Oxford Policy Management

# UK Perspectives on Program Evaluations and Generative Artificial Intelligence

Final report

Paul Jasper, Tom Wagstaff, Humaira Hansrod

February 2025

# About Oxford Policy Management

**Our vision is for fair public policy that benefits both people and the planet. Our purpose is to improve lives through sustainable policy change in low- and middle-income countries.**

Through our global network of offices, we work in partnership with national stakeholders and decision makers to research, design, implement, and evaluate impactful public policy. We work in all areas of economic and social policy and governance, including health, finance, education, climate change, and public sector management. We have cross-cutting expertise in our dedicated teams of monitoring and evaluation, political economy analysis, statistics, and research methods specialists. We draw on our local and international sector experts to provide the very best evidence-based support.

# Table of contents

# List of tables, figures, and boxes

# List of abbreviations and technical terminology

## Abbreviations

| | |
|---|---|
| AEA | American Evaluation Association |
| AGI | Artificial General Intelligence |
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| GPT | Generative Pre-trained Transformer |
| LLM | Large Language Model |
| NLP | Natural Language Processing |
| ODK | Open Data Kit |
| OPM | Oxford Policy Management |
| RAG | Retrieval Augmented Generation |
| RCT | Randomized Controlled Trial |
| SSC | Shared Services Canada |

## Technical terms

| | |
|---|---|
| Agent | A large language model (LLM) with access to tools so that it can not only generate text but take actions e.g. book a flight |
| Application Programming Interface | A service giving online access to a large language model (LLM), allowing the creation of applications that use its features |
| Artificial General Intelligence | An artificial intelligence with a level of problem solving ability comparable to a human |
| Chatbot | A computer program designed to simulate human conversation |
| Context window | The number of tokens (words) a large language model (LLM) can retain in memory |
| Deductive coding | Identifying the topic of a statement from a predefined list |
| Fine-tuning | Training an artificial intelligence (AI) model using new data, allowing some parts of the model to learn while holding other parts fixed. Often used to speed up the learning process of large language models (LLMs) |

| | |
|---|---|
| Foundation model | The core large language model (LLM) - text-generating model – itself, as opposed to more complex products built using one or more LLMs. In this paper, we also use this term to refer to the most direct interfaces with minimal additional software engineering e.g. we refer to both GPT-4o and ChatGPT as “foundational models” |
| Generative AI | Artificial Intelligence (AI) models primarily designed to generate text, images etc. |
| Hallucination | The phenomenon where Generative AI models present invented material as fact |
| Inductive coding | Identifying the topic of a statement flexibly, using judgement and discretion rather than referring to a list |
| Large language model | A model designed to understand human language and employing neural networks trained on vast amounts of data |
| Open Data Kit | A technical software engineering standard for creating and deploying surveys |
| Natural Language Processing | The field of developing computer programs that understand human language |
| Neural network | A machine learning model that learns to make predictions using a complex series of intermediate relationships in a network relationship |
| Prompt | Detailed instructions for a large language model (LLM), designed to guide its behaviour and its output |
| Prompt chaining | Asking repeated prompts in an attempt to narrow down the large language model’s response |
| Prompt engineering | The art and science of designing a good prompt |
| Retrieval Augmented Generation | A system where a large language model (LLM) is given access to a large corpus of text to inform its response |
| Token | The unit of language used by large language models (LLMs) – often a word, but can be a suffix like -ing or punctuation like a full stop. |
| Transformer | The specific neural network architecture used to build large language models (LLMs). Uses an attention mechanism to allow the model to focus on a subset of relevant words in the dialogue. |

# 1 Introduction

## 1.1 Context: objectives and research questions

This project aims to explore and evaluate the potential of using Large Language Models (LLMs) to improve the efficiency and effectiveness of the evaluation process at Shared Services Canada (SSC). Specifically, it focuses on identifying opportunities for LLMs (like ChatGPT 4.0 and others) to automate or facilitate routine tasks within the evaluation process, thereby freeing up SSC evaluators to focus on more complex, high-value analytical work.

This report begins by reviewing the steps in the current SSC evaluation process and addresses the following key research questions:

1. **How efficient are each of the steps in the current SSC evaluation process?** This involves identifying bottlenecks or areas where processes are time-consuming or inefficient. While precise quantification of time savings proved challenging due to varying levels of Artificial Intelligence (AI) proficiency and trust among the evaluators interviewed, the same interviewees highlighted how their use of LLMs demonstrated notable efficiency gains in key evaluation tasks. This is presented in Box 3 of section 2.2.
2. **Where are the opportunities to leverage LLM tools?** This focuses on pinpointing specific tasks within the evaluation process that could be automated or assisted by LLMs. Section 2.2 draws from the interviews with SSC evaluators to identify current LLM-supported tasks (see Box 1 and Box 2 for examples) and discusses areas where there is potential for further leveraging LLMs in evaluation tasks.
3. **What LLM tools are available?** This involves researching and identifying existing LLM tools that could potentially be useful.
4. **What tools would we recommend adopting, in what priority order?** This involves providing prioritized recommendations for tool adoption based on their potential impact and feasibility.

Additionally, the report will consider potential risks or downsides associated with adopting the recommended LLM tools.

The intended audience of this report is not limited to SSC management but rather includes leaders of evaluation functions across government, academia, and any other providers of rigorous evaluation services.

## 1.2 Methods

### 1.2.1 Data collection and analysis

The project team employed a mixed-methods approach involving qualitative data collection and analysis techniques in parallel with practical testing of specific LLM tools to answer the above questions. More specifically, this involved the following activities:

### 1) Document review:

- We reviewed the SSC Evaluation Division's project methodology document, standard evaluation examples, internal guidance documents, evaluation strategy, framework, matrix, interview guides, and survey questionnaires.
- We reviewed publicly available guidance documents on the use of generative Artificial Intelligence (generative AI) and responsible use of AI from the Canadian government.

### 2) Interviews:

- We conducted semi-structured interviews with eight evaluation staff members at SSC. These interviews explored current evaluation practices, experiences with LLMs, and the potential for LLM integration.
- We also carried out semi-structured interviews with key informants working in the AI/LLM space. These interviews sought the perspectives of leaders in the AI for evaluation space on the current applications of LLMs, their practicality for adoption into SSC evaluations, and future opportunities with LLMs that SSC should monitor for potential applications.

### 3) Practical testing/tool assessment:

- We tested selected LLM tools using publicly accessible demos.
- We developed a rubric to assess the performance of each tool on specific evaluation tasks identified in the document review and interviews.

The insights from these activities were used to answer our research questions as follows:

- To answer research questions 1 and 2, we used the results from our document review and interviews with SSC evaluators.
- To answer research question 3, we used the results from our summary document review and from our interviews with key informants in the AI/LLM space to produce a list of relevant LLM tools.
- To answer research question 4, we used insights gained in interviews, as well as our own assessments from practical testing activities. To be able to assess tools, we developed a rubric (described in more detail in section 1.2.2) that we then applied to a generic evaluation workflow based largely on the ‘Conducting’ phase of the SSC process: data collection, deriving insights, synthesizing findings, and disseminating learnings. We also included training evaluators as a standalone step.

## 1.2.2 Criteria for appraising LLM tools in evaluations

To be able to produce recommendations on the use of AI tools, and to be able to answer research question 5, we developed criteria to assess their usefulness for evaluations. There are a wide range of benchmarks and metrics for assessing the performance of the core models, and leaderboards abound online. Some, like artificialanalysis.ai (leaderboard) compare them based on their specifications and cost, some like LM Arena based on user votes, while still others like MTEB and Vellum compare them based on their ability to perform a specific task or set of tasks.

