasks per occupation based on their average exposure scores and importance. Of the 66 tasks, the final sample included 21 tasks across the 11 occupations, all rated as having high productivity potential through ChatGPT assistance [46, SI Appendix, Section 1.A].

From this list of high-exposure tasks, we performed qualitative thematic analysis [72] to categorize the types of tasks where ChatGPT offers the greatest productivity gains. Each author independently reviewed the 32 tasks and grouped them by thematic similarity based on the cognitive processes and outputs involved. We iteratively discussed and refined these groupings until reaching consensus on six recurring AI capabilities. These categories are not mutually exclusive; individual tasks were often associated with multiple capabilities reflecting the overlapping nature of real-world work.

See Table A1 for a complete mapping of each task to its assigned capabilities. This preliminary classification serves as the basis for developing capability-aligned AI evaluation metrics.

2.2 Validating Capabilities with Real-World Usage

To test the robustness of the AI capability categories found during the qualitative thematic analysis, we validated them using a second data source [71]. This source mapped over four million Claude.ai prompts to O*NET tasks and grouped them by occupational categories.

We accessed the dataset provided by Anthropic, extracting the specific O*NET tasks, their associated occupational categories, and the percentage of total prompts linked to each task. In total, the dataset included 35,014 unique tasks. A cumulative distribution plot of task frequency is shown in Figure 1. Notably, the top 100 most frequent tasks accounted for just over 50% of all usage.

We therefore focused our analysis on these top 100 tasks, which represent over two million real-world AI prompts. This approach enabled a broad coverage of usage while keeping the qualitative review tractable. Each task was manually reviewed by the authors and mapped to one or more of the six AI capabilities identified earlier.

2.3 Constructing Prompts and Usage Mapping

For each of the top 100 tasks, we assigned at least one AI capability, with many tasks mapping to multiple capabilities. To illustrate how each capability manifests in practice, we constructed an example prompt for every task–capability pairing. These prompts were designed to reflect how a user might realistically request assistance from a language model when performing the given task.

Creating example prompts served two purposes. First, it ensured that the mapping between tasks and capabilities was grounded in real-world usage, not just abstract categorisation. By imagining a plausible user prompt, we were able to validate whether a capability truly applied to the task and how it might be operationalized in a typical human–AI interaction. Second, the prompts provided an interpretable bridge between occupational tasks and the kinds of language-based inputs that LLMs are designed to handle. This step enhanced the interpretability of our mapping and ensured that our subsequent evaluation remained anchored in naturalistic language use. The full table for the top 100 tasks and example prompts given by the authors for each task can be found in Table C3

Using the usage percentage data provided by Anthropic, we then computed the cumulative usage share for each AI capability. This allowed us to quantify how heavily each capability featured in real-world AI interactions. The resulting distribution is shown as a comparative bar chart in Figure 2, highlighting the relative importance of each capability based on actual usage frequency.

2.4 Evaluation of Metrics

To assess how well existing AI benchmarks reflect real-world usage, we evaluated a set of widely used metrics through the lens of the six AI capabilities in Appendix B Table 1. This step completes the methodological framework by mapping current benchmarks to the practical functions users expect from language models.

We based our evaluation on five objective qualities commonly expected in human–AI interactions: coherence, accuracy, clarity, relevance, and efficiency. These criteria serve as a human-centered lens for assessing whether a benchmark realistically captures how users interact with AI systems. Each criterion is defined below

- **Coherence:** Does the metric assess AI performance using prompts and formats that resemble genuine human interaction, rather than artificial test structures, such as multiple choice?
- **Accuracy:** Is the information verifiably correct according to trusted sources or gold-standard answers? We examined whether each metric had a clear standard for correctness and the types of sources it referenced.
- **Clarity:** Does the metric measure whether outputs are easy to understand and clearly worded? Even a technically correct output may be unhelpful if it is confusing or overly complex.
- **Relevance:** Does the benchmark test a broad and meaningful range of content within the capability domain? For instance, a programming benchmark focused solely on python would not necessarily help with coding problems in Java.
- **Efficiency:** Does the metric reflect how much time or cognitive effort the AI helps save? While often overlooked, this is critical to assessing practical utility in workplace settings.

We reviewed the release notes and technical specifications of recent foundation models from OpenAI [73], Google Gemini [74], Deepseek [75], xAI [76], Qwen [77], Meta [78], and Anthropic [79]. From these sources, we compiled a list of benchmarks commonly used to evaluate text-to-text generative capabilities:

> *MMLU, AIME, Codeforces, GPQA, SWE-bench Verified, FACTS Grounding, MRCR, HumanEval, SimpleQA, GSM8K, Math-500, Bird-SQL, LiveCodeBench, Humanities Last Exam, MMLU-Pro, AGIEval English, CRUXEval-I/O, SimpleQA, Chatbot Arena, Webdev Arena* [80, 20, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 19, 94, 95, 96, 97, 98]

Each metric was then mapped to the AI capability it most directly tests. To assess alignment, we reviewed the benchmark's design, inputs, outputs, and scoring methodology to determine how well it addressed each of the five human-centered evaluation criteria: coherence, accuracy, clarity, relevance, and efficiency. This set of five criteria constitutes our evaluation framework.

For each capability, we then selected the benchmark that best aligns with all five dimensions, based on our independent reviews of the benchmark documentation and, where applicable, third-party validation studies. Full assessments for each benchmark are included Appendix A Table B2.

3 Results

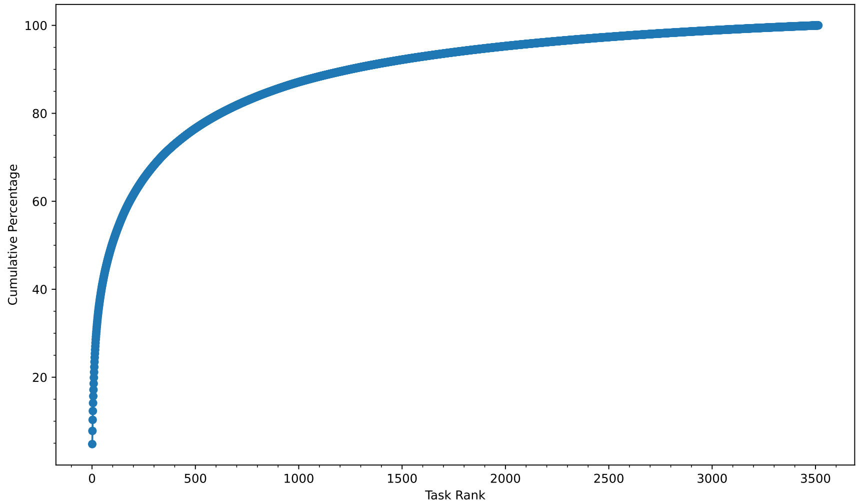

Figure 1: Cumulative distribution of AI usage across 35,014 occupational tasks. Each task is ranked by the percentage of total Claude.ai prompts associated with it, using data from Handa et al. (2025) [71]. The x-axis shows task rank in descending order of frequency, and the y-axis represents the cumulative percentage of total prompts. A small number of tasks dominate usage: the top 100 tasks account for just over 50% of all prompts.

In Figure 1, we observe a cumulative distribution curve illustrating how AI usage is concentrated across 35,014 occupational tasks, based on Claude.ai prompt data from Handa et al. (2025) [71]. The x-axis ranks tasks from most to least frequently prompted, while the y-axis shows the cumulative percentage of total prompts.

The steep initial rise of the curve reveals a strong skew: a small subset of tasks accounts for a disproportionately large share of usage. Notably, the top 100 tasks alone comprise just over 50% of all prompts, while the top 500 account for just under 80%. This long-tail distribution indicates that although AI is applied across a broad range of tasks, actual usage is highly concentrated in a narrow band of frequently occurring requests.

3.1 AI Capabilities

In Table 1, we present six core AI capabilities that emerged from thematic analysis of occupational tasks marked as highly automatable, as well as validation using real-world usage data from Claude.ai. These capabilities: Summarization, Technical Assistance, Reviewing Work, Data Structuring, Generation, and Information Retrieval, capture the main functions AI performs to support workers across diverse domains. Each reflects a distinct mode of interaction:

for example, Information Retrieval supports fact-finding without producing new and original content, while Generation involves the creation of novel material. The examples included illustrate how the same capability can span both objective and subjective domains. Summarization, for instance, may involve extracting price trends from datasets (objectively measurable) or summarizing an author's viewpoint (subjective).

The table also distinguishes between objective and subjective types of evaluation. A key distinction between these two types is that objective metrics can often times be automated, and thus are easier for metrics to measure, while Subjective measurements require human evaluation. These distinctions offer insight into where AI performance can be reliably benchmarked and where human input remains essential. We validated the robustness of these categories using the top 100 most frequently used occupational tasks which account for over 50% of four million Claude.ai prompts. Each of the top 100 most frequently used tasks could be linked to at least one of the six capabilities, and many to multiple.

AI Capability	**Explanation**	**Type of Evaluation**	**Examples**
Summarization	Analysing large amounts of content to extract key information and present concise summaries	Subjective	Summarise the author's main arguments in the web pages [99]
Technical Assistance	Providing clear instructions, diagnosing issues, and suggesting step-by-step solutions for software and hardware problems	Objective	Write optimized code [100]
Reviewing Work	Identifying issues, evaluating systems or processes, and recommending improvements or solutions	Subjective and Objective	Objective correct math homework answers [101] Subjective: review my email for tone and clarity and suggest improvements
Data Structuring	Efficiently managing, logging, and maintaining accurate records of transactions, interactions, or documentation	Objective	Change academic reference list from Chicago to APA style [102]
Generation	Creating original ideas, drafting written content, or suggesting creative solutions based on provided parameters or goals	Subjective	Create content for an About Us page for a small bakery.
Information Retrieval	Finding and delivering relevant factual or background information in response to queries, often based on pre-trained knowledge or Retrieval-augmented Generation (RAG)	Objective	What is the current price of mangoes in Australia

Table 1: This table presents a multidimensional framework connecting six real-world AI capabilities with key evaluation criteria. Each capability (e.g., Summarization, Generation) is assessed using both objective and subjective criteria—reflecting whether the capability can be evaluated is based on observable outputs (objective) or user perception (subjective). This structure highlights how current benchmarks align with, or overlook, practical AI uses by showing which criteria are essential for evaluating each capability. The intersecting dimensions are necessary to capture the full complexity of AI performance in real-world tasks.

Figure 2 shows the relative importance of each AI capability across the top 100 most commonly used occupational tasks, as identified in over four million real-world Claude.ai prompts. Each bar reflects the cumulative percentage of task–capability associations, normalized by the maximum cumulative importance observed. Importantly, tasks can be linked to more than one capability—for example, the task "modify existing software to correct errors, to adapt it to new hardware, or to upgrade interfaces and improve performance" requires both Technical Assistance and Reviewing Work. This approach captures not just how many tasks are associated with each capability, but how heavily those tasks feature in real-world AI use.