We will, for the most part, avoid these metrics (which are readily available in any case) and pursue a rubric-based approach, asking how a tool is likely to impact an evaluation task,

based on the following criteria, drawn from American Evaluation Association (AEA) evaluation standards[1]:

- Utility – stakeholders find the evaluation useful
- Feasibility – the evaluation is carried out efficiently and effectively. In particular we expect impacts on:
  - Cost – will it make the process cheaper?
  - Time – will it make the process faster?
- Accuracy – are the evaluation findings dependable and truthful? As an adjunct to this, and given the promise of LLMs, we dwell in particular on:
  - Richness – is it likely to unearth fresh insights beyond the traditional process, for example by expanding the scope of the evaluation (e.g. including a wider range of respondents or more extensive responses) or enhancing the analysis itself (e.g. identifying more complex patterns or surfacing nuance in the responses)
- Propriety – the evaluation is carried out in a just and fair manner
- Accountability – the evaluation approach is transparent, with the steps taken well documented

We will discuss the impacts qualitatively in the narrative sections of the report, but for convenience we also distill this into a traffic light rating summarizing the impact of each technology on each of these criteria. Green indicates significant expected improvements with low to no risk; red indicates that the tool is likely to reduce evaluation quality on this dimension; and amber indicates the potential for improvements but with significant risks of failure or harm. Finally, we also use a grey/transparent category where we do not expect any impact of the tool on that criterion.

### 1.2.3 Terminology used in this report

In this section, we clarify some of the terminology used in this report. We present key terms and their definitions in the list of abbreviations and technical terminology. Moreover, for the purposes of identifying and evaluating relevant LLM-based tools, it will be helpful to divide them into three categories. Our discussion of use-cases in section 3 will be organized around these categories.

The first we will refer to as **foundation models**, the underlying transformer models developed since the seminal 2017 paper Attention Is All You Need, which solve the problem of 'what word comes next?' This includes OpenAI's GPT-3, GPT-4, and GPT-4o, Anthropic's Claude models, Meta's Llama models, Google's Gemini, and so on. We stretch the definition by also including in this category the fine-tuned 'instruct / chat' versions of these models and their chatbot interfaces (i.e. ChatGPT, Claude.ai, etc.). While these interfaces actually supplement the base model with a significant amount of traditional software engineering, we can regard this as interacting with the model 'out of the box', with no domain-specific tailoring. We also include in this category extremely general applications with a near universal market, such as LLM-assisted search. Examples include Perplexity.ai and integrations like Microsoft Copilot, which offers search but also an LLM plug-in to a wide range of applications. Minimally tailored instances of these generic tools, like CanChat, also fall into this first category. For each application in the following sections, we start by

[1] These standards have not been formally adopted by the AEA but were rather developed by the Joint Committee on Standards for Educational Evaluation, of which the AEA is a member.

considering what can be achieved by 'doing it yourself' with a foundation model or generic LLM chatbot.

Secondly there is **traditional qualitative research software**, which has begun to incorporate functionality provided by LLMs. For example, ATLAS.ti has long been used by qualitative researchers for manual analysis, but is now incorporating a growing range of automated functions powered by OpenAI's GPT models. A similar trend is happening with NVivo. Exactly how these integrations work is unclear, but they are likely to involve more than a simple Application Programming Interface (API) call so it might be quite difficult to replicate the functionality by working with a foundation model directly. Moreover, it is a strength of these offerings that this software was always designed to be used by a human researcher who retains control of the process. (Our sense is that LLMs can be very effective at accomplishing narrow tasks but they require close supervision. They are not so-called 'Artificial General Intelligence' (AGI) that is truly capable of the abstract reasoning necessary to faithfully execute the whole evaluation process without intervention.)

Our final category are **off-the-shelf AI tools** designed specifically for evaluation. These are usually provided by small, single-product vendors – often start-ups. They use an LLM as their underlying technology but typically aim to provide evaluation functions 'at the push of a button' with minimal user tailoring (although this varies). These products are powered by a range of techniques and technologies. Sometimes they will be based on a fine-tuned instance of the foundation model, which has been exposed to a large corpus of evaluation documents. Often, they will be a Retrieval Augmented Generation (RAG) system, i.e. a generic LLM pointed at a specific corpus of text to provide the knowledge to inform the response but falling short of retraining the core model itself. Similarly, they could be an 'out of the box' LLM prompted specifically to achieve certain tasks, especially as part of a larger piece of software – using traditional software engineering techniques – which employs the LLM at certain points to accomplish certain tasks, but without the LLM doing all the work. This type of product could be approximated in-house using a workflow. Sometimes they will use agents (i.e. multiple LLM-based tools that interact with each other) to get a job done. Vendors are often not transparent about the underlying technologies, but we suspect fine-tuned models are relatively rare, whereas incorporation in a (hard-coded) workflow using extensive prompting is quite common. How well this works in practice varies depending on the application. Many of these tools – while useful for program evaluations – are aimed at a wider market of product developers and market researchers reviewing customer feedback. They may therefore be somewhat limited use in other contexts.

## 1.3 Structure of the remainder of the report

The remainder of the report is structured as follows:

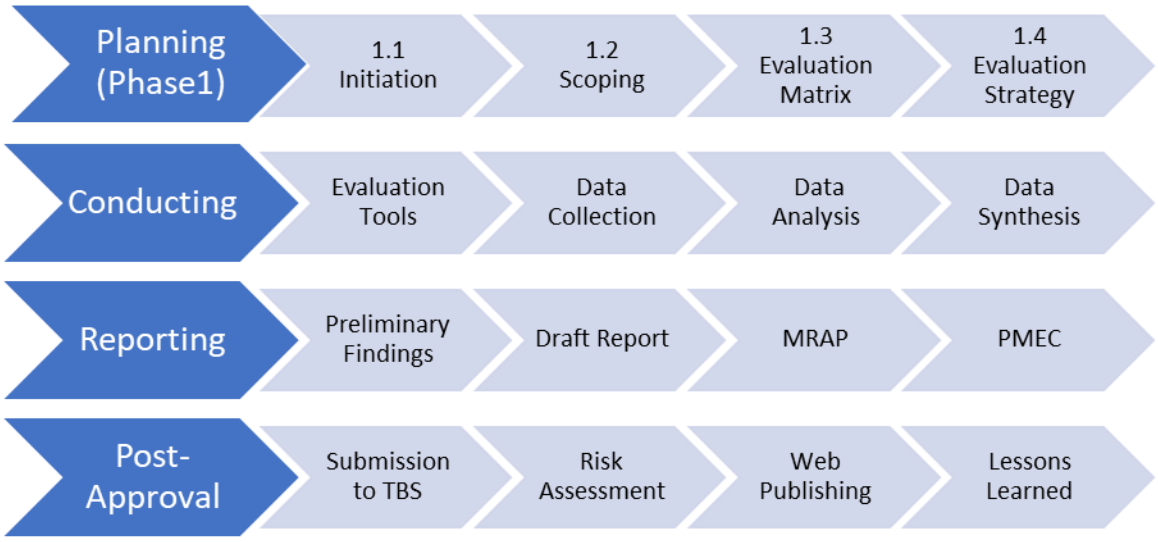
- In section 2, we tackle research questions 1 and 2 by presenting a summary description of current evaluation processes at SSC, including the ways that eight evaluators described utilizing LLMs in evaluation tasks (note this is valid as at November 2024, when the interviews with evaluators were completed).
- In section 3, we tackle research question 3 by reviewing and assessing a set of available AI and LLM tools. Note that we use the criteria presented in section 1.2.2 for this.
- In section 4, we tackle research question 4 by bringing together our assessments of the individual tools and sketching out a roadmap for adoption considering what complementary investments in processes and training might be required.
- In section 5 we conclude, providing summary answers to all the research questions.

# 2 SSC evaluation processes and the current use of AI

## 2.1 Current SSC evaluation processes

The evaluation process at SSC, which is illustrated in Figure 1, involves several phases, including: planning, conducting (e.g. data collection, analysis), reporting, and post-approval (e.g. public dissemination). During the planning phase, evaluators review internal documents, develop evaluation strategies (i.e. workplans or terms of reference), and data collection tools, as well as creating presentations for senior management. In the data collection phase, they gather and analyze both qualitative and quantitative data, using tools such as interviews, surveys, and administrative data. Data synthesis is performed using a results matrix, which helps in forming conclusions and recommendations. In the reporting phase, evaluators contribute to report writing and, after final approval of the report, may also create dissemination products such as publications for the website. Throughout the evaluation process, evaluators collaborate with colleagues and may at any time be working on 2–3 evaluations concurrently.