The results highlight Technical Assistance (65.1%) and Reviewing Work (58.9%) as the most prevalent capabilities, suggesting that AI is most often used to support problem-solving, diagnosis, and improvement-related activities. In contrast, capabilities such as Generation (25.5%), Information Retrieval (16.6%), and Summarization (16.6%) are moderately represented, while Data Structuring (4.0%) is relatively rare among the top-used tasks. These findings

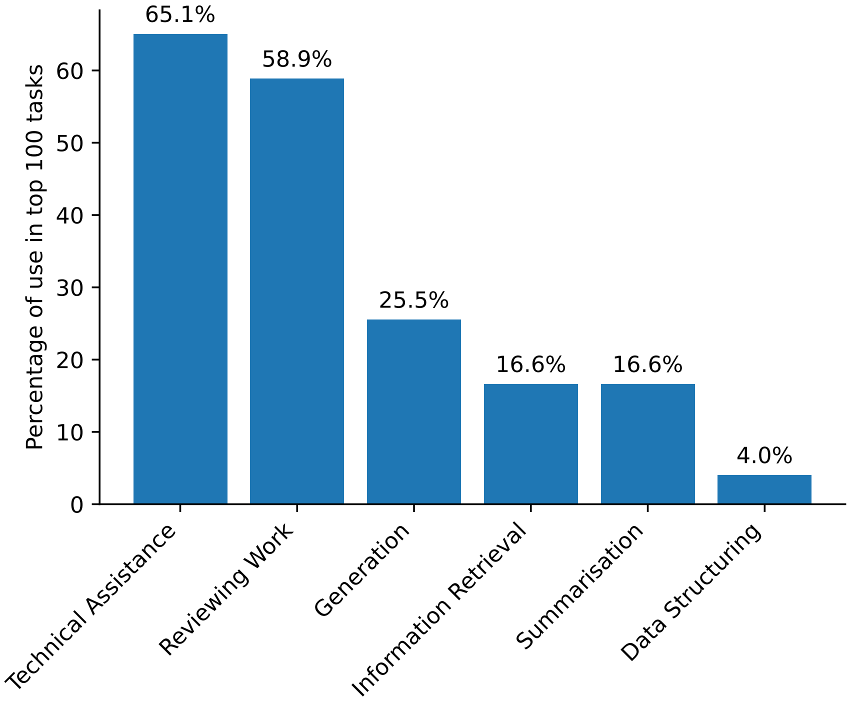

Figure 2: Percentage of the top 100 occupational tasks for which each AI capability is relevant. A task can be associated with multiple capabilities. For example, the task *"modify existing software to correct errors, to adapt it to new hardware, or to upgrade interfaces and improve performance"* involves both Technical Assistance and Reviewing Work. The percentages reflect the sum of the percentage scores (derived from over 4 million prompts) associated with each task–capability pair, divided by the maximum cumulative importance across the top 100 tasks. This highlights how commonly each capability appears among the tasks people are most likely to use AI for.

suggest that users most frequently rely on AI for tasks that require operational or evaluative reasoning, while tasks involving routine formatting play a less prominent role.

3.2 Metrics Used by AI Companies

Of the six AI capabilities we identified, we found that existing benchmarks covered only three of them: *Information Retrieval, Technical Assistance*, and *Summarization*. For these three, we were able to identify relevant and widely-used metrics. For instance, MMLU and AGIEval English were most aligned with Information Retrieval, while SWE-bench and HumanEval were relevant to Technical Assistance.

To assess whether the remaining capabilities were covered by existing benchmarks, we conducted a structured search of academic literature, leaderboard repositories, and benchmark evaluations for newer foundation models. Our goal was to identify any widely used benchmarks that could be reasonably mapped to the remaining capabilities: Generation, Reviewing Work, and Data Structuring.

This search revealed WritingBench as a suitable candidate for evaluating the Generation capability [96]. WritingBench includes structured tasks across multiple writing domains—such as persuasive, technical, and informative writing—and uses a fine-tuned critic model validated against human judgments to evaluate output quality. These features align closely with the Generation capability as defined in our framework, which involves creating original content across diverse goals and genres.

In addition, we identified the Creative Writing section of Chatbot Arena as another relevant benchmark for this capability [98]. It evaluates open-ended Generation tasks such as storytelling and poetry through pairwise human preference voting, providing a comparative score that reflects subjective judgments of human preference. Together, these two benchmarks represent the only benchmarks found that reflect the Generation capability.

For Reviewing Work and Data Structuring, however, we found no widely adopted benchmarks that explicitly test these capabilities. In the absence of suitable benchmarks, we did not attempt to assign scores to these capabilities. Instead, for the four capabilities where relevant benchmarks did exist—Information Retrieval, Technical Assistance, Summarization, and Generation—we collected publicly reported benchmark scores from leading models developed by Google, Anthropic, OpenAI, xAI, Meta, DeepSeek, and Alibaba. For each capability, we selected the highest-scoring model per company, using confidence intervals or reported variance where available to facilitate meaningful comparisons.

We now examine the extent to which existing benchmarks capture these core capabilities, identifying areas of alignment as well as major blind spots in practical utility.

3.2.1 Technical Assistance

Among the evaluated benchmarks for Technical Assistance, WebDev Arena stands out as the most realistic and user-centered according to our criteria [97]. In terms of coherence, WebDev Arena uses open-ended prompts that closely resemble how humans naturally ask for help in practical settings (e.g., “Build a chess game” or “Clone the WhatsApp UI”). Unlike benchmarks such as HumanEval or CRUXEval-I/O, which the authors of this paper examined and found to focus on narrowly defined tasks implemented as Python functions, WebDev Arena reflects real-world requests that are broader, more ambiguous, and closer to how LLMs are actually deployed. For accuracy, while WebDev Arena does not rely on gold-standard answers, it uses human preference voting as a judgment mechanism, allowing outputs to be evaluated in a context-sensitive manner. This contrasts with benchmarks like SWE-bench Verified, where nearly one-third of correct solutions involved cheating or weak test cases [103], and Codeforces, which suffers from dataset leakage over time [104]. On the dimension of clarity, WebDev Arena does not explicitly penalize unclear output, but human preference may implicitly reflect whether responses are well-structured and comprehensible—an approach that is arguably more aligned with end-user needs than automated tests like pass@k used in HumanEval. Regarding relevance, WebDev Arena’s major strength is its focus on front-end web development using HTML, CSS, and JavaScript across eleven broad categories and many subcategories. In contrast, most other benchmarks—such as Codeforces, HumanEval, LiveCodeBench, CRUXEval-I/O, and SWE-bench Verified—evaluate tasks exclusively in Python. This makes WebDev Arena more representative of real-world UI and client-side development but also limits its relevance for back-end or systems-level programming. Finally, while none of the benchmarks measure efficiency directly, WebDev Arena indirectly reflects it through human preference, capturing how helpful or time-saving an LLM-generated solution feels to a user. By contrast, Codeforces is one of the few to report task completion time, and Bird-SQL introduces a reward-based efficiency score, though both rely on more narrowly scoped or language-specific tasks. Overall, WebDev Arena offers the most practical and human-aligned assessment of Technical Assistance, even while acknowledging its trade-offs in backend evaluation and objective scoring.

3.2.2 Information Retrieval

Among the benchmarks evaluated for Information Retrieval, SimpleQA emerges as the most well-rounded according to our criteria. It performs better than other benchmarks on coherence, as it presents concise, fact-based prompts and expects short-form answers, rather than relying on multiple-choice formats which diverge from how humans typically interact with language models. In contrast, benchmarks like MMLU, GPQA, and AGIEval all adopt rigid multiple-choice structures with fixed answer formats (for example, ”The answer is A”), which limit natural interaction and reduce applicability in real-world retrieval settings. Accuracy in SimpleQA, while not flawless, has undergone independent review with 94.4% agreement among expert annotators, providing a relatively high standard of validation[87]. This is notably more robust than GPQA, which has only expert agreement of 74% [82], and GSM8K, which initially included a high proportion of incorrect questions before refinement in the Platinum version [105]. MMLU has an overall accuracy of 6.49%, however these errors are not distributed uniformly among topics with the Virology section having an error rate of 57%. This error rate is an issue as most modern models are getting close to 90% [75]. For clarity, SimpleQA implicitly encourages clear, succinct answers, though like most other benchmarks—including Humanities Last Exam and Math-500—it does not explicitly assess how understandable the outputs are for human users. On relevance, SimpleQA covers a broad and diverse range of knowledge domains such as STEM, politics, art, geography, sports, and more, setting it apart from domain-specific benchmarks like GSM8K, AIME, and Math-500, which are focused solely on mathematics, or GPQA, which is limited to science subjects. Even the broader MMLU and Humanities Last Exam, though more interdisciplinary, often reflect a US-centric or highly academic framing that may not translate well to everyday queries. Finally, although efficiency is not directly measured in SimpleQA, this is a limitation shared by nearly all benchmarks in this space—only AIME provides an explicit latency-to-answer measure. On balance, SimpleQA offers the clearest, most coherent, and domain-relevant framework for evaluating Information Retrieval capabilities in language models.

3.2.3 Summarization

Among the limited benchmarks available for evaluating the Summarization capability, MRCR provides a more rigorous and direct test of a model's ability to condense and retrieve information from large contexts. Unlike FACTS Grounding, which relies on domain-specific materials such as legal texts, Wikipedia entries, and scientific documents, MRCR employs synthetic data that is carefully constructed to test core summarization skills—particularly the ability to distinguish and retrieve closely related pieces of information based on structure and context. This makes MRCR less reliant on background knowledge and more focused on the underlying cognitive capability being measured. While both benchmarks are comparable in coherence and clarity, FACTS Grounding introduces an epistemological issue by relying on an AI model from a different family to assess correctness. Although the designers take steps to avoid evaluation bias by not using models from the same family, the absence of a consistent comparison to human judgement raises concerns about the validity of its accuracy scores. MRCR, by contrast, provides more objective performance signals based on retrieval of precisely planted information. Neither benchmark explicitly measures efficiency, and both could be improved by incorporating user-centric metrics like time saved or effort reduced. Overall, despite the limited number of summarization benchmarks compared to other capabilities, MRCR offers a more principled and interpretable framework for evaluating summarization in language models.

3.2.4 Generation

Among the very limited benchmarks currently available for evaluating the Generation capability, Chatbot Arena: Creative Writing offers the most human-aligned and comparative approach. It uses head-to-head human preference voting to evaluate model responses to creative prompts, such as writing poems, short stories, or humorous narratives. This evaluation method aligns closely with how real users experience AI-generated content, making it particularly strong on coherence and clarity—two dimensions that are naturally captured through direct human judgment. Moreover, Chatbot Arena provides an aggregated score for each model, allowing comparisons across systems, which is valuable for tracking progress in open-ended Generation tasks. However, WritingBench also brings important strengths. Unlike the open-ended and sometimes subjective setup of Chatbot Arena, WritingBench is structured around clearly defined writing tasks spanning six core domains and one hundred subdomains, including persuasive, technical, and informative writing. It uses a fine-tuned critic model validated against human judgments to assess performance based on explicit criteria, including clarity, structure, and genre appropriateness. One key limitation shared by both benchmarks is the absence of an explicit measure of efficiency—neither evaluates how much cognitive effort or time is saved for the user. Nonetheless, given the field's current landscape, Chatbot Arena stands out for its human-grounded evaluation method and breadth of comparative scoring, while WritingBench remains a promising complement for more structured and evaluable writing tasks. Overall, the lack of established benchmarks in this area highlights that content Generation is an under-researched area within AI capabilities.

3.2.5 Mismatch Between Benchmarks and Real-World Use

Among the six real-world capabilities we evaluate, Reviewing Work and Data Structuring are not captured at all by the metrics used in current AI benchmarking. Meanwhile, Summarization and Generation, capabilities widely used by non-technical professionals are only weakly supported. This highlights a consistent mismatch between what AI benchmarks measure and how AI is actually used in practice. The gap suggests that existing evaluation tools underprioritise the capabilities most relevant to the broader AI user base.

3.3 Model evaluation

Figure 3 presents the performance of the highest-ranking models across four distinct AI capabilities, revealing several notable trends. Most prominently, Gemini 2.5 appears in the top two positions across all four tasks, ranking first in Summarization (89.1%), Generation (Elo score of 1458), and Technical Assistance (Elo score of 1420). This consistent high performance highlights Gemini's versatility and strength across a diverse range of tasks, including factual grounding, creative writing, and Technical Assistance. Anthropic models also demonstrate strong results: Claude ranks second in Summarization (79.4%) and second in Technical Assistance (Elo score of 1357).

4 Discussion

4.1 Metrics and Real–World Usage: A Fundamental Mismatch

The Results section demonstrates how the metrics employed by AI companies align with various functional capabilities. However, most technical reports released by AI companies, along with academic evaluations and industry

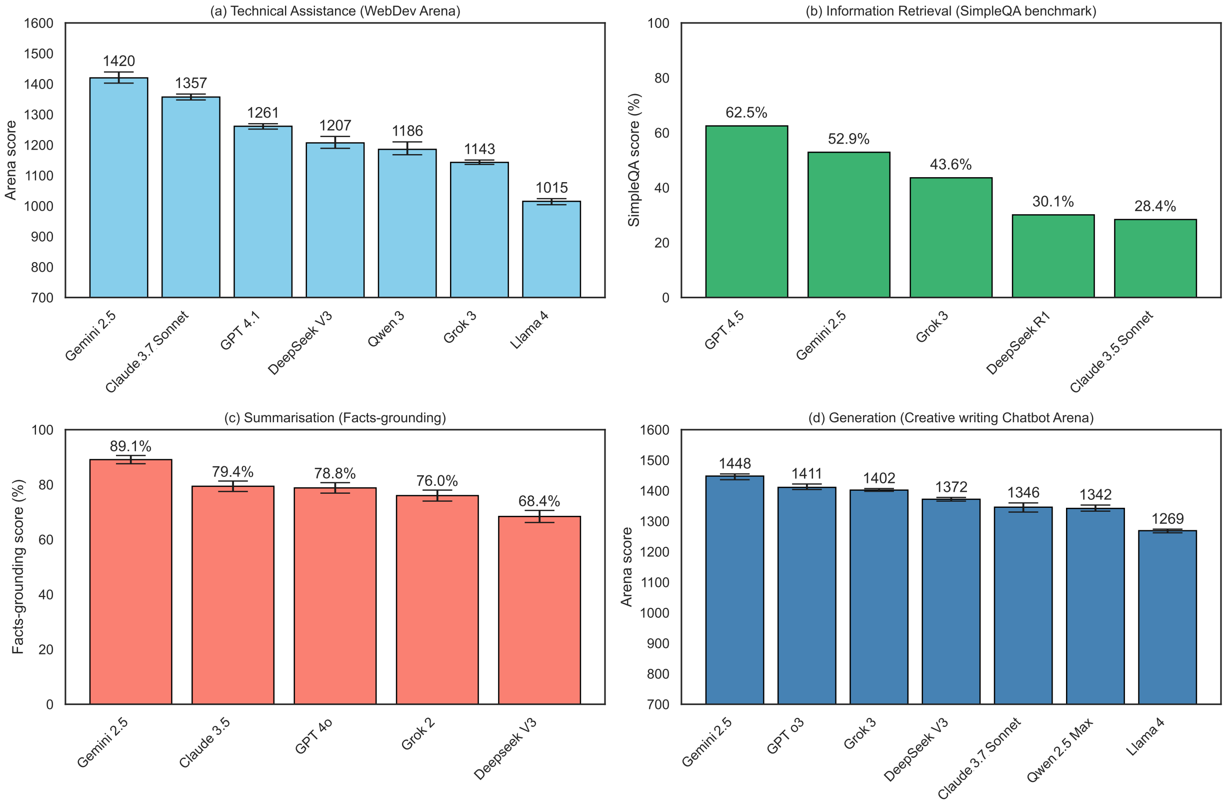

Figure 3: Evaluation results for highest ranking models, we only take the highest scoring model from one company, so they may have models that score higher than other companies. Scores are correct as of 12th May 2025. Across four of the six different AI capabilities outlined in this paper. **(a) Technical Assistance (WebDev Arena):** Elo-style Arena scores, derived from head-to-head matchups in web development tasks, reflect relative performance. Higher values indicate better comparative results: error bars show 95% confidence intervals. **(b) Information Retrieval (SimpleQA benchmark):** Accuracy percentage on a benchmark testing factual correctness in short-answer question responses. The y-axis shows the proportion of correct answers; higher values indicate better factual precision. **(c) Summarization (Facts-grounding benchmark):** Percentage of grounded content in model-generated summaries, based on human judgments of factual consistency with source documents. Error bars indicate 95% confidence intervals. **(d) Generation (Creative Writing Arena):** Elo scores from subjective human preferences in open-ended Generation tasks such as storytelling and creative writing. Higher scores denote consistent wins in pairwise comparisons; error bars show 95% confidence intervals.

leaderboards, rely heavily on aggregate benchmark scores to assess model performance [1, 26, 79]. These benchmarks are typically designed to evaluate models on narrowly defined tasks such as multiple-choice exams, coding problems, or symbolic logic puzzles—tasks that are meant to measure traditional conceptions of intelligence as the ability to reason abstractly, solve problems, or retrieve factual knowledge. Despite measuring AI capability in limited tasks, this framing obscures an important reality: LLMs are not just tools for solving bounded tasks with single correct answers—they are often deployed in open-ended, dialogic, and highly contextual workflows. Our analysis shows that, alongside technical tasks like coding and data manipulation, users commonly rely on LLMs for co-authoring emails, providing stylistic feedback, reformatting citations, summarizing long documents, and brainstorming ideas [71, 46]. These tasks are better understood not as tests of intelligence, but as forms of collaborative cognition where the AI plays the role of assistant, or editor, where applying the same objective standards to evaluate their strengths and weaknesses proves challenging.

This divergence between benchmark design and everyday use suggests that prevailing evaluations fail to capture the practical, human-centered goals that motivate the adoption of generative AI. By privileging abstract measures of machine intelligence over grounded assessments of utility, these benchmarks risk misrepresenting both the strengths and limitations of current models.This suggests a need to evaluate benchmarks not against abstract tasks alone, but through the lenses of real-world utility—where usefulness, clarity, and time savings take precedence over syntactic correctness or answer matching.

Most public benchmarks focus on technical tasks—code Generation, mathematics, and factual recall—while linguistically rich or socially situated tasks (e.g., tone editing, policy summarization) remain underrepresented. This skew reflects not just a preference for measurable outputs, but a deeper misalignment between benchmark design and the diverse ways people use AI in practice. While these tasks may be easier to score automatically, they capture only a narrow slice of the real-world interactions that users engage in daily—interactions that often involve open-ended reasoning, contextual sensitivity, and human judgment.

This mismatch is particularly evident when considering workforce data. Survey findings indicate that 88% of employees using AI on the job are non-technical workers—educators, healthcare staff, and retail associates—compared to just 12% in software development or engineering roles [40]. Yet benchmark coverage and public prompt datasets disproportionately reflect the workflows of developers. In Anthropic's prompt-to-task dataset, for example, nearly two-thirds of high-use prompts relate to Technical Assistance [71]. One hypothesis for this imbalance stems from differences in prompt frequency: developers often engage in multi-turn, iterative interactions with AI tools, whereas non-technical users may issue fewer but more varied prompts aimed at business communication, summarization, or formatting. The result is a feedback loop in which coding use cases dominate evaluation design, while broader categories of human–AI collaboration remain undervalued, despite their wide-use. However, a study would need to be conducted to confirm this.

4.2 The Vanishing Human in the Loop

Although metrics such as HumanEval and MMLU purport to assess model accuracy and reasoning, they depend almost entirely on automated scoring pipelines and, in some cases, on other LLMs for adjudication [10, 11]. This approach contrasts sharply with the human-mediated evaluation practices that govern professional and academic settings, —such as peer review of scholarly writing, editorial judgment in journalism, or code reviews in software engineering where domain experts judge the quality of summaries, translations, or code reviews [106]. By eliminating human raters, large-scale benchmarks can obscure ambiguous or context-sensitive errors—hallucinations, cultural misinterpretations, or stylistic misalignments—that are readily detected in real-world use. Reintroducing human judgement into evaluation frameworks, even at the expense of smaller test sets, would more faithfully reflect the collaborative nature of human–AI workflows.

4.3 Coverage Gaps Across Capabilities

Our thematic analysis identified six core AI capabilities—Summarization, Technical Assistance, Reviewing Work, Data Structuring, Generation, and Information Retrieval—yet only four of these map onto established benchmarks. Notably, Reviewing Work, which encompasses proofreading, tone adjustment, and structural feedback, appears in 58.9% of high-frequency occupational tasks but lacks a dedicated evaluation suite. Similarly, Data Structuring tasks such as citation conversion or table formatting account for 4% of real-world prompts yet are unrepresented in mainstream test collections. The absence of metrics for these capabilities not only deprives users of guidance when selecting models but also enables vendors to emphasize strengths in well-measured areas while neglecting essential collaborative functions.

4.4 Blind Spots in Current Benchmarks

Building on the benchmark reviews in Section 3.2, several blind spots emerge that constrain their applicability to real-world workflows. While current benchmark suites offer valuable insights into certain model capabilities, they also exhibit notable blind spots that limit their applicability to real-world workflows. First, the emphasis on Python in Technical Assistance benchmarks—such as HumanEval and MBPP—overrepresents support for a single programming language. In contrast, few benchmarks evaluate assistance with other widely used tools and languages such as R, Excel, LaTeX, or domain-specific scripting environments, despite their prevalence in professional contexts. As a result, models may appear well-rounded in technical performance while providing limited support for non-Python workflows.

Second, most existing benchmarks prioritize correctness while largely overlooking efficiency. For example, a model may return an accurate citation or summary, but the time or cognitive effort required to parse or integrate the output remains unmeasured. In real-world settings, a key value proposition of generative AI is its ability to reduce user workload. Yet benchmarks rarely quantify this benefit, failing to ask: How much faster did the user complete their task? or How many revision cycles were saved?

Finally, there is limited attention paid to interpretability from the user's perspective; an aspect we categorize under subjective criteria. Evaluation pipelines often treat the presence of a correct answer as sufficient, without assessing whether the output is understandable, logically structured, or appropriately formatted for human consumption. This is especially problematic for tasks involving technical documentation, policy analysis, or instructional content, where clarity and usability are as important as correctness. Without incorporating interpretability metrics—such as alignment with domain conventions or ease of follow-up—the utility of model outputs in collaborative settings remains under-evaluated.

Together, these blind spots point to a deeper issue: benchmarks that foreground abstract correctness over contextual relevance risk mischaracterizing what it means for a model to be "useful." Expanding evaluation frameworks to include tool coverage, task efficiency, and interpretability would offer a more holistic picture of model capability and better align with user needs.

4.5 Towards Capability–Aligned, Human–Centred Metrics

To close the gap between benchmark performance and user value, we propose constructing bespoke evaluation suites for each of the six identified capabilities. Each suite should be designed around the five criteria of coherence, accuracy, clarity, relevance, and efficiency, with efficiency defined in terms of time or cognitive effort saved. Tasks must permit multi-turn interaction, enabling iterative refinement, and should integrate human raters for any dimension that involves subjective judgement. Smaller, carefully curated datasets, transparent annotation protocols, and public prompt repositories will provide more reliable and actionable signals than the sprawling, opaque corpora that currently dominate the field.

This proposal addresses a growing tension in the field: while alignment methods like RLHF have improved how closely model outputs reflect human preferences [18, 17], evaluation remains largely decoupled from these goals—focusing instead on abstract problem-solving benchmarks that overlook real-world use.

4.6 Interpreting Leaderboards: Recency and Transparency Effects

Following the release of DeepSeek-R1 in January 2025, a wave of competing models was announced in rapid succession—OpenAI, Google, Anthropic, Alibaba, and others each launched models in February and March that narrowly outperformed DeepSeek on various benchmark metrics [73, 74, 77, 79]. Each successive release claimed marginal improvements over its predecessors, often by one or two percentage points, suggesting a pattern of incremental gains driven by awareness of previous scores. This phenomenon highlights what we term a ”recency advantage," where developers of newer models can fine-tune against published benchmarks and adjust hyperparameters or test-time configurations to outperform the latest leader. However, this advantage does not necessarily reflect genuine architectural advances.

Compounding this issue is the lack of transparency in evaluation practices. Many performance results—such as those we analyze in this paper—are self-reported by companies without standardized auditing, making it difficult to determine whether models were evaluated under comparable conditions. For example, it is often unclear whether models are run multiple times with only the best-performing result reported, or whether evaluations were specifically optimized to highlight strengths. This leaderboard arms race incentivizes superficial metric gains rather than meaningful improvements aligned with real-world user needs.

Crowd-sourced evaluations such as the Chatbot Arena address some of these issues by incorporating human-in-the-loop comparisons and being run by an external party. However, the platform still lacks full transparency: the prompt set is not publicly disclosed, rater demographics are unknown, and ratings may be influenced by style preferences or familiarity with certain model behaviors. Moreover, the comparative nature of the judgments—asking which model is better rather than how well a model performs against a defined standard—limits interpretability. Recent research has highlighted additional concerns, such as the ability of certain providers to test multiple model variants before public release and selectively disclose performance results, leading to biased Arena scores due to selective disclosure of performance results. Furthermore, proprietary closed models are sampled at higher rates and have fewer models removed from the arena than open-weight and open-source alternatives, resulting in data access asymmetries that can distort leaderboard rankings [107]. In the absence of an objective gold standard or fine-grained error analysis, leaderboard rankings risk overstating differences while obscuring common failure modes across models.