**Figure 1: SSC Evaluation Process Map**




Source: Shared Services Canada Evaluation Methodology: GCDOCS #86838759

## 2.2 Current use of AI and LLMs in evaluations at SSC

Evaluators at SSC are increasingly using AI and LLM tools, particularly the in-house CanChat tool, to support various evaluation tasks involving unclassified data. CanChat is an enterprise version of ChatGPT where the data resides within the organization and is not used to train the LLM. Other tools mentioned by the evaluators include Copilot and ChatGPT. Evaluators have used CanChat for tasks such as summarizing interview transcripts, extracting themes from documents, generating evaluation questions, and generating preliminary slide decks and written products. Some evaluators mentioned using

Copilot for transcription and generating speaker notes and DeepL for translations and language quality checks. Box 1 highlights how SSC evaluators have used AI tools to conduct qualitative tasks.

As at November 2024,[2] the tools were not fully meeting the needs and expectations of SSC evaluators. For example, one evaluator mentioned using CanChat to summarize a 107-page document section by section and found the generated summaries not detailed enough, even when the tool was prompted with a follow-up for more details. Nevertheless, after doing a comparison between Copilot and CanChat, the evaluator decided that, while both tools were disappointing, CanChat performed the same task slightly better. Evaluators also tried using CanChat to analyze interview transcripts. One interviewee noted that, before using AI, evaluators would look at the full set of interview notes then code. However, based on another team's experience they decided to 'generate summaries for interview notes then focus on coding and analyzing just the summaries to speed up that process'. While this did speed up the process, the evaluator found they still needed to go back to the interview notes for details that were key to qualitative analysis and that simply could not be picked up by CanChat.

### Box 1: Using AI tools to support qualitative tasks

Interviewees described using AI for qualitative tasks in the following ways:

- **Summarizing interview transcripts:** Several interviewees mentioned using CanChat or Copilot to summarize interview transcripts, which can help evaluators quickly identify key themes and findings.
- **Extracting themes and codes from interviews:** Some interviewees described using CanChat to extract themes and codes from interview excerpts, which can help with qualitative data analysis and synthesis.
- **Generating interview guides and survey questions:** CanChat has been used by at least three evaluators to generate potential interview questions or survey questions based on evaluation findings or other inputs, helping evaluators develop data collection tools.
- **Improving the quality of written reports:** Interviewees mentioned using CanChat to improve the clarity, conciseness, and professionalism of their written reports, as well as to explore different ways of expressing findings.
- **Transcribing interviews:** Copilot Pro has been used to transcribe interviews by at least two evaluators, with these individuals noting that it more effectively transcribes and creates summaries of virtual interviews, saving evaluators time and effort.

Additionally, albeit to a lesser extent than for qualitative tasks, evaluators also described using AI to support quantitative analysis. One interviewee noted using CanChat more for idea generation, such as helping to define an efficiency metric, rather than to complete specific quantitative analyses or tasks. Box 2 outlines how some evaluators have used AI (i.e. CanChat) to support quantitative analysis.

[2] Interviews with SSC evaluators were conducted in October and November 2024, meaning findings reflect evaluators' perspectives at that time. Given the rapid evolution of LLM technology, some observations on usage may not represent current practices by the time of this report's submission in February 2025.

### Box 2: Using AI tools to support quantitative tasks

Interviewees described using AI for quantitative analysis in the following ways:

- **Generating formulas in Excel:** One interviewee mentioned using CanChat to generate formulas for calculations in Excel, which they found helpful because they did not know many formulas themselves.
- **Identifying good indicators for evaluation questions:** An evaluator suggested that AI could be used to identify good indicators for evaluation questions, which could help make the evaluation approach and methodology more rigorous.
- **Providing instructions for answering analysis-related questions:** An interviewee shared an example of asking CanChat an analysis-related question and receiving instructions, which they verified and found correct, on how to answer it using Excel.

### 2.2.1 Challenges and concerns

Despite the potential benefits, evaluators also face challenges and have concerns about using AI and LLMs. Some have trust issues regarding the accuracy and completeness of AI-generated outputs, leading to a need for extensive review and verification. For instance, one evaluator expressed hesitation in using AI for most tasks after she tried to use it and found that CanChat missed important details and reported some inaccurate findings. Another evaluator noted that being unable to attach documents to the current CanChat tool prevents evaluators from using the tool to conduct document retrieval and other tasks.

There were also concerns about AI tools missing important nuances in qualitative data and the potential oversimplification of complex issues. When describing the process of inputting interview data into an evaluation matrix for analysis, an evaluator described it as a 'boring, long, and tedious' step but explained that, when they tried using CanChat to speed it up, they found disappointing results in the form of diluted answers. Additionally, there were concerns about data privacy and security when using external AI tools, as well as the need for clear guidelines and policies on AI use in evaluations. As one evaluator explained, '*I'm hesitant because we work for the government, so it's protected data that we cannot risk other people having access to*'.

### 2.2.2 Potential benefits and future opportunities

Despite some questions and challenges, evaluators recognize the potential benefits of AI and LLMs in enhancing the efficiency and quality of evaluations at SSC. These tools can save time on tasks such as document review, qualitative data analysis, and report writing. For example, one evaluator estimated saving a full day's work by using CanChat to avoid reading a document word by word. Another noted that SSC Assistant had limited capacity but was nonetheless useful to look at internal SSC-related information. They felt CanChat has been most useful, especially in 'summarizing, generating portions of material, finessing writing, and suggesting new things to include or identifying gaps in written material'.

Drawing from the interviews, it is clear that evaluators at SSC believe AI can support them in day-to-day tasks in every phase of the evaluation process, such as helping generate new ideas, improving writing quality, and supporting data visualization. One evaluator used Copilot to generate speaker notes based on a slide deck. Another used CanChat to summarize an interview transcript, prompting the tool to provide bullet points and a thematic analysis of the transcript. Based on what SSC evaluators are using LLMs for, as well as what is likely to facilitate their tasks, future opportunities in evaluations include using – and

guiding the evaluation team to use – AI for tasks such as developing evaluation matrices, identifying indicators, reviewing documents, and tailoring evaluation products for different audiences. As one evaluator highlighted, there is potential to use an AI tool like CanChat to produce various communication products aimed at specific audiences. For example, what is currently used as a final written product to the President of the department is also the same product used to inform the broader Canadian public on the results of an evaluation. But as this evaluator noted, 'if we ask the AI tool to tailor and adjust the language of our written product, we could do a one-page summary of a report to put on the website [for the broader public]' that is much more accessible and increases the reach of SSC's evaluation findings. Given that evaluators have different fluencies in the use of AI and LLMs, and with varying levels of trust in these tools, it was not possible to reliably quantify efficiency (i.e. using amount of time saved doing tasks manually versus using an LLM). Nonetheless, Box 3 summarizes how evaluators saved time using LLM tools to speed up certain key evaluation tasks.

### Box 3: How AI and LLM use has improved efficiency at SSC

**Planning phase**

- **Identifying relevant internal documents**: AI and LLM tools can analyze large volumes of internal documents and identify those relevant to a specific evaluation, saving evaluators time and effort.
- **Developing evaluation strategies and data collection tools**: AI and LLM tools can assist in generating evaluation questions, developing interview guides and survey questionnaires, and creating evaluation matrices.
- **Creating presentations for senior management**: AI and LLM tools can generate presentations for senior management, including speaker notes.

**Conducting phase**

- **Summarizing interview transcripts**: AI and LLM tools can summarize interview transcripts, identify key themes and findings, and extract relevant quotes.
- **Analyzing qualitative data**: AI and LLM tools can support qualitative data analysis by extracting themes and codes from interview excerpts and generating thematic analyses of transcripts.
- **Analyzing quantitative data**: AI and LLM tools can assist in quantitative data analysis by generating formulas for calculations in Excel, identifying good indicators for evaluation questions, and providing instructions for answering analysis-related questions.