A further limitation is that many of these evaluations rely on single-turn prompts. Yet in real-world applications, text-to-text models are rarely used in isolation. Instead, they function within multi-turn, iterative workflows—refining outputs, responding to clarification requests, and adapting to changing context. A single prompt evaluation fails to capture these dynamics, leaving critical aspects of interaction quality unmeasured. To move beyond superficial comparisons, we need benchmark designs that integrate objective reference answers, track error margins, and account for the conversational, evolving nature of human–AI collaboration.

4.7 Implications, Limitations, and Future Work

The divergence between benchmark scores and actual user workflows underscores the need to reorient evaluation culture toward human-centred measures of utility. Rather than focusing narrowly on correctness or static tasks, evaluations should consider whether models meaningfully streamline or support the tasks users care about most. Such a shift is essential for informing responsible procurement, deployment, and regulation of generative AI systems—ensuring that advances in model architecture translate into genuine improvements in human productivity and satisfaction.

Although our analysis integrates survey data, real-world usage logs, and benchmark documentation, it is subject to several limitations. First, the process of mapping raw Claude.ai prompts to predefined occupational tasks may oversimplify user intent and overlook nuanced or cross-functional activities that do not fit neatly into O*NET categories. Second, our assessment of benchmark design draws on publicly available documentation; proprietary evaluation protocols and unpublished internal benchmarks may follow different conventions. Third, while we have proposed capability-aligned criteria, we have not yet empirically validated a concrete evaluation suite for any single capability.

Future work should address these limitations by designing, piloting, and refining new benchmarks aligned with the six capabilities and five evaluative criteria proposed in this paper. Each benchmark should adopt a human-in-the-loop methodology. For example, a Reviewing Work benchmark might present professional documents (e.g., emails, reports, or manuscripts) alongside model-generated edits, with expert annotators rating improvements in style, accuracy, and coherence. A Data Structuring benchmark could supply raw reference lists or unformatted tables and ask human judges to assess correctness and efficiency in standard citation formats or tabular layouts. Benchmarks for Summarization, Technical Assistance, Generation, and Information Retrieval should incorporate multi-turn dialogues, domain-specific tasks, and metrics reflecting time saved or cognitive load reduced. By publicly releasing prompts, annotation protocols, and evaluation rubrics, future research can help establish transparent and reproducible standards that better reflect real-world utility and support more human-centered model development.

5 Conclusion

This paper has explored how generative AI is being used in practice and identified a meaningful opportunity to better align evaluation metrics with real-world workflows. While current benchmarks provide useful insights into specific model capabilities—particularly in technical domains—they often focus on narrow, isolated tasks that do not capture the full spectrum of how people interact with AI. By analyzing both survey and usage data, we distilled AI use into six key capabilities: Summarization, Technical Assistance, Reviewing Work, Data Structuring, Generation, and Information Retrieval. These capabilities reflect the diverse, collaborative ways users engage with language models across professional and personal contexts. Rather than proposing a new evaluation framework, our analysis surfaces where existing benchmarks align with these capabilities—and, notably, where they do not. Many capabilities used frequently in real-world tasks, such as Reviewing Work and Data Structuring, currently lack dedicated benchmarks. Even where coverage exists, we find that commonly used metrics often prioritize abstract notions of correctness over qualities that matter to users, such as clarity, efficiency, and contextual relevance. In addition, many benchmarks rely on decontextualized tasks, offering little variation in domain, audience, or user intent. This limits their ability to capture how model performance varies across the rich range of real-world use cases.

To examine these shortcomings, we evaluate current benchmarks through five practical lenses—coherence, accuracy, clarity, relevance, and efficiency—derived from linguistic principles and user expectations. These dimensions are not presented as a new framework, but as tools to reveal blind spots in current practice and identify areas for improvement.

By highlighting these mismatches between benchmark design and user needs, this paper provides a road-map for more human-centered evaluation. This approach supports more meaningful comparisons between models, helps organizations make better-informed decisions, and encourages developers to focus on the capabilities users care about most. As generative AI continues to become embedded in everyday work, improving how we evaluate these systems is essential to ensuring they deliver genuinely helpful experiences.

References

[1] Reuters, “Deepseek rushes to launch new ai model as china goes all in,” *Reuters*, 2025, accessed: 2025-03-04. [Online]. Available: https://www.reuters.com/technology/artificial-intelligence/deepseek-rushes-launch-new-ai-model-china-goes-all-2025-02-25/

[2] BBC News, “China’s deepseek ai reportedly matches openai’s gpt-4o in key tasks,” 2025, accessed: 2025-03-04. [Online]. Available: https://www.bbc.com/news/articles/c5yv5976z9po

[3] T. Guardian, “Diving into deepseek: inside the 7 february guardian weekly,” *The Guardian*, 2025, accessed: 2025-03-04. [Online]. Available: https://www.theguardian.com/news/2025/feb/05/diving-into-deepseek-inside-the-7-february-guardian-weekly

[4] CNBC, “How china’s new ai model deepseek is threatening u.s. dominance,” 2025, accessed: 2025-03-04. [Online]. Available: https://www.cnbc.com/2025/01/24/how-chinas-new-ai-model-deepseek-is-threatening-us-dominance.html

[5] Yahoo Finance UK, “Deepseek: The mysterious chinese ai that’s beating openai?” 2025, accessed: 2025-03-04. [Online]. Available: https://uk.finance.yahoo.com/news/deepseek-mysterious-chinese-ai-becomes-120630369.html

[6] DeepSeek, “Deepseek-coder-v2: Advancements in ai coding capabilities,” *DeepSeek*, 2024, accessed: 2025-03-04. [Online]. Available: https://www.deepseek.com/deepseek-coder-v2

[7] J. Smith and J. Doe, “A comprehensive framework for evaluating multilingual tokenizer quality,” *arXiv preprint arXiv:2410.12989*, 2024, accessed: 2025-03-04. [Online]. Available: https://arxiv.org/abs/2410.12989

[8] C. for AI Safety and S. AI, “Humanity’s last exam,” https://lastexam.ai, 2025, accessed: 2025-03-04.

[9] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang, “Phi-4 technical report,” *arXiv preprint arXiv:2412.08905*, 2024. [Online]. Available: https://arxiv.org/abs/2412.08905

[10] E. Miller, “Adding error bars to evals: A statistical approach to llm evaluations,” *arXiv preprint arXiv:2402.03091*, 2024. [Online]. Available: https://arxiv.org/abs/2402.03091

[11] J. He, J. Du, G. Neubig, Z. Tan, and K. Duh, “On the blind spots of model-based evaluation metrics for text generation,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.871

[12] Y. Wang, W. Chen, S. Chen, C. Xu, Y. Wang, Z. Liu, L. Wang, and M. Huang, “Large language models are not fair evaluators,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. [Online]. Available: https://aclanthology.org/2024.acl-long.152

[13] S. Balloccu, P. Schmidtová, M. Lango, and O. Dusek, “Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 67–93. [Online]. Available: https://aclanthology.org/2024.eacl-long.5/

[14] J. Gallifant, T. Cowen, G. Brockman *et al.*, “Peer review of the gpt-4 technical report,” *PLOS Digital Health*, 2024. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000291

[15] L. Sottana, L. F. R. Ribeiro, and I. Gurevych, “Evaluation metrics in the era of gpt-4: Can we trust reference-based scores?” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. [Online]. Available: https://aclanthology.org/2023.emnlp-main.99

[16] P. Calais, V. Lopes, A. Freire, S. Jariwala, D. Kiyuna, L. Ribeiro *et al.*, “Beyond accuracy: Performance of llms on human exams,” *arXiv preprint arXiv:2403.05004*, 2024. [Online]. Available: https://arxiv.org/abs/2403.05004

[17] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2203.02155

[18] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, T. Henighan, T. Hume, D. Krueger, R. De Haas, B. Scheurer, A. Jones, K. Chen, C. Conerly, S. DasSarma, D. Drain, J. Gonzalez, T. Hennigan, S. Johnston, C. Lovitt, A. Malave, B. Mann, C. McKinnon, V. Mikulik, N. Mu, K. Ndousse, D. Nyarko, E. Perez,

M. Petrov, E. Pfaff, S. Ringer, W. Saunders, B. Shlegeris, J. Uesato, G. Vinay, D. Ziegler, and D. Amodei, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2212.08073

[19] Y. Wang, X. Ma, G. Zhang *et al.*, "Mmlu-pro: A more robust and challenging multi-task language understanding benchmark," *arXiv preprint arXiv:2406.01574*, June 2024, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2406.01574

[20] Vals AI, "Aime benchmark," March 2025, accessed: 2025-04-09. [Online]. Available: https://www.vals.ai/benchmarks/aime-2025-03-26

[21] I. Gabriel, "Artificial intelligence, values and alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411–437, 2020. [Online]. Available: https://doi.org/10.1007/s11023-020-09539-2

[22] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative ai," *Business & Information Systems Engineering*, vol. 66, no. 1, pp. 111–126, 2024. [Online]. Available: https://doi.org/10.1007/s12599-023-00834-7

[23] A. Vaswani, N. Shazeer, and N. e. a. Parmar, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[24] C. Raffel, N. Shazeer, and A. e. a. Roberts, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, 2020.

[25] F. A. Poltronieri and M. Hänska, "Technical images and visual art in the era of artificial intelligence: from gofai to gans," in *Proceedings of the 9th International Conference on Digital and Interactive Arts*, 2019, pp. 1–8.

[26] OpenAI, "Gpt-4 technical report," *arXiv*, 2023.

[27] H. Touvron, T. Lavril, and G. e. a. Izacard, "Llama 3: Open and efficient foundation models," *arXiv*, 2024.

[28] D. AI, "Deepseek v3: Large-scale mixture-of-experts language model," *arXiv*, 2024.

[29] M. Cazzaniga, M. F. Jaumotte, L. Li, M. G. Melina, A. J. Panton, C. Pizzinelli, E. J. Rockall, and M. M. M. Tavares, *Gen-AI: Artificial intelligence and the future of work*. International Monetary Fund, 2024.

[30] M. T. R. Laskar, S. Alqahtani, M. S. Bari, M. Rahman, M. A. M. Khan, H. Khan, I. Jahan, A. Bhuiyan, C. W. Tan, M. R. Parvez *et al.*, "A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 13 785–13 816.

[31] M. Wahde and M. Virgolin, "Conversational agents: Theory and applications," in *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*. World Scientific, 2022, pp. 497–544.

[32] K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, 2023.

[33] S. K. Dam, C. S. Hong, Y. Qiao, and C. Zhang, "A complete survey on llm-based ai chatbots," *arXiv preprint arXiv:2406.16937*, 2024.

[34] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative ai and chatgpt: Applications, challenges, and ai-human collaboration," pp. 277–304, 2023.

[35] I. J. Akpan, Y. M. Kobara, J. Owolabi, A. A. Akpan, and O. F. Offodile, "Conversational and generative artificial intelligence and human–chatbot interaction in education and research," *International Transactions in Operational Research*, vol. 32, no. 3, pp. 1251–1281, 2025.

[36] J. Hu, "Race and gender bias in midjourney," 2023, accessed: 2025-03-17. [Online]. Available: https://medium.com/@hujason/race-and-gender-bias-in-midjourney-c43e92f515f

[37] C. Jeong, "A study on the implementation of generative ai services using an enterprise data-based llm application architecture," *arXiv preprint arXiv:2309.01105*, 2023.

[38] A. Algaba, V. Holst, F. Tori, M. Mobini, B. Verbeken, S. Wenmackers, and V. Ginis, "How deep do large language models internalize scientific literature and citation practices?" *arXiv preprint arXiv:2504.02767*, 2025, accessed: 2025-04-11. [Online]. Available: https://arxiv.org/abs/2504.02767

[39] A. Singla, A. Sukharevsky, L. Yee, M. Chui, and B. Hall, "The state of ai: How organizations are rewiring to capture value," *McKinsey & Company*, 2025. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai/

[40] A. D. Smet, S. Durth, B. Hancock, M. Mugayar-Baldocchi, and A. Reich, “The human side of generative ai: Creating a path to productivity,” *McKinsey Quarterly*, March 2024. [Online]. Available: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-human-side-of-generative-ai-creating-a-path-to-productivity

[41] Edge Delta Team, “AI Adoption by Companies: 5 Statistics You Should Know,” March 2024, accessed: 2025-03-26. [Online]. Available: https://edgedelta.com/company/blog/ai-adoption-by-companies

[42] C. McClain, “Americans’ use of chatgpt is ticking up, but few trust its election information,” *Pew Research Center*, 2024. [Online]. Available: https://www.pewresearch.org/short-reads/2024/03/26/americans-use-of-chatgpt-is-ticking-up-but-few-trust-its-election-information

[43] A. Polyportis, “A longitudinal study on artificial intelligence adoption: Understanding the drivers of chatgpt usage behavior change in higher education,” *Frontiers in Artificial Intelligence*, vol. 6, p. 1324398, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frai.2023.1324398/full

[44] A. not specified, “Nearly 1 in 3 college students have used chatgpt on written assignments,” *Intelligent.com*, 2023. [Online]. Available: https://www.intelligent.com/nearly-1-in-3-college-students-have-used-chatgpt-on-written-assignments/

[45] Stack Overflow, “Ai 2024 stack overflow developer survey,” https://survey.stackoverflow.co/2024/ai, 2024.

[46] A. Humlum and E. Vestergaard, “The unequal adoption of chatgpt exacerbates existing inequalities among workers,” *Proceedings of the National Academy of Sciences*, vol. 122, no. 1, p. e2414972121, 2024. [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.2414972121

[47] M. McTear, *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature, 2022.

[48] P. Limna, T. Kraiwanit, K. Jangjarat, P. Klayklung, and P. Chocksathaporn, “The use of chatgpt in the digital era: Perspectives on chatbot implementation,” *Journal of Applied Learning and Teaching*, vol. 6, no. 1, pp. 64–74, 2023.

[49] M. L. R. Rasmussen, A.-C. Larsen, Y. Subhi, and I. Potapenko, “Artificial intelligence-based chatgpt chatbot responses for patient and parent questions on vernal keratoconjunctivitis,” *Graefe’s Archive for Clinical and Experimental Ophthalmology*, vol. 261, no. 10, pp. 3041–3043, 2023.

[50] M. M. Carlà, G. Gambini, A. Baldascino, F. Giannuzzi, F. Boselli, E. Crincoli, N. C. D’Onofrio, and S. Rizzo, “Exploring ai-chatbots’ capability to suggest surgical planning in ophthalmology: Chatgpt versus google gemini analysis of retinal detachment cases,” *British Journal of Ophthalmology*, vol. 108, no. 10, pp. 1457–1469, 2024.

[51] C. F. Durach and L. Gutierrez, ““hello, this is your ai co-pilot”–operational implications of artificial intelligence chatbots,” *International Journal of Physical Distribution & Logistics Management*, vol. 54, no. 3, pp. 229–246, 2024.

[52] H. P. Grice, “Logic and conversation,” in *Speech acts*. Brill, 1975, pp. 41–58.

[53] ——, “Further notes on logic and conversation,” *Syntax and semantics*, vol. 9, pp. 113–127, 1978.

[54] H. H. Clark and S. E. Brennan, “Grounding in communication,” in *Perspectives on Socially Shared Cognition*, L. B. Resnick, J. M. Levine, and S. D. Teasley, Eds. Washington, DC: American Psychological Association, 1991, pp. 127–149.

[55] Q. Yang and R. Liu, “Understanding the application of utility theory in robotics and artificial intelligence: A survey,” *arXiv preprint arXiv:2306.09445*, 2023.

[56] H. Levesque. (2021) Grice’s conversational maxims. Accessed: March 12, 2025. [Online]. Available: https://levesque.medium.com/grices-conversational-maxims-189370125917

[57] E. Goffman, *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

[58] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *language*, vol. 50, no. 4, pp. 696–735, 1974.

[59] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Books, 2011.

[60] C. R. Berger, M. E. Roloff, and D. R. Ewoldsen, *The handbook of communication science*. Sage, 2010.

[61] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou, R. Goel, S. Yang, and A. Raju, “On evaluating and comparing conversational agents,” in *Proceedings of the NeurIPS 2017 Workshop on Conversational AI*, 2017. [Online]. Available: https://arxiv.org/abs/1801.03625

[62] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 295–305.

[63] J. Euchner, "Generative ai," *Research-Technology Management*, vol. 66, no. 3, pp. 71–74, 2023.

[64] M. Gerlich, "Ai tools in society: Impacts on cognitive offloading and the future of critical thinking," *Societies*, vol. 15, no. 1, p. 6, 2025.

[65] B. Reeves and C. Nass, "The media equation: How people treat computers, television, and new media like real people," *Cambridge, UK*, vol. 10, no. 10, pp. 19–36, 1996.

[66] G. Bansal, V. Chamola, A. Hussain, M. Guizani, and D. Niyato, "Transforming conversations with ai—a comprehensive study of chatgpt," *Cognitive Computation*, vol. 16, no. 5, pp. 2487–2510, 2024.

[67] A. J. Guydish and J. E. F. Tree, "Good conversations: Grounding, convergence, and richness," *New Ideas in Psychology*, vol. 63, p. 100877, 2021.

[68] K. Ishii, M. M. Lyons, and S. A. Carr, "Revisiting media richness theory for today and future," *Human behavior and emerging technologies*, vol. 1, no. 2, pp. 124–131, 2019.

[69] X. Chen, J. Li, and Y. Ye, "A feasibility study for the application of ai-generated conversations in pragmatic analysis," *Journal of Pragmatics*, vol. 223, pp. 14–30, 2024.

[70] S. Gupta and P. H. Tu, *Practical Philosophy Of Ai-assistants, The: An Engineering-humanities Conversation*. World Scientific, 2023.

[71] K. Handa, A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, and D. Ganguli, "Which economic tasks are performed with ai? evidence from millions of claude conversations," *arXiv preprint arXiv:2503.04761*, 2025. [Online]. Available: https://arxiv.org/abs/2503.04761

[72] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[73] OpenAI, "Introducing gpt-4.5," https://openai.com/index/introducing-gpt-4-5/, Feb. 2025, accessed: 2025-04-09.

[74] Google DeepMind, "Gemini 2.5: Our most intelligent ai model," Mar. 2025, accessed: 2025-04-09. [Online]. Available: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/gemini-2-5-pro

[75] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, January 2025, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2501.12948

[76] xAI, "Grok 3 beta — the age of reasoning agents," February 2025, accessed: 2025-04-09. [Online]. Available: https://x.ai/news/grok-3

[77] Qwen Team, "Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model," January 2025, accessed: 2025-04-09. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-max/

[78] Meta AI, "The llama 4 herd: The beginning of a new era of natively multimodal ai innovation," April 2025, accessed: 2025-04-09. [Online]. Available: https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[79] Anthropic, “Claude 3.7 sonnet,” February 2025, accessed: 2025-04-09. [Online]. Available: https://www.anthropic.com/claude/sonnet

[80] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” *arXiv preprint arXiv:2009.03300*, September 2020, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2009.03300

[81] S. Quan, J. Yang, B. Yu, B. Zheng, D. Liu, A. Yang, X. Ren, B. Gao, Y. Miao, Y. Feng, J. Yang, Z. Cui, Y. Fan, Y. Zhang, B. Hui, and J. Lin, “Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings,” *arXiv preprint arXiv:2501.01257*, January 2025, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2501.01257

[82] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, “Gpqa: A graduate-level google-proof q&a benchmark,” *arXiv preprint arXiv:2311.12022*, November 2023, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2311.12022

[83] Princeton University, “Swe-bench verified,” March 2025, accessed: 2025-04-09. [Online]. Available: https://hal.cs.princeton.edu/swebench

[84] A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, C. Saroufim, C. Fry, D. Marcus, D. Kukliansky, G. S. Tomar, J. Swirhun, J. Xing, L. Wang, M. Gurumurthy, M. Aaron, M. Ambar, R. Fellinger, R. Wang, Z. Zhang, S. Goldshtein, and D. Das, “The facts grounding leaderboard: Benchmarking llms’ ability to ground responses to long-form input,” *arXiv preprint arXiv:2501.03200*, January 2025, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2501.03200

[85] K. Vodrahalli, S. Ontanon, N. Tripuraneni, K. Xu, S. Jain, R. Shivanna, J. Hui, N. Dikkala, M. Kazemi, B. Fatemi, R. Anil, E. Dyer, S. Shakeri, R. Vij, H. Mehta, V. Ramasesh, Q. Le, E. Chi, Y. Lu, O. Firat, A. Lazaridou, J.-B. Lespiau, N. Attaluri, and K. Olszewska, “Michelangelo: Long context evaluations beyond haystacks via latent structure queries,” *arXiv preprint arXiv:2409.12640*, September 2024, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2409.12640

[86] OpenAI, “Humaneval: Hand-written evaluation set,” July 2021, accessed: 2025-04-09. [Online]. Available: https://github.com/openai/human-eval

[87] J. Wei, S. Papay, K. Nguyen, H. W. Chung, A. Glaese, J. Schulman, Y. J. Jiao, and W. Fedus, “Measuring short-form factuality in large language models,” *arXiv preprint arXiv:2411.04368*, November 2024, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2411.04368

[88] K. Cobbe, V. Kosaraju, M. Bavarian *et al.*, “Training verifiers to solve math word problems,” *arXiv preprint arXiv:2107.03374*, July 2021, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2107.03374

[89] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan, “tau-bench: A benchmark for tool-agent-user interaction in real-world domains,” *arXiv preprint arXiv:2406.12045*, June 2024, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2406.12045

[90] Vals AI, “Math 500 benchmark,” April 2025, accessed: 2025-04-09. [Online]. Available: https://www.vals.ai/benchmarks/math500-04-07-2025

[91] J. Li, B. Hui, G. Qu *et al.*, “Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls,” *arXiv preprint arXiv:2305.03111*, May 2023, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2305.03111

[92] LiveCodeBench, “Livecodebench: Holistic and contamination-free evaluation of large language models for code,” April 2025, accessed: 2025-04-09. [Online]. Available: https://livecodebench.github.io/

[93] L. Phan, A. Gatti, Z. Han *et al.*, “Humanity’s last exam,” *arXiv preprint arXiv:2501.14249*, January 2025, accessed: 2025-04-09. [Online]. Available: https://arxiv.org/abs/2501.14249

[94] R. Cui *et al.*, “Agieval: A human-centric benchmark for evaluating foundation models,” April 2025, accessed: 2025-04-09. [Online]. Available: https://github.com/ruixiangcui/AGIEval

[95] S. Gu *et al.*, “Cruxeval-i/o: Benchmarking the code reasoning of large language models,” in *Proceedings of Machine Learning Research*, vol. 235, 2024. [Online]. Available: https://proceedings.mlr.press/v235/gu24c.html

[96] Y. Wu, J. Mei, M. Yan *et al.*, “Writingbench: A comprehensive benchmark for generative writing,” *arXiv preprint arXiv:2503.05244*, March 2025. [Online]. Available: https://arxiv.org/abs/2503.05244

[97] A. Vichare, A. N. Angelopoulos, W.-L. Chiang, K. Tang, and L. Manolache, “Webdev arena: A live llm leaderboard for web app development,” 2025.

[98] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, “Chatbot arena: An open platform for evaluating llms by human preference,” 2024.

[99] Google Assistant Help, "Use google assistant to summarize web pages," 2023, accessed: 2025-03-25. [Online]. Available: https://support.google.com/assistant/answer/14163109?hl=en

[100] I. Busu, A.-M. Enescu, H.-R. Enescu, and C. Bîră, "Using chatgpt to write program-space optimized source-code," in *2024 Advanced Topics on Measurement and Simulation (ATOMS)*. IEEE, 2024, pp. 279–282.

[101] R. Rogers, "Generative ai transformed english homework. math is next," *WIRED*, August 2024, accessed: 2025-03-25. [Online]. Available: https://www.wired.com/story/gauth-ai-math-homework-app/

[102] M. Khalifa and M. Albadawy, "Using artificial intelligence in academic writing and research: An essential productivity tool," *Computer Methods and Programs in Biomedicine Update*, p. 100145, 2024.