**Reporting phase**

- **Writing evaluation reports**: AI and LLM tools can contribute to report writing by generating outlines, drafting sections of the report, and improving the clarity, conciseness, and professionalism of the writing.
- **Tailoring evaluation products for different audiences**: AI and LLM tools can tailor evaluation products for different audiences, such as creating one-page summaries of reports for the broader public.

**Post-approval phase**

- **Public dissemination**: AI and LLM tools can support public dissemination of evaluation findings by generating various communication products tailored to specific audiences.

### 2.2.3 Training and capacity building

To effectively leverage AI and LLMs, all evaluators interviewed identified the need for training and capacity building. The interviews suggested training on prompt engineering, understanding AI capabilities and limitations, and addressing ethical considerations. One evaluator mentioned that the Treasury Board Secretariat had some information on handling sensitive information, but that these were 'vague guidelines'. They also highlighted that having clear guidance from senior management within SSC will be most useful in helping evaluators understand what exactly they can and cannot use AI tools for. Additionally, evaluators noted that training should cover best practices for using AI in specific evaluation tasks and how to integrate AI outputs with human expertise. 'If more tools became approved or there were more functions in CanChat, I'd be open to use them', added one evaluator, who was unsure on using AI in evaluations—especially since they believe that, eventually, the Director of Evaluation will train CanChat to provide feedback on their logic models.

Monitoring how team members use LLMs in their evaluations will be key to sustainable capacity building and ensuring that the broader team of evaluators can learn from each others' experiences in using AI tools on certain evaluation tasks. SSC has already taken steps in that direction, such as through the 'use-case tracking' document used by some evaluators who trialled Microsoft Copilot between June and September 2024. Similarly, SSC may want to consider building a prompt library, which is essentially a curated collection of effective prompts and prompt templates, as a recipe book for getting the best results from LLMs. Box 4 highlights key uses and benefits for SSC in building and maintaining such a library.

#### Box 4: Uses and benefits of building and maintaining a prompt library

A well-maintained prompt library can be a valuable asset for harnessing the power of LLMs. It can enable evaluators to optimize their workflow and drive consistent, high-quality results. Key uses and benefits include:

**Improved efficiency and productivity:** Instead of evaluators repeatedly crafting prompts from scratch (which can be time-consuming and require expertise), they can access a library of pre-tested prompts, significantly speeding up their workflow.

**Enhanced consistency and quality:** Standardized prompts can ensure consistent outputs from the LLMs, reducing variability in quality and making it easier to compare and analyze results across different projects or team members.

**Knowledge sharing and collaboration:** A central prompt library acts as a repository of knowledge, allowing teams of evaluators to learn from each other's experiences and build upon successful prompting strategies. It facilitates collaboration and fosters a culture of continuous improvement.

**Faster onboarding and training:** New team members can quickly get up to speed on how to effectively use LLMs for evaluation tasks by studying and utilizing the existing prompt library.

**Reduced errors and biases:** Carefully crafted prompts can help mitigate biases and errors in LLM outputs. The library can include guidelines and best practices for writing prompts that minimize these risks.

**Increased experimentation and innovation:** By providing a foundation of effective prompts, the library frees up staff to experiment with more advanced prompting techniques and explore new ways to leverage LLMs for evaluation work.

**Scalability:** As the company's use of LLMs expands, the prompt library can be scaled to accommodate new tasks, tools, and evaluation processes, ensuring that best practices are shared and maintained across the organization.

**Specific to SSC's evaluation work, a prompt library could contain templates for:**

- Summarizing evaluation reports

- Identifying key findings and recommendations
- Generating interview questions
- Analyzing qualitative data for themes and patterns
- Creating data visualizations from evaluation results
- Drafting sections of evaluation reports
- Comparing and contrasting different evaluation methodologies

The findings from the interviews with evaluators reveal significant potential for AI and LLMs to enhance efficiency and effectiveness across all phases of the evaluation process at SSC. Evaluators are already using these tools for various tasks, from generating ideas to improving writing quality and data visualization. The technology shows promise in streamlining document review, developing evaluation matrices, and tailoring communication products for different audiences. However, the varying levels of AI fluency and trust among evaluators highlight the need for structured training and clear guidelines on AI usage. While quantifying efficiency gains remains challenging, the potential for time-saving and improved output quality is evident.

# 3 What's new: promising use-cases of AI in evaluations

## 3.1 How does generative AI improve worker productivity?

The internet abounds with breathless predictions about the impact of generative AI on enterprise performance in the form of revenues, costs and profits, and individual worker productivity. Many of these are authored by the major AI vendors and often there turns out to be scant evidence for the extravagant claims.

Google's AI Trends 2025 surveys a whole range of applications of generative AI, but says very little about impact, restricting itself to estimates of generative AI adoption and the size of the AI market. It refers to other studies like its ROI of GenAI report, based on a survey of 2,500 enterprise leaders; however, even here, the results are less impressive than they might first appear. Apparently, '86% of organizations using gen AI in production and seeing revenue growth estimate 6% or more gains to overall annual company revenue'; the reader is left to calculate for themselves that less than half of adopters saw revenue growth (and there is no data on how many saw revenue drop). Likewise, '45% of organizations that report improved productivity have seen employee productivity double or more as a result of gen AI', but it turns out only 70% of organizations did report such an increase. It is also interesting that two-thirds claim improved worker productivity but only one-third saw any uplift in revenue: this might reflect the fact that the business leaders who responded to the survey have a better handle on company revenues than their employees' day-to-day working practices.

OpenAI released an early, if obscurely titled, forecast of the impact of generative AI called GPTs are GPTs. This turns out to limit itself mainly to estimates of adoption rather than impact, and we are not aware of any more recent literature on impacts from OpenAI. Microsoft has released a Total Economic Impact Study of Copilot, which makes stronger claims: a 4% increase in revenue, 8% increase in profit over three years = a 457% return on investment in generative AI. However, it is unclear how this estimate – which appears to be a forecast – was arrived at. Apparently, it was 'based on a composite organisation', which we take to mean it was simulated in some way.

It is important to recognize that high-level reviews of generative AI's impact also include detractors. For example, Goldman Sachs' Gen AI, too much spend, too little benefit? expressed concern in mid-2023 that tech companies appear to be over-investing in a technology with few proven benefits.

Rigorous estimates are hard to find, and are more circumspect. Some randomized controlled trials (RCTs) have been performed and found strong impacts across a range of tasks in a simulated setting. This one held back from offering any overall productivity estimates, but reported 20% improvements in speed and quality in professional writing tasks. A series of quasi-experimental case studies, reviewed here, claim productivity increases on average of 66%. While the methodologies are not obviously unsound, if these results are true we wonder why the vendors of AI themselves are not claiming anything like such transformational impacts on the economy.

Although not a regular task of SSC's evaluation department, it is relevant to examine the effect of LLMs on software developer productivity, since it should be the most obvious – both

because it is easy to measure and estimate the impact and because this is a promising application of LLMs. Since ChatGPT excels at generating one series of text from another, translation between natural language and programming language should be among the canonical applications of the technology.

Here, the evidence is stronger but it is still limited. Cui *et al.* (2024) report productivity improvements based on RCTs at Microsoft, Accenture, and a third major company. Their headline result is a 26% increase in completed tasks among developers, but the report skips over cases where the generative AI seems to have been detrimental (e.g. on quality). In the Microsoft case, a significant proportion of the supposed control group was using generative AI anyway, meaning it is unclear what the results really measure. In the case of Accenture, where there was a much clearer separation between the two groups, the generative AI adopters completed far more new software builds, which then went on to fail at a much higher rate.