[103] R. Aleithan, H. Xue, M. M. Mohajer, E. Nnorom, G. Uddin, and S. Wang, "Swe-bench+: Enhanced coding benchmark for llms," *arXiv preprint arXiv:2410.06992*, 2024. [Online]. Available: https://arxiv.org/abs/2410.06992

[104] Y. Huang, Z. Lin, X. Liu, Y. Gong, S. Lu, F. Lei, Y. Liang, Y. Shen, C. Lin, N. Duan, and W. Chen, "Competition-level problems are effective llm evaluators," *arXiv preprint arXiv:2312.02143*, 2023. [Online]. Available: https://arxiv.org/abs/2312.02143

[105] J. Vendrow, E. Vendrow, S. Beery, and A. Madry, "Do large language model benchmarks test reliability?" *arXiv preprint arXiv:2502.03461*, 2025, available at https://arxiv.org/abs/2502.03461.

[106] M. Gašić, V. Rieser, and M. Walker, "How to do human evaluation: A brief introduction to user studies in nlp," *Natural Language Engineering*, vol. 28, no. 4, pp. 485–504, 2022. [Online]. Available: https://doi.org/10.1017/S135132492200025X

[107] S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. Smith *et al.*, "The leaderboard illusion," *arXiv preprint arXiv:2504.20879*, 2025. [Online]. Available: https://arxiv.org/abs/2504.20879

A Appendix: Capability First Analysis

Table A1: Mapping of 21 high-exposure occupational tasks to AI capabilities based on qualitative coding of tasks rated as offering large time savings through ChatGPT assistance. Tasks span 11 occupations and reflect the cognitive processes and outputs best supported by generative AI.

Occupation	Job Task	Explanation	Capability
Accountants & Auditors	Prepare detailed reports on audit findings.	ChatGPT can draft and structure reports based on audit findings.	Generation, Data Structuring
Accountants & Auditors	Examine and evaluate financial and information systems, recommending controls to ensure system reliability and data integrity.	ChatGPT can compile reports on financial and IT systems from provided data and text, identifying potential issues and suggesting controls for system reliability and data integrity.	Summarization, Reviewing Work
Accountants & Auditors	Prepare, examine, or analyze accounting records, financial statements, or other financial reports to assess accuracy, completeness, and conformance to reporting and procedural standards.	ChatGPT can draft accounting documents and analyze accounting information and financial reports.	Generation
Accountants & Auditors	Compute taxes owed and prepare tax returns, ensuring compliance with payment, reporting, or other tax requirements.	ChatGPT can provide guidance on tax legislation, calculate tax liabilities, and generate drafts of tax returns.	Information Retrieval
Customer Service Rep.	Keep records of customer interactions or transactions, recording details of inquiries, complaints, or comments, as well as actions taken.	ChatGPT can assist with logging and reporting customer contacts based on data from customer support.	Data Structuring

Occupation	Job Task	Explanation	Capability
Customer Service Rep.	Check to ensure that appropriate changes were made to resolve customersâ€™ problems.	ChatGPT can prepare a structured report on whether the actions taken resolved the customer complaint.	Summarization, Generation
Customer Service Rep.	Contact customers to respond to inquiries or to notify them of claim investigation results or any planned adjustments.	ChatGPT can suggest responses to customer inquiries and complaints.	Generation
Financial Advisors	Recommend to clients strategies in cash management, insurance coverage, investment planning, or other areas to help them achieve their financial goals.	ChatGPT can develop and formulate financial strategies and plans based on a client's financial situation and goals.	Generation
Financial Advisors	Implement financial planning recommendations, or refer clients to someone who can assist them with plan implementation.	ChatGPT can provide step-by-step instructions for implementing a client's financial plan and suggest agents who can assist with the implementation.	Generation
Financial Advisors	Analyze financial information obtained from clients to determine strategies for meeting clientsâ€™ financial objectives.	ChatGPT can suggest and describe suitable financial strategies based on clientsâ€™ financial situations and goals.	Generation
Financial Advisors	Answer clientsâ€™ questions about the purposes and details of financial plans and strategies.	ChatGPT can suggest answers to typical questions about financial plans and strategies.	Information Retrieval
HR Professionals	Inform job applicants of details such as duties and responsibilities, compensation, benefits, schedules, working conditions, or promotion opportunities.	ChatGPT can generate detailed descriptions of jobs and employment terms.	Summarization
HR Professionals	Interpret and explain human resources policies, procedures, laws, standards, or regulations.	ChatGPT can interpret and explain complex HR policies and regulations in easily accessible language.	Summarization
IT Support	Answer user inquiries regarding computer software or hardware operation to resolve problems.	ChatGPT can provide step-by-step instructions for solving typical hardware and software problems.	Technical Assistance
IT Support	Read technical manuals, confer with users, or conduct computer diagnostics to investigate and resolve problems or to provide technical assistance and support.	ChatGPT can summarize technical manuals and assist with technical support by suggesting questions to users and possible solutions.	Summarization, Technical Assistance
IT Support	Maintain records of daily data communication transactions, problems and remedial actions taken, or installation activities.	ChatGPT can structure log files and notes into coherent reports.	Summarization
Journalists	Write commentaries, columns, or scripts.	ChatGPT can generate drafts, suggest changes, and provide ideas for articles, etc.	Generation, Reviewing Work

Occupation	Job Task	Explanation	Capability
Journalists	Examine news items of local, national, and international significance to determine topics to address, or obtain assignments from editorial staff members.	ChatGPT can analyze and summarize news content and suggest topics to cover.	Summarization
Journalists	Analyze and interpret news and information received from various sources to broadcast the information.	ChatGPT can analyze, summarize, and translate news from various sources.	Summarization, Information Retrieval
Legal Professionals	Prepare affidavits or other documents, such as legal correspondence, and organize and maintain documents in paper or electronic filing system.	ChatGPT can suggest templates and drafts for legal documents and provide guidance on filing.	Generation, Information Retrieval
Legal Professionals	Prepare legal documents, including briefs, pleadings, appeals, wills, contracts, and real estate closing statements.	ChatGPT can deliver drafts of legal documents based on entered details.	Generation
Marketing Professionals	Prepare reports of findings, illustrating data graphically and translating complex findings into written text.	ChatGPT can write and structure reports from data and text, and can also suggest presentation forms for data.	Generation
Office Clerks	Communicate with customers, employees, and other individuals to answer questions, disseminate or explain information, take orders, and address complaints.	ChatGPT can suggest responses to typical inquiries, complaints, and orders.	Generation
Office Clerks	Compile, copy, sort, and file records of office activities, business transactions, and other activities.	ChatGPT can prepare records following complex instructions and assist with filing and sorting documents by summarizing and editing text.	Reviewing Work, Summarization
Office Clerks	Compute, record, and proofread data and other information, such as records or reports.	ChatGPT can prepare and check records and reports based on predefined guidelines.	Generation
Software Developers	Write, analyze, review, and rewrite programs, using workflow chart and diagram, and applying knowledge of computer capabilities, subject matter, and symbolic logic.	ChatGPT can assist with writing code and analyzing errors in programs based on software developers' preferences and program outputs.	Technical Assistance
Software Developers	Correct errors by making appropriate changes and rechecking the program to ensure that the desired results are produced.	ChatGPT can identify code errors and suggest corrections and checks based on error messages and other program outputs.	Technical Assistance
Software Developers	Perform or direct revision, repair, or expansion of existing programs to increase operating efficiency or adapt to new requirements.	ChatGPT can provide code suggestions for auditing, debugging, and extending programs, and can also suggest ways to optimize the code.	Technical Assistance, Information Retrieval

Occupation	Job Task	Explanation	Capability
Software Developers	Conduct trial runs of programs and software applications to be sure they will produce the desired information and that the instructions are correct.	ChatGPT can suggest code changes and debug programs, as well as explain program output in a reader-friendly format.	Technical Assistance, Summarization
Software Developers	Consult with and assist computer operators or system analysts to define and resolve problems in running computer programs.	ChatGPT can identify code errors and suggest corrections based on error messages, program output, and user input.	Technical Assistance
Teachers	Adapt teaching methods and instructional materials to meet studentsâ€™ varying needs and interests.	ChatGPT can tailor teaching methods and materials based on each student's learning style and interests.	Generation
Teachers	Prepare objectives and outlines for courses of study, following curriculum guidelines or requirements of states and schools.	ChatGPT can suggest and structure learning objectives and courses in accordance with curricula or similar requirements.	Generation

B Appendix: Metric Evaluation

Table B2: Comparison of AI Benchmarks by Capability and 5 Evaluation Criteria

Metric	Capability	Coherence	Accuracy	Clarity	Relevance	Efficiency
MMLU	Information Retrieval	Questions follow a multiple-choice format but lack natural dialogue or conversational flow	6.5% of questions contain incorrect answers	Format is rigid; only exact answer options are considered correct	Focused on subjects common in the US education system	Not explicitly measured
AIME	Information Retrieval	Questions ask for one-off numerical answers, with little explanation or reasoning	Based on official math competition questions	Answers are expected to be exact numbers	Covers only mathematics problems	Measured indirectly through response time
Codeforces	Technical Assistance	Tasks are described in plain English programming problems	Performance drops for questions written after the model's training period	Output is executable code that can be used by a developer	All tasks are code-related, most suited for C++ and Python	Time to solve is measured
GPQA	Information Retrieval	Standard multiple-choice questions without much context	36% of questions have disagreement over the correct answer	Exact formatting of answers is required	Covers only biology, chemistry, and physics	Not explicitly measured
SWE-bench	Technical Assistance	Based on real bug reports from software projects	32% of solutions used information found in public comments; 31% had questionable test cases	Requires the model to suggest a working patch that passes automated tests	Focuses entirely on Python software development	Not explicitly measured

Metric	Capability	Coherence	Accuracy	Clarity	Relevance	Efficiency
FACTS Grounding	Summarization	Large text passages are followed by detailed factual questions	Model answers are rated by a different model or human reviewer, with 62–85% agreement	Answers must be understandable by other models or humans	Based on dense material such as law, medicine, or STEM writing	Not explicitly measured
MRCR	Summarization	Tasks include similar sentences where models must retrieve the correct one	Measures whether the correct sentence is chosen from among distractors	Emphasizes accurate retrieval over natural phrasing	Based on synthetic data but intended to reflect real-world understanding	Not explicitly measured
HumanEval	Technical Assistance	Each task includes a function description in plain English	Model is judged on whether its function passes test cases	Code does not need to be clear, only correct	All tasks are in Python	Not explicitly measured
SimpleQA	Information Retrieval	Short factual questions that expect short factual answers	94.4% agreement between original reviewers and a third reviewer	Prioritizes clear, correct, concise answers	Covers diverse topics from science to culture	Not explicitly measured
GSM8K	Information Retrieval	Math problems are written in plain language and require multi-step reasoning	About 50% of questions were incorrect in early versions; improved in later release	Focus is on producing the correct final answer, not explanation quality	Covers elementary and middle school mathematics	Not explicitly measured
Math-500	Information Retrieval	Involves complex math questions requiring multi-step logic	No known major errors in question set	Explanation quality is not rated	Focused entirely on mathematics	Not explicitly measured
Bird-SQL	Technical Assistance	Natural language questions must be converted into SQL code	Some inconsistencies in dataset, but no published error rate	No penalties for unclear output	Covers a wide range of database queries	Measures efficiency using a custom score
LiveCodeBench	Technical Assistance	Coding tasks from popular programming websites	Only uses tasks published after model training period	Focus is on whether code runs correctly, not how it's written	All problems are Python-based	Not explicitly measured
Humanity's Last Exam	Information Retrieval	Combination of multiple-choice and short-answer questions	Written by experts, including PhDs	Short answers must match exactly to count as correct	Wide coverage, including niche historical and technical content	Not explicitly measured
MMLU-Pro	Information Retrieval	Similar to MMLU but with more answer options per question	Reviewed for robustness with revised questions	Same rigid answer format as MMLU	Broader focus beyond US education	Not explicitly measured
AGIEval	Information Retrieval	Objective questions only; no open-ended responses	Based on official entrance exams like SAT and LSAT	Only exact matches are scored as correct	Covers diverse academic entrance exams	Not explicitly measured
CRUXEval	Technical Assistance	Models must predict input or output of a Python function	Functions adapted from tutorials	Clarity not judged; correctness is all that matters	Functions are short and educational	Not explicitly measured
WritingBench	Generation	Models generate full passages in various writing styles	Judged by a trained model validated against human scores	Penalizes vague or incoherent writing	Includes creative, persuasive, technical, and informative writing	Not explicitly measured