An extensive quasi-experimental study from GitHub estimated the impact of GitHub's Copilot tool by comparing countries where the tool had been launched with those where it had not. They found a significant impact on the proportion of the population who were software developers and increased rates of GitHub activity, including the number of repositories (coding projects) and the number of pushes (adding a single piece of code to a repository). The impacts – at the level of national populations over such a short time – seem implausible to us. Moreover, they only measure activity and it seems plausible to us that the availability of GitHub Copilot may have encouraged more enthusiasts to create amateur code bases and contribute to them. This in itself is a good thing – in terms of democratizing access to advanced technology – but we do not see strong evidence of a productivity improvement here.

Even less has been said specifically about the evaluation sector. Jasper *et al.* (2023) estimate fairly widespread adoption, with impacts on most stages in the evaluation workflow, but were not in a position to estimate the resulting savings or improvements in quantitative terms. Moreover, the predictions were based on forecasts by expert informants in the evaluation sector, rather than actual trials.

In sum, evaluators are knowledge workers and this review has shown that there is not (yet) a large, high-quality evidence base for significant productivity impacts of AI/LLM adoption in knowledge work. Hence, it is also possible that the adoption of AI and LLMs will not significantly affect evaluations. However, to assess this in more detail we proceed to review these tools in light of the assessment criteria described in section 1.2.2.

## 3.2 Currently available applications of LLMs in evaluations as at February 2025

### 3.2.1 Evaluation design

Many of the more promising applications of generative AI are based on the human evaluator and AI co-creating evaluation materials, with the human using the AI as a tool for ideation, drafting of content, and feedback. Indeed, learning from the AI how to do complete tasks or what end result is expected seems to be a typical experience, as evidenced by the fact that inexperienced or junior workers tend to be those who benefit most from employing generative AI.

In the longer term, generative AI might make it possible for a junior evaluator to simulate the entire process of conducting an evaluation end-to-end before they even begin, designing interviews, simulating the responses, and drawing conclusions. This would allow them to rehearse each of the steps and tailor or amend the design to take account of issues that are likely to arise. A key concern here would be that generative AI could simulate a lot of reasonable-looking material that is in fact far removed from what would be created in the real world, thus misleading the trainee. Having said that, there are some promising prototypes in development, including Stanford's simulation of survey respondents, which did quite accurately replicate real-world responses using generative AI.

For now, we focus on less ambitious applications of generative AI to assist an evaluator in evaluation design, particularly focusing on the data collection stage.

### Foundation models

Evaluation design is an area where simple access to a foundation model might be all that is needed. For ideation purposes, you can simply ask the LLM a question and see what response you get. Examples of potentially useful applications include:

- Providing an overarching evaluation topic and asking the LLM to propose questions;
- Providing a set of questions and asking it to identify blind spots; and
- Asking the LLM what data collection modality would suit an evaluation topic.

The crucial thing is that the human evaluator should take the ideas generated under advisement. They will often be plausible and based on similar situations that have arisen in the past but it will be up to the evaluator to decide if they are right for the task at hand (or, more likely, how wrong they are and how far they have to adapt them).

### SurveyCTO

SurveyCTO is an established product aimed at academics and official agencies for data collection, primarily via surveys, which has introduced an AI Assistant to guide users in how to make effective use of the tool or design a good survey. This can translate a natural language description of a question into the tabular format required by Open Data Kit (ODK) form builders (see section 3.2.2) and sensibly fill in the gaps in the specification by, for example, suggesting question types or choices if these have not been defined. It can also make suggestions on broader queries (e.g. what questions should be included) or give general advice on survey design (e.g. how many questions should be used).

### Fatima

Fatima positions itself as an ethical data collection platform. It is in some ways an end-to-end evaluation software, in that, as well as designing an evaluation, it can also be used to carry out the data collection, store the results, and promises some automated analysis. We include it here because of its pleasing emphasis on co-creation with the human evaluator, which intends to guide evaluators of various skill and experience levels in the design of survey instruments, with a particular emphasis on ethical compliance and respondent consent.

### BlockSurvey

BlockSurvey is another end-to-end evaluation platform, promising to automate many of the sorts of steps in analysis we will consider below. Again, we include it here because it includes a survey design tool whereby the AI can generate a survey instrument based on a short prompt, but then the user can override the choices and finalize the design for themselves.

### Summary

We identify two basic processes that these tools offer. The first – and the only one where mature tools are available on the market – assists evaluators in the process of evaluation design. We believe this will enhance utility (by improving evaluation design), feasibility (by increasing speed), and accuracy (by asking better questions). There are, however, some risks to propriety and accountability if evaluators become over-reliant on these tools and unthinkingly adopt their recommendations.

The second tool, which is still in development, goes one step further and promises to effectively rehearse the evaluation by simulating responses. This could super-charge the benefits listed above for survey design, accelerating and improving a complex process. However, here there are risks to accuracy – if simulated responses are not actually representative of real respondents' feedback, they could be misleading to the point that the wrong questions are asked. This would give the illusion of reducing design costs and timescales while actually just leading to botched designs, which is a risk to feasibility and utility. Moreover, there are more serious risks to accountability and propriety if we remove an entire feedback loop from real respondents.

**Table 1: Evaluation design: traffic light ratings**

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| Survey design | | | | | |
| Stakeholder gaming | | | | | |

## 3.2.2 Data collection

Data collection is a step in the evaluation workflow that has already been transformed by the employment of digital technologies. Surveys can be implemented at much larger scale and lower cost than paper-based systems allowed, and there is a mature market for digital surveys with a range of suppliers, many using the ODK standards.

Some might question whether the availability of these tools has made implementing a survey too easy, with the end result being a lot of organizations running their own surveys without proper regard for sampling and statistical validity. Another drawback might be a premature rush to scale, rolling out survey instruments that are asking fundamentally the wrong questions due to insufficient ideation and testing at the start of the process. These surveys would benefit from an initial phase of open, qualitative interviews with key informants to establish the relevant questions to include in the survey.

Could LLMs open up the same opportunity to scale open-ended, loosely structured qualitative interviews? Could this even go some way to improving the way existing tools for quantitative surveys are used? The prospects here appear strong: this is both a more

established use-case than you might expect and there is a growing field of products and suppliers catering specifically to evaluators.

ELIZA, developed in the 1960s, was one of the first chatbots. It was technically very simple and had no AI as we would now understand it: rather it used cleverly devised templates (regular expressions) to take user inputs and feed them back to the user in a way that created the illusion of understanding and invited further input. The key finding from ELIZA was that people were willing to engage with these systems and indeed seemed more willing to share intimate information with them than with a human interlocutor (despite attributing human-like intelligence to the computer program).

In a sense, then, we have had the capability to perform fully open-ended unstructured interviews via computer for a long time. The intriguing capability that LLM-powered chatbots offer is to conduct a semi-structured interview, allowing the user to feed back freely, but also keeping the conversation on topic and asking pertinent probe questions.

The key question is whether an LLM can truly facilitate a rich dialogue while 'sticking to the script'. In our experience, and based on the demos of tools we have had access to, there does exist a proof of concept wherein the LLM can execute a short series of questions and probe where there is limited feedback from the user. What remains to be demonstrated, and we remain skeptical of, is the ability of such a system to replace a qualitative researcher in a true long-form interview that deals in a series of complex concepts.

## Using foundation models

It is not difficult to prompt a foundation model like ChatGPT to conduct a short interview with the user. Once prompted, the chatbot performs well enough at posing the questions and asking somewhat pertinent probes. It would be relatively trivial to turn this functionality into a web app that interviewees could visit to provide the feedback, with the app storing the responses for later analysis. This could already be a very useful application, but what remains unproven, both here and in the bespoke software reviewed below, is the ability to hold an extended dialogue. We anticipate problems with staying on track – most fundamentally, asking every question and not repeating any questions, but more subtly knowing when and how to probe – and formulating appropriate reactions based on the context and correctly inferring the meaning and significance of the responses. For example, when testing the interview in Box 5 below, we tried answering question 2 by saying, 'I don't know, but if I had to guess…' In its final summary, ChatGPT reports the guess, admittedly noting that is a guess, whereas to us it seemed the key finding would be that an implementer of the project is unaware of the objectives. Perhaps the purpose of the interview could be communicated in a more extensive prompt, but our experience suggests that there are limits to how far a prompt, however long, can shape the chatbot's behaviour.

### Box 5: Example of a key informant interview prompt

# Role

You are a qualitative researcher conducting a key informant interview.

# Context

The user – the interviewee – works in the Clean Cook Consortium (CCC). Your task is to answer the following questions as fully as possible.

# Questions

1. What activities of the CCC are you involved with?

2. What are the objectives of the CCC?
3. How do your activities contribute to the overall objective?

# Instructions

Start the interview when the user types START.

Then process each of the above questions in order.

First, pose each question directly to the user. If their response is relevant but incomplete, you should generate up to two probe questions to obtain extra detail. If their response is irrelevant, ask the original question with a different form of words.

Once the user has finished responding to the final question, output answers to the questions in the following format:

# Output format

## Question 1

- The interviewee is a project manager, co-ordinating the field work of engineers installing new infrastructure
- They also organize contracting of suppliers and donor finance
- They are not involved with the research or policy advocacy arms of the project

## Convo

Convo is designed to be a drop-in solution for an analytics team, aimed primarily at market researchers for applications like customer feedback on products. It uses LLMs to automate most steps in the process, from devising interview questions to conducting the interviews and carrying out a high-level thematic analysis. It produces, as so often with LLM-based tools, a passable first pass at all these tasks, and may be accurate enough for relatively narrow tasks like product analysis. However, it seems to perform all these steps in one go, with limited opportunities for manual intervention and control of the process. Convo apparently has some academic users, but in our assessment an evaluation team will need more control over the process to ensure accurate insights are drawn from long and multifaceted interviews about complex projects.

## Fortell

Fortell is an AI-based solution designed to conduct surveys in the field, i.e. to collect structured quantitative and categorical data that would otherwise be collected by a field technician filling out a form. The value-add of the generative AI in this case is to create an avatar so that the interviewee can dialogue with them on a phone, rather than filling out a webform. There are potential advantages to this tool: it allows data collection in dangerous and conflict-affected areas, expands the number of respondents, ensures the inclusion of people with poor literacy skills, and, potentially, facilitates dialogue on sensitive subjects. However, enumerators are rarely major cost drivers in evaluations, whereas LLM-based tools are expensive to develop and run. Thus, this risks being a high-cost replacement for a job already performed well by humans.

## Qualia

Qualia is aimed at the academic evaluation market and offers more information and control over data security than most other applications we reviewed. It is somewhat similar to Convo's interview module, in that it allows users to set up an interview with certain questions, then share that link with respondents, and have an LLM conduct the interview

with them. It also stores the responses but does not itself analyse the results, at least not to the extent promised by Convo. This is something we regard as a wise design choice on the part of Qualia. Nonetheless, there is some processing; for example, we took a demo interview and at the end, as well as saving the transcript, Qualia output a set of binary pairs that would be suitable for creating a mind map of how the various issues raised by the interviewee are connected.

### Other chatbot-facilitated interview tools

There are countless tools of this type on the market that we were unable to review in detail. For example, Reveal requires you to book a call to see a demonstration of the app, which the provider cancelled. Strella also offers similar functionality, but requires prospective users to sign up to a waiting list.

### Transcription and translation

Machine transcription and translation predates LLMs, although they have certainly enhanced the performance of these tools, so the market of suppliers is relatively mature. One leading provider that tends to be positively reviewed is Otter, which offers transcription of meetings in English, French, and Spanish and can also generate minutes/summaries. (As an aside, this is a suitably low-risk application of auto-summaries: Otter can take a first pass but participants will know what was said in the meeting and swiftly be able to verify and correct the minutes.)

There are also a range of options for accessing the new LLM-based transcription models more directly. Whisper, OpenAI's open-source transcriber, can be downloaded for free and run on your own infrastructure. It can also be accessed as an API for a fee, which is likely to be cheaper than paying for the infrastructure to self-host the model. In our experience, these models perform to an extremely high standard on recordings in English, French, and other high-resource languages. However, they have not worked to the required standard on lower-resource languages, even some of those the model claims to support, and we found them to not be robust to real-world issues such as poor quality recordings, diverse accents, technical terminology, and so on.

### Summary

We identify three fundamental new data collection technologies powered by LLMs. The first are traditional surveys conducted remotely by avatar. These could increase feasibility by allowing data collection from inaccessible areas, but might harm it by inflating costs and dependency on power and data infrastructure. They could increase accuracy by facilitating responses on difficult subjects, but it is possible that respondents will be less forthcoming without the intimacy and trust built up by a human interlocutor. There is a risk to propriety, in the form of shifting resources for data collection away from the local workforce and toward already high-paid developers. There is also a risk to accountability, given that the technology reduces real human contact between project beneficiaries and implementers.

The second are semi-structured interviews conducted by chatbot. These have significant potential to increase the utility of evaluations by unearthing new issues and concerns, as well as to increase feasibility by significantly expanding data collection. We also rate them green for accuracy since they have such potential to enhance the richness of evaluation findings, although hallucination always poses some risk that the wrong questions will be asked. We assign amber for propriety – some of the same concerns about centralizing and

monopolizing knowledge work apply, but we do not see an equally vulnerable group of providers at risk. We rate red for accountability, since the tools do reduce direct contact between respondents and evaluators. We therefore encourage evaluators to offer alternative routes to raising issues about the project and evaluation if using this technology.

Finally, we consider machine transcription and translation. This has great potential to increase feasibility by reducing costs, and increasing accuracy by providing instant written records of evaluation data collection, so we assign green for both. There are still some issues around propriety – particularly the risk of destroying jobs – so we assign red here.

**Table 2: Data collection: traffic light ratings**

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| Avatar-conducted surveys | | | | | |
| Chatbot-facilitated interviews | | | | | |
| Machine transcription / translation | | | | | |

### 3.2.3 Deriving insights

This task is in a sense the heart of the evaluation process and would be the biggest prize if it could be automated via AI. It is also the area where the introduction of AI is the most challenging and poses the biggest risks.

The promise would be that AI can do the grunt work of reviewing long documents and extracting the key content of interest, whether that be in the form of a summary or abstract, answers to specific questions, or by simply extracting the key passages. Demos and tools abound that apparently do this, instantly producing documents that look the part.

Here is where the core workings of LLMs are crucially important to bear in mind. They have been described as ‘auto-predict on steroids’ because the problem they are built to serve is ‘which word comes next?’ This is not to say that they are crude models – their complex architecture encodes a lot of subtle linguistic understanding, and it is worth pointing out that they solve the core problem very well. Further, the text generated by LLMs is always grammatically correct, as the systems learn the rules needed to reproduce language very accurately. The problem, rather, is that this is the *only* problem the LLM is built to solve: the LLM produces plausible, typical text, similar to text it has seen before; it has no independent frame of reference and no view on the truth or falsehood of the statements it makes, as pointed out by Alan Blackwell in Oops, we automated bullshit. While there are many useful applications where the truth value does not matter (e.g. translation of a statement from one language to another, changing the style of a statement, generating ideas, or inviting input from users), this is a fundamental problem when the truth value of the statements does matter. For this reason, every LLM tool still features a prominent disclaimer that all the content generated is liable to be inaccurate or misleading (euphemisms for false).

This is not to say that LLMs cannot be applied for these purposes at all, but we must bear in mind that when we ask an LLM to summarize an interview, code a response to a category, or answer a question about a document, the criteria it uses for the output is the ‘most pleasing’ (or plausible, or typical) response, based on similar examples it has seen before.