Metric	Capability	Coherence	Accuracy	Clarity	Relevance	Efficiency
WebDev Arena	Technical Assistance	Prompts ask models to build entire websites or interfaces	No gold standard; judged by human preferences	Clarity depends on how useful output is to humans	Tasks involve real-world web development problems	Efficiency inferred from human preferences
Chatbot Arena	Generation	Creative prompts (e.g., poems, stories) are given to models	Responses compared by direct human voting	Writing quality judged indirectly via preference	Focused on imaginative, creative writing	Not explicitly measured

C Appendix: Prompts for all applicable capabilities

Table C3: Top 100 tasks given by users or Anthropic, showing the aligned AI capability, example prompts, and the percentage of prompts associated with each task. Tasks are sorted by frequency, with the highest-ranked tasks appearing first.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Modify existing software to correct errors, to adapt it to new hardware, or to upgrade interfaces and improve performance.	Software developers, systems software	4.79	1	Technical Assistance, Reviewing Work	• My Python program runs really slowly when i open a large file. How can i optimise it? • Can you check this c++ code for compatibility with newer hardware?
Correct errors by making appropriate changes and rechecking the program to ensure that the desired results are produced.	Computer programmers	3.0	2	Reviewing Work, Technical Assistance	• Can you help me debug this for loop? It's not producing the correct output. • I'm getting a 'null pointer exception' in Java. How do i fix it?
Modify existing software to correct errors, allow it to adapt to new hardware, or to improve its performance.	Software developers, applications	2.52	3	Reviewing Work, Technical Assistance	• Identify performance bottlenecks in my TensorFlow code. • Update my app to run efficiently on M1 Macs
Perform initial debugging procedures by reviewing configuration files, logs, or code pieces to determine breakdown source.	Software quality assurance engineers and testers	2.01	4	Technical Assistance, Reviewing Work	• Can you find common causes for 'segfault' in c programs? • Here's my server log — what caused the crash?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Perform routine system administrative functions such as troubleshooting, back-ups, and upgrades.	Bioinformatics technicians	1.81	5	Technical Assistance	• How do I automate backups of my Linux server every night?
Diagnose, troubleshoot, and resolve hardware, software, or other network and system problems, and replace defective components when necessary.	Network and computer systems administrators	1.56	6	Technical Assistance	• My wi-fi disconnects randomly. What are some possible causes?
Write new programs or modify existing programs to meet customer requirements, using current programming languages and technologies.	Data warehousing specialists	1.46	7	Technical Assistance	• Write SQL code to create a sales summary table by region.
Select and edit documents for publication and display, applying knowledge of subject, literary expression, and presentation techniques.	Archivists	1.41	8	Reviewing Work, Generation, Summarization	• Can you edit this paragraph to make it more formal and clear? • Write an exhibition description for this archived letter • Summarise this historical document in 100 words.
Review and analyze computer printouts and performance indicators to locate code problems, and correct errors by correcting codes.	Computer systems analysts	1.35	9	Reviewing Work, Technical Assistance	• Here's the output from my script. Why is it failing? • My server report shows high cpu usage at midnight. What could be the cause?
Write, analyze, review, and rewrite programs, using workflow chart and diagram, and applying knowledge of computer capabilities, subject matter, and symbolic logic.	Computer programmers	1.24	10	Reviewing Work, Technical Assistance	• Review this flowchart and suggest improvements to the algorithm. • Generate a Python program that implements a binary search tree.
Design, build, or maintain web sites, using authoring or scripting languages, content creation tools, management tools, and digital media.	Web developers	1.19	11	Technical Assistance	• Help me write javascript to make this image carousel auto-scroll.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Write, update, and maintain computer programs or software packages to handle specific jobs such as tracking inventory, storing or retrieving data, or controlling other equipment.	Computer programmers	1.15	12	Technical Assistance	• Write a Python script that tracks inventory in a csv file.
Prepare, rewrite and edit copy to improve readability, or supervise others who do this work.	Editors	1.0	13	Reviewing Work, Generation	• Edit this paragraph to be more concise and academic. • Write an article to appeal to high school readers.
Determine sources of web page or server problems, and take action to correct such problems.	Web administrators	0.89	14	Reviewing Work, Technical Assistance, Information Retrieval	• Look at these logs, can you tell what's wrong with my apache config? • My website is showing a 500 error. What could be causing it? • What are common causes of slow server load times?
Edit, standardize, or make changes to material prepared by other writers or establishment personnel.	Technical writers	0.85	15	Reviewing Work, Summarization	• Edit this user manual section to follow our company style guide. • Summarise the key points of this technical spec document.
Review class material with students by discussing text, working solutions to problems, or Reviewing Worksheets or other assignments.	Tutors	0.79	16	Generation, Summarization	• Create a worksheet that will test year 7 students english ability • Summarise this chapter on plant biology for year 10 students.
Modify existing programs to enhance efficiency.	Computer numerically controlled machine tool programmers, metal and plastic	0.78	17	Technical Assistance	• Optimise this g-code script for faster cnc milling
Edit or rewrite existing copy as necessary, and submit copy for approval by supervisor.	Copy writers	0.76	18	Reviewing Work	• Check this sales pitch paragraph for tone and clarity.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Develop or apply data mining and machine learning algorithms.	Bioinformatics technicians	0.63	19	Technical Assistance	• Help me write a code for k-means clustering.
Write supporting code for web applications or web sites.	Web developers	0.6	20	Technical Assistance	• Write javascript code to validate a contact form
Organize material and complete writing assignment according to set standards regarding order, clarity, conciseness, style, and terminology.	Technical writers	0.59	21	Generation, Reviewing Work	• Write a concise summary for the introduction of this user manual. • Can you revise this technical paragraph to improve clarity?
Modify existing databases and database management systems or direct programmers and analysts to make changes.	Database administrators	0.55	22	Technical Assistance	• How do i update all customer emails in a PostgreSQL database?
Provide private instruction to individual or small groups of students to improve academic performance, improve occupational skills, or prepare for academic or occupational tests.	Tutors	0.54	23	Reviewing Work, Generation, Summarization	• Check my short essay on climate change for grammar and structure. • Write a lesson plan for a student with the following needs • Summarise the key points of this shakespeare sonnet.
Guide clients in the development of skills or strategies for dealing with their problems.	Mental health counselors	0.53	24	Generation	• Can you generate daily affirmations for anxiety?
Write original or adapted material for dramas, comedies, puppet shows, narration, or other performances.	Actors	0.5	25	Generation	• Write a comedic monologue about being stuck in traffic.
Write advertising copy for use by publication, broadcast, or internet media to promote the sale of goods and services.	Copy writers	0.49	26	Generation, Reviewing Work	• Create a product description for a smart water bottle • Can you rewrite this ad copy to make it more persuasive?
None		0.48	27		

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Read from sacred texts such as the bible, torah, or koran.	Clergy	0.46	28	Summarization, Information Retrieval	• Summarise the parable of the good samaritan. • What is the historical background of the torah?
Analyze information to determine, recommend, and plan layout, including type of computers and peripheral equipment modifications.	Computer hardware engineers	0.43	29	Technical Assistance	• How do i plan a workstation layout for high-throughput computing?
Compute dimensions, areas, volumes, and weights.	Patternmakers, wood	0.42	30	Information Retrieval	• Calculate the volume of a cylinder with radius 10cm and height 25cm.
Conduct logical analyses of business, scientific, engineering, and other technical problems, formulating mathematical models of problems for solution by computers.	Computer and information research scientists	0.41	31	Technical Assistance, Information Retrieval	• Help me write a simulation model for disease spread using Python. • Please explain the travelling salesman problem and how it relates to graph theory
Prepare and deliver lectures to undergraduate or graduate students on topics such as how to speak and write a foreign language and the cultural aspects of areas where a particular language is used.	Foreign language and literature teachers, postsecondary	0.38	32	Information Retrieval, Generation	• What are common words i should know before visiting korea? • Create a lesson plan about chinese idioms.
Develop factors such as themes, plots, characterizations, psychological analyses, historical environments, action, and dialogue, to create material.	Poets, lyricists and creative writers	0.37	33	Generation, Reviewing Work, Summarization	• Write a poem about lost time using abab rhyme • Review this character sketch and suggest improvements. • Summarise the main themes of hamlet
Analyze system performance or operational requirements.	Photonics engineers	0.36	34	Reviewing Work, Technical Assistance	• Evaluate this optical system for energy loss. • How can i improve the performance of a photonic integrated circuit?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Read written materials, such as legal documents, scientific works, or news reports, and rewrite material into specified languages.	Interpreters and translators	0.35	35	Information Retrieval, Generation	• Translate this paragraph into spanish • What's the difference between 'rapport' and 'relationship' in spanish?
Analyze organizational, occupational, and industrial data to facilitate organizational functions and provide technical information to business, industry, and government.	Compensation, benefits, and job analysis specialists	0.35	36	Data Structuring, Summarization	• Convert these notes on who reports to who, into an org chart • Summarise the key compensation trends in the tech industry.
Review software documentation to ensure technical accuracy, compliance, or completeness, or to mitigate risks.	Software quality assurance engineers and testers	0.34	37	Technical Assistance	• Check this API documentation for clarity and completeness.
Break systems into their component parts, assign numerical values to each component, and examine the mathematical relationships between them.	Operations research analysts	0.33	38	Information Retrieval, Technical Assistance	• What are standard methods for queuing theory analysis? • Model this supply chain problem using linear programming.
Write, design, or edit web page content, or direct others producing content.	Web developers	0.32	39	Generation, Reviewing Work	• Create content for an about us page for a small bakery. • Check this blog draft for grammar and tone.
Compile and write documentation of program development and subsequent revisions, inserting comments in the coded instructions so others can understand the program.	Computer programmers	0.31	40	Technical Assistance, Reviewing Work	• Write docstrings for this Python function. • Check if these comments clearly explain what the code does.
Use models to simulate the behavior of animated objects in the finished sequence.	Multimedia artists and animators	0.31	41	Technical Assistance	• Write code to simulate realistic movement for a cloth blowing in the wind.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Observe and evaluate students' work to determine progress, provide feedback, and make suggestions for improvement.	Vocational education teachers, postsecondary	0.31	42	Reviewing Work, Summarization	• Evaluate this student's woodworking project based on provided rubric. • Summarise overall student progress from these weekly reports.
Analyze user needs and software requirements to determine feasibility of design within time and cost constraints.	Software developers, applications	0.3	43	Technical Assistance, Information Retrieval	• How do i assess whether a feature request can be implemented in under 2 weeks? • What features do users prioritise in this type of software.
Determine supplementary virtual features, such as currency, item catalog, menu design, and audio direction.	Video game designers	0.3	44	Generation, Information Retrieval	• Design a virtual currency system that discourages hoarding. • What are ways successful video game franchises have organised their menu layout to be less annoying for the players?
Store, retrieve, and manipulate data for analysis of system capabilities and requirements.	Software developers, applications	0.29	45	Technical Assistance	• Write Python code to load and clean JSON data
Interact with clients to assist them in gaining insight, defining goals, and planning action to achieve effective personal, social, educational, and vocational development and adjustment.	Clinical psychologists	0.29	46	Generation	• Suggest 3 goals for a cbt therapy plan for anxiety.
Test machines, components, materials, or products to determine characteristics such as performance, strength, or response to stress.	Mechanical engineering technologists	0.28	47	Technical Assistance, Information Retrieval	• Generate a script to log temperature and pressure test data. • How do i measure tensile strength in composite materials?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Compile reports, charts, or graphs that describe and interpret findings of analyses.	Statistical assistants	0.27	48	Summarization, Generation, Technical Assistance	• Summarise this dataset's key trends in 3 bullet points. • Generate a short paragraph for the executive summary section of this report. • Create a graph of age distribution from this csv
Compose letters in reply to correspondence concerning such items as requests for merchandise, damage claims, credit information requests, delinquent accounts, incorrect billing, or unsatisfactory service.	Correspondence clerks	0.26	49	Reviewing Work, Generation	• Check this response letter for tone and grammar. • Write a polite response to a customer requesting a refund.
Design databases to support business applications, ensuring system scalability, security, performance and reliability.	Database architects	0.25	50	Information Retrieval, Technical Assistance	• What best practices should I follow to ensure high performance and reliability in a cloud-based SQL database used by a financial application? • Can you help me model a normalized relational database for a hospital management system
Confer with clients regarding the nature of the information processing or computation needs a computer program is to address.	Computer systems analysts	0.25	51	Technical Assistance	• Help me evaluate client requirements and translate them into specs.
Reformat documents, moving paragraphs or columns.	Word processors and typists	0.25	52	Data Structuring	• Move all paragraphs starting with 'summary' to the end of the document
Consult with multiple stakeholders to define requirements and implement online features.	Video game designers	0.24	53	Generation, Technical Assistance	• Suggest online features for a multiplayer puzzle game. • How do I implement a player ranking system in unity?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Test programs or databases, correct errors, and make necessary modifications.	Database administrators	0.24	54	Reviewing Work, Technical Assistance	• Review this database update script for problems. • Write code that will help me run tests on my database to determine any vulnerabilities
Provide technical guidance or support for the development or troubleshooting of systems.	Computer systems engineers/architects	0.24	55	Technical Assistance, Information Retrieval	• Help me design an API server that won't crash under load? • What are common causes of memory leaks in node.js?
Analyze operations to evaluate performance of a company or its staff in meeting objectives or to determine areas of potential cost reduction, program improvement, or policy change.	Chief executives	0.24	56	Summarization, Reviewing Work	• Summarize the key weaknesses in our quarterly performance report. • Evaluate this staff productivity data and suggest improvements.
Test, maintain, and monitor computer programs and systems, including coordinating the installation of computer programs and systems.	Computer systems analysts	0.24	57	Technical Assistance	• How do I monitor server health in real-time?
Explain general financial topics to clients, such as credit report ratings, bankruptcy laws, consumer protection laws, wage attachments, or collection actions.	Credit counselors	0.23	58	Information Retrieval	• What does a credit score of 620 mean?
Instruct individuals in career development techniques such as job search and application strategies, resume writing, and interview skills.	Educational, guidance, school, and vocational counselors	0.23	59	Generation, Reviewing Work	• Generate 5 interview questions for a software developer role. • Review this resume and suggest changes