There are a number of strategies for addressing this weakness in LLMs. Below, in our horizon scan in section 3.3, we will consider the latest generation of ‘reasoning models’ (which attempt to improve the models’ ‘out of the box’ performance). The established

approach currently is to create a workflow, or perhaps even LLM agents, where each LLM is sandboxed, or given a relatively narrow, low-risk task, and given access to other models that are better suited to other tasks. One of the first of these was to connect GPT to models that were better suited to quantitative calculations (LLMs are notoriously bad at simple arithmetic, counting, etc.) and much of this functionality is now built into the ChatGPT product. The problem with this approach is that it represents in large part a return to traditional software engineering, where specific workflows have to be crafted for specific tasks. That means these applications will require the continuous efforts of high-skilled programmers, leading to high, ongoing redevelopment costs – not to a new era where a general-purpose AI can be used for these purposes by anyone with minimal training.

## Using foundation models

Our experience using foundation models to perform substantive qualitative analysis tasks has been challenging. When asked to summarize documents of a wide variety of types – including formal reports, articles, and key informant interviews – the output is almost always plausible and surfaces some relevant content. However, achieving a comprehensive summary that includes all the relevant details is very difficult, perhaps because what is relevant and to whom is very context dependent and may not be easily inferred from the source material itself. We found that while prompt design can quite easily alter the format of the text output, even long carefully crafted prompts often fail to elicit the right substantive content, and the output generated can vary erratically in response to small changes in the prompt (recognizing the fact that the output always varies randomly to some extent).

There are some tactics that work better, such as prompt chaining where, for example, the initial task is to cite relevant sections of the document, then to create a list, and finally to write this up in narrative form. The idea here is that since the only 'thinking' an LLM does is to generate the next word, by having the LLM 'talk through the problem' it effectively increases the processing power applied to the problem and guides the LLM to a better response than simply answering a question cold. We have had some success using this approach, but still found each step in this process to be littered with inaccuracies that have to be corrected manually.

A tactic that has been particularly disappointing is to break down the summarization task into specific questions to be answered. We find that not only are the answers to these questions often incorrect, moreover they are often irrelevant and contain content that is adjacent to the topic of the question asked but does not actually address the question. This seems to reflect the model architecture – they generate plausible responses based on quite superficial semantic similarity and it shows.

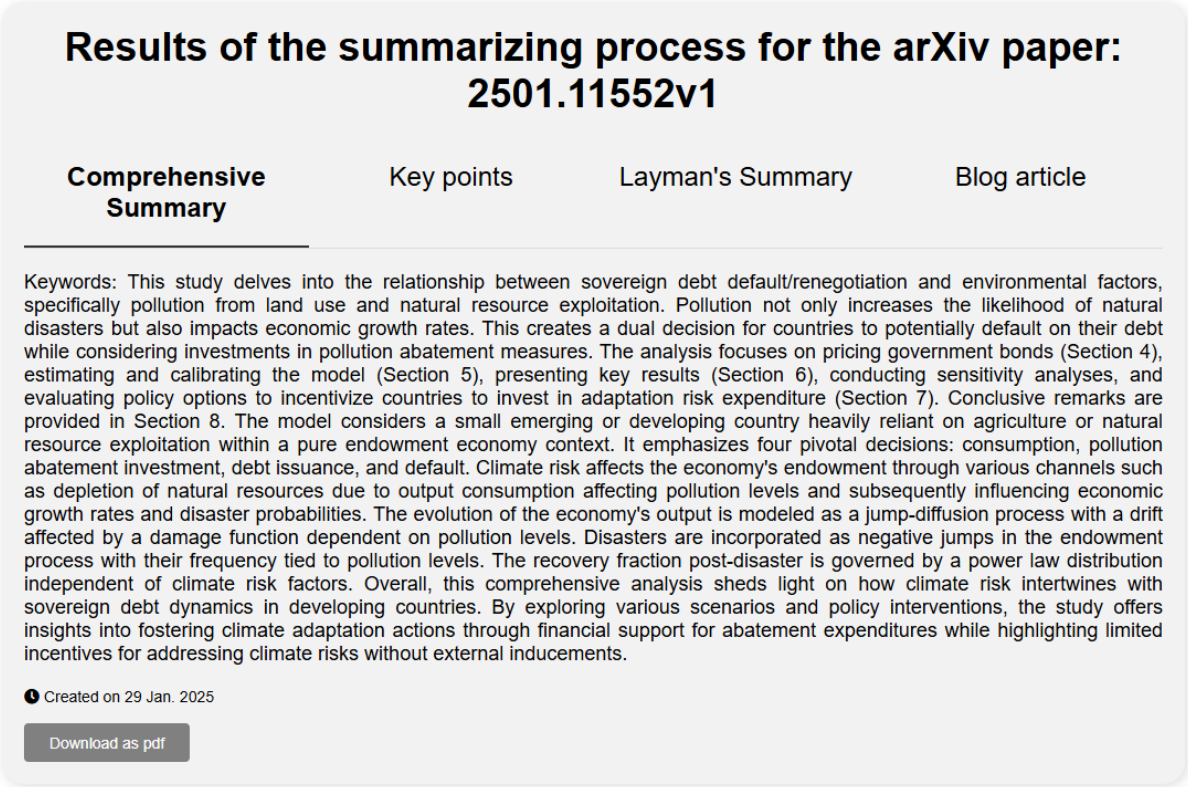
## SummarizePaper

SummarizePaper is an early tool based on GPT-3.5 that can summarize any article published on arXiv. We tested it using an economics paper on national debt and pollution: the result is a rough but reasonably accurate overview, although it is essentially just a shortened version of the text in the article itself. The AI is not capable of critically evaluating the concepts used in the paper or reformulating them, as demonstrated by the layman's summary it produced. This is little more than a table of contents for the paper rather than a true explanation.

More recent models do a noticeably better job of understanding and using the concepts in play. For example, we uploaded the same paper to Claude and it produced a more

meaningful summary that explained the paper rather than just described the contents. The model was also able to engage in a critical dialogue about the paper. So, we have observed that good performance is possible, but there seems to be no guarantee: Copilot did a poor job of summarizing the same article (comparable to early generations of GPT), while Claude did a poor job of summarizing some other articles.

**Figure 2: Example output from SummarizePaper**




Many of the off-the-shelf tools are simply not mature or well-maintained products, which limits their usefulness in practice. We embarked on a project involving a lot of document summarization using ChatGPT, but found that the file upload function was broken. OpenAI was unable to resolve this simple software engineering failure over the course of several weeks, so we had to abandon ChatGPT in favour of Claude to complete the project.

## AILYZE

AILYZE promises to be a drop-in replacement for a qualitative evaluator, offering summaries and thematic overviews of interviews and surveys, either individually or as a body, and is squarely aimed at the official evaluation market. Recently, it has also added some data collection functionality.

Our experience with the tool is that, in common with other tools, it very quickly produces plausible-looking results and these do contain some true insights, but they require careful manual review. It seems to us that the value of a product that generates instant outputs without 'showing its working' in the form of intermediate outputs (e.g. a coded transcript) is limited for fully fledged evaluation teams that will be held to high standards of transparency and expected to be methodical.

However, our main problem with AILYZE was simple platform maintenance – using the web interface the model simply stalled and we had to instead email documents to technical support staff to run the models on our behalf. We soon gave up on this and cancelled our subscription when we saw we had been overcharged for the service.

### Causal Map

Causal Map is another tool aimed squarely at official evaluation, but rather than trying to replicate the entire evaluation workflow it is specifically designed to unearth causal claims and validate theories of change in a set of qualitative interviews. It combs the transcript for causal claims, pulls out the (posited) cause and effect, and displays these relationships graphically.

The approach here seems to be a workflow where separate modules, in each of which the LLM is prompted to perform a different task, are combined together in a single tool. This approach of using the LLM input sparingly in a highly constrained fashion to accomplish a well-defined task seems to us a much more promising approach than many of the tools that promise wide functionality based largely on simply asking an untailored LLM instance. Having only seen a brief demo, we cannot speak for the performance of the app and we anticipate it might struggle to unearth *all* the causal claims that might be implied or indirectly referenced by a respondent, especially when this requires intimate knowledge of the context. Crucially, though, Causal Map does allow the user to override choices made by the LLM, so while the LLM has a first pass the researcher uses it as a tool, rather than it replacing the researcher. This seems to us the appropriate way to be developing these tools with the current level of technology.