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Analyze organic or inorganic compounds to determine chemical or physical properties, composition, structure, relationships, or reactions, using chromatography, spectroscopy, or spectrophotometry techniques.	Chemists	0.22	60	Information Retrieval, Summarization	• What is the procedure for running a uv-vis spectroscopy test? • Summarise the findings of this report
Plan, evaluate, and revise curricula, course content, course materials, and methods of instruction.	Area, ethnic, and cultural studies teachers, postsecondary	0.22	61	Summarization, Generation, Information Retrieval, Reviewing Work	• Summarise the objectives of this course in 3 bullet points. • Generate lesson plans for a 6-week course on indigenous literatures. • Explain the effects of colonisation on different cultures. • Evaluate this syllabus for coverage of key asian american history topics.
Consult with and assist computer operators or system analysts to define and resolve problems in running computer programs.	Computer programmers	0.22	62	Technical Assistance, Reviewing Work	• Generate Python code that is able to handle memory leaks. • Read through my code and identify any issues with it that may cause issues
Prepare or edit organizational publications, such as employee newsletters or stockholders' reports, for internal or external audiences.	Public relations specialists	0.21	63	Generation, Reviewing Work, Summarization, Data Structuring	• Write a newsletter update on our recent merger • Review this shareholder report draft for clarity. • Summarise key takeaways from our annual report for employees.
Develop, implement, modify, and document recordkeeping and accounting systems, making use of current computer technology.	Accountants	0.21	64	Data Structuring, Technical Assistance	• Format these notes into a table that i can paste into excel. • How do i implement gaap-compliant reports in quickbooks?"

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Generate standard or custom reports summarizing business, financial, or economic data for review by executives, managers, clients, and other stakeholders.	Business intelligence analysts	0.2	65	Summarization, Data Structuring	• Summarise the findings from this financial paper • Take the data from this spreadsheet and put it into paragraph form
Participate in the work of subordinates to facilitate productivity or to overcome difficult aspects of work.	First-line supervisors of office and administrative support workers	0.2	66	Reviewing Work, Technical Assistance, Generation	• Review this task workflow and suggest improvements for speed. • What are some ways to automate excel data entry tasks? • What are some ways to imrpove office productivity?
Interpret and analyze policies, public issues, legislation, or the operations of governments, businesses, and organizations.	Political scientists	0.2	67	Summarization, Reviewing Work, Information Retrieval	• Summarise the main arguments in this policy proposal • Review this op-ed on healthcare reform for bias. • What does the eu ai act regulate?
Monitor database performance and perform any necessary maintenance, upgrades, or repairs.	Bioinformatics technicians	0.2	68	Technical Assistance	• How do i write a PostgreSQL that will give me the data with the quickest downtime?
Address the relationships of quantities, magnitudes, and forms through the use of numbers and symbols.	Mathematicians	0.2	69	Information Retrieval	• What's the difference between eigenvalues and singular values?
Design, develop and modify software systems, using scientific analysis and mathematical models to predict and measure outcome and consequences of design.	Software developers, applications	0.19	70	Technical Assistance	• Model a queueing system in Python using discrete event simulation.
Write articles, manuals, and other publications, and assist in the distribution of promotional literature about facilities and programs.	Education administrators, elementary and secondary school	0.19	71	Generation	• Write a brochure paragraph about our new after school care program

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Review instructional content, methods, and student evaluations to assess strengths and weaknesses, and to develop recommendations for course revision, development, or elimination.	Adult basic and secondary education and literacy teachers and instructors	0.19	72	Reviewing Work, Summarization	• Evaluate this adult literacy curriculum for engagement. • Summarise the curriculum for this adult education class
Select material most pertinent to presentation, and organize this material into appropriate formats.	Broadcast news analysts	0.19	73	Summarization	• Summarise this press release into a 2-minute news script
Collaborate with system architects, software architects, design analysts, and others to understand business or industry requirements.	Database architects	0.19	74	Information Retrieval, Technical Assistance	• What data model suits a large-scale medical records system? • Help me define schema constraints for multi-user access.
Review published materials and recommend revisions or changes in scope, format, content, and methods of reproduction and binding.	Technical writers	0.18	75	Reviewing Work, Generation	• Check this user guide for outdated references. • Write installation steps for the new software version
Design database applications, such as interfaces, data transfer mechanisms, global temporary tables, data partitions, and function-based indexes to enable efficient access of the generic database structure.	Database architects	0.18	76	Technical Assistance	• Write code that uses function-based indexing to improve query speed.
Conduct materials test and analysis using tools and equipment and applying engineering knowledge.	Civil engineering technicians	0.18	77	Information Retrieval	• What astm standards apply to asphalt testing?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Use mediation techniques to facilitate communication between disputants, to further parties' understanding of different perspectives, and to guide parties toward mutual agreement.	Arbitrators, mediators, and conciliators	0.18	78	Information Retrieval, Generation, Reviewing Work	• What are some evidence backed ways to ensure a successful mediation • Generate 3 open-ended questions to encourage dialogue between parties. • Evaluate this mediation script for balance and neutrality
Organize information for publication and for other means of dissemination, such as use in cd-roms or internet sites.	Historians	0.18	79	Summarization, Data Structuring	• Summarise this archive record • Organise this list of events in the cold war into a chronological order.
Prepare and deliver lectures to undergraduate or graduate students on topics such as anatomy, therapeutic recreation, and conditioning theory.	Recreation and fitness studies teachers, postsecondary	0.17	80	Information Retrieval, Generation	• What are the key physiological effects of interval training. • Write lecture notes on muscle anatomy for undergraduate students.
Plan parties or other special events and services.	Hosts and hostesses, restaurant, lounge, and coffee shop	0.17	81	Generation	• Generate a birthday party schedule for a 10-year-old with a superhero theme.
Write programs in the language of a machine's controller and store programs on media such as punch tapes, magnetic tapes, or disks.	Computer numerically controlled machine tool programmers, metal and plastic	0.17	82	Technical Assistance	• Help me write g-code for drilling 5 holes evenly spaced along a 10 cm line.
Write articles, bulletins, sales letters, speeches, and other related informative, marketing and promotional material.	Copy writers	0.17	83	Generation	• Write a sales letter for a new eco-friendly detergent.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Draw conclusions or make predictions based on data summaries or statistical analyses.	Biostatisticians	0.17	84	Summarization, Technical Assistance, Information Retrieval	• Summarise key findings from this results section from clinical trial dataset. • Write code in r that will run a logistic regression on this dataset to predict disease occurrence. • What does a p-value of 0.01 mean in this context?
Assist students who need extra help with their coursework outside of class.	Biological science teachers, postsecondary	0.17	85	Reviewing Work, Information Retrieval	• Check my lab report for accuracy • Explain how photosynthesis works in 2–3 sentences.
Analyze business operations, trends, costs, revenues, financial commitments, and obligations to project future revenues and expenses or to provide advice.	Accountants	0.17	86	Reviewing Work, Data Structuring	• Review my conclusions of this financial report and make sure they follow a logical conclusion • Take my cells in in excel and convert the text in them into a format that is appropriate for a word document
Document findings of study and prepare recommendations for implementation of new systems, procedures, or organizational changes.	Management analysts	0.16	87	Summarization, Reviewing Work, Generation	• Summarise the issues identified in this customer service audit. • Evaluate this draft proposal for restructuring. • Generate a list of actionable steps to improve efficiency.
Prepare reports, manuscripts, proposals, and technical manuals for use by other scientists and requestors, such as sponsors and customers.	Materials scientists	0.16	88	Generation, Reviewing Work	• Draft the methods section of a technical report on alloy testing. • Check this abstract for clarity and conciseness.

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Evaluate sources of information to determine any limitations in terms of reliability or usability.	Statisticians	0.16	89	Reviewing Work, Information Retrieval	• Evaluate the conclusion of this report based on the results section • How do i determine the sample size required to measure a population?
Help programmers and systems analysts test and debug new programs.	Computer operators	0.16	90	Technical Assistance, Reviewing Work	• Write a a program that will debug my Java code • Review this error log and suggest fixes.
Provide assistance to students in college writing centers.	English language and literature teachers, postsecondary	0.16	91	Generation, Reviewing Work	• Write a conclusion paragraph for an essay on climate change. • Check this essay introduction for clarity and flow.
Answer customers' questions, and provide information on procedures or policies.	Cashiers	0.16	92	Information Retrieval	• What's this store's return policy on electronics?
Create, analyze, report, convert, or transfer data, using specialized applications program software.	Geospatial information scientists and technologists	0.16	93	Data Structuring, Technical Assistance	• Organise satellite imagery by resolution and date. • Convert these shapefiles to geojson format
Develop briefings, brochures, multimedia presentations, web pages, promotional products, technical illustrations, and computer artwork for use in products, technical manuals, literature, newsletters, and slide shows.	Multimedia artists and animators	0.16	94	Generation	• Create ideas for how i could design a visual diagram of a solar panel.
Read manuals, periodicals, and technical reports to learn how to develop programs that meet staff and user requirements.	Computer systems analysts	0.16	95	Summarization, Information Retrieval	• Summarise this whitepaper on cloud architecture. • What's the difference between rest and graphql apis?

Task	Profession	Pct of prompts	Rank	AI Capability	Example Prompts
Edit instructional materials, such as books, simulation exercises, lesson plans, instructor guides, and tests.	Instructional designers and technologists	0.15	96	Reviewing Work, Generation	• Check this multiple-choice question for bias. • Write an introduction to a module on computer literacy.
Write changes directly into compositions, or use computer software to make changes.	Music composers and arrangers	0.15	97	Generation, Reviewing Work	• Add a bridge section to this melody in c major • Review this transition between chorus and verse.
Design or develop software systems, using scientific analysis and mathematical models to predict and measure outcome and consequences of design.	Software developers, systems software	0.15	98	Information Retrieval, Technical Assistance	• How can mathematical models assist in system development? • Model a predictive maintenance system for industrial sensors.
Analyze test data, making computations as necessary, to determine test results.	Inspectors, testers, sorters, samplers, and weighers	0.15	99	Summarization, Technical Assistance	• Summarise these qa results for executive reporting. • Write code to calculate the mean and standard deviation for a dataset of product weights
Develop database architectural strategies at the modeling, design and implementation stages to address business or industry requirements.	Database architects	0.14	100	Technical Assistance	• What's the best indexing strategy for a highly relational database and show me how to implement it