### Julius

While not the focus of this study, we should mention tools like Julius which are designed for the analysis of quantitative data using natural language. The idea is that they provide a friendly, natural language interface to something like a spreadsheet, which can perform calculations and create visualizations on request, without the need for the user to know how to write code or formulas. This is a popular application and indeed has been mainstreamed into general-purpose chatbots like ChatGPT, at least for paying users.

We have no reason to question the reliability of the calculations performed by Julius. We do wonder what the true value added is by such tools, since a large part of the role of a quantitative analyst is not simply to perform calculations but to critically assess the results and identify the interesting questions to ask of the data. We are skeptical that a qualitative researcher armed with a tool like Julius would be a good substitute for a quantitative researcher.

### ATLAS.ti

ATLAS.ti is a well-established qualitative research tool designed to for the storage, coding, and analysis of long-form interviews and other documents. It was an early adopter of LLM functionality, introducing an auto-coding function that relied on OpenAI's GPT API. This essentially shares the body text with the model and for each sentence asks 'What is this about?', with ChatGPT then providing a one- or two-word label. This is a kind of summarization task at a micro scale and should be a lower-risk application of an LLM than skipping straight to high-level tasks at scale. Moreover, it readily fits into an existing

evaluation workflow and provides intermediate outputs (a coded transcript) that can be verified before proceeding to draw conclusions.

It is worth noting that the approach to coding implicitly followed by this tool is inductive coding: there is no predefined list of categories and ChatGPT simply proposes a label that seems to describe the content of the sentence based on its lexical understanding. While good results have been reported from auto-coding in this way, it is not hard to anticipate this failing in practice: given there are myriad (infinite?) potential categories a piece of text could belong to or be accurately labelled by, knowing which ones are relevant depends on the context. The alternative approach, deductive coding, could also be achieved by using a chatbot and prompting it to choose among a number of categories, or alternatively by using a classifier model that learns how to categorize texts based on examples (this might include transformer architecture to understand the text, but it lacks any ability to generate text of its own – it simply estimates the probability of the various categories).

## CoLoop

CoLoop is an emerging AI-powered qualitative research tool designed to streamline the analysis of interview transcripts, focus group discussions, and other textual data. Developed by Genei.io, CoLoop positions itself as an 'AI Copilot' for qualitative researchers, offering functionalities that span from transcription to advanced analysis. CoLoop integrates AI capabilities throughout the qualitative research workflow. It begins with an AI-powered transcription feature, whereby researchers upload audio or video files for automatic conversion to text. The system then automatically identifies speakers and allows for the specification of roles (e.g. researcher, participant), which enhances subsequent analysis capabilities. CoLoop uses an LLM to analyse the specific data uploaded by users for each project. The tool offers two main interfaces for analysis: the Analysis Grid and the Chat function. The Analysis Grid allows researchers to pose questions about the data and receive AI-generated summaries across different participants or segments, while the Chat function enables a more conversational approach to data exploration.

CoLoop's approach to qualitative analysis is primarily inductive, generating summaries and themes directly from the data without predefined categories. This is similar to ATLAS.ti's auto-coding function but operates at a higher level of abstraction. Instead of labelling individual sentences, CoLoop aims to provide thematic summaries and extract relevant quotes based on researcher prompts. CoLoop claims one of its strengths is its ability to maintain context across large datasets, supposedly handling up to 100 hours of interview data accurately (although this is unverified). It also offers features like automatic quote extraction and video clip generation, which can significantly speed up the reporting process. However, like all AI tools in qualitative research, CoLoop's outputs should be critically examined. While it can rapidly generate summaries and identify themes, researchers need to ensure these align with their research objectives and interpretive frameworks. The tool is best viewed as an aid to the researcher's analytical process rather than a replacement for in-depth qualitative interpretation.

## Summary

We have identified five fundamental evaluation tasks that these qualitative analysis tools offer: document summarization, interview coding, extracting claims, writing a thematic summary, and facilitating quantitative analysis.

In general, the impacts we expect across these different tools are very similar. In all cases we regard these as neutral for utility since they do not alter the evaluation design or the questions being asked. Likewise, they are neutral for propriety, since they do not mediate the relationship between the evaluation, the substantive project operations, and the wider community.

They all offer the potential to increase feasibility by accelerating timescales and reducing costs. On the other hand, they all pose a risk to accountability if these tools are used uncritically by evaluators who substitute their use for real substantive reflection on the evaluation questions.

The most controversial criterion is accuracy. Quantitative analysis gets a green pass because the LLM typically facilitates the evaluator's own work by drafting code or proposing formulas, which the evaluator will naturally verify; there is little risk of a direct hallucination. We rate interview coding and claim extraction as amber, because, while there is some risk of hallucination, the LLM is given a very narrow task and the results are readily verified by the evaluator themselves. We rate document summarization and thematic summary as red. This is not to say there are no viable applications of these tools, just that there is a substantial risk that the LLM invents or misrepresents material, so careful human verification will be needed for the foreseeable future.

**Table 3: Deriving insights: traffic light ratings**

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| Document summarization | | | | | |
| Interview coding | | | | | |
| Extracting claims | | | | | |
| Thematic summary | | | | | |
| Quantitative analysis | | | | | |

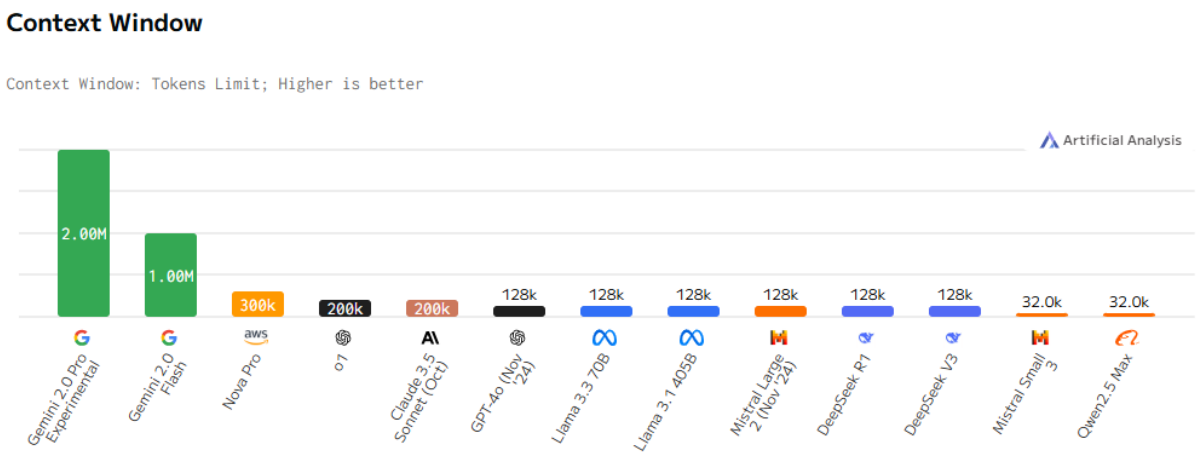
### 3.2.4 Synthesizing findings

Summarizing, coding, and extracting information from a single interview or document is one thing: performing this for an entire corpus of documents is quite another. For that reason, we separate out tools that focus on deriving insights from whole bodies of evidence from the tools above that primarily focus on analyzing an individual piece of evidence.

While the strength of transformer-based LLMs as opposed to previous generations of language models (like recurrent neural networks) is that they can handle long sequences of text and retain the relevant information to understand, say, references, the early versions still had very limited context windows. Crudely, the context window is how far back in the conversation the LLM can store in memory. When it launched in 2023, ChatGPT initially had a context window of 4,096 tokens, equating to roughly 3,500 words. Thus, while it had a very sophisticated understanding of language and could generate plausible text, it could only sustain a short conversation or answer questions about a single, short document.

That said, progress in addressing this limitation has been rapid and by early 2025 the leading models could handle over 100,000 words, with some reaching the millions.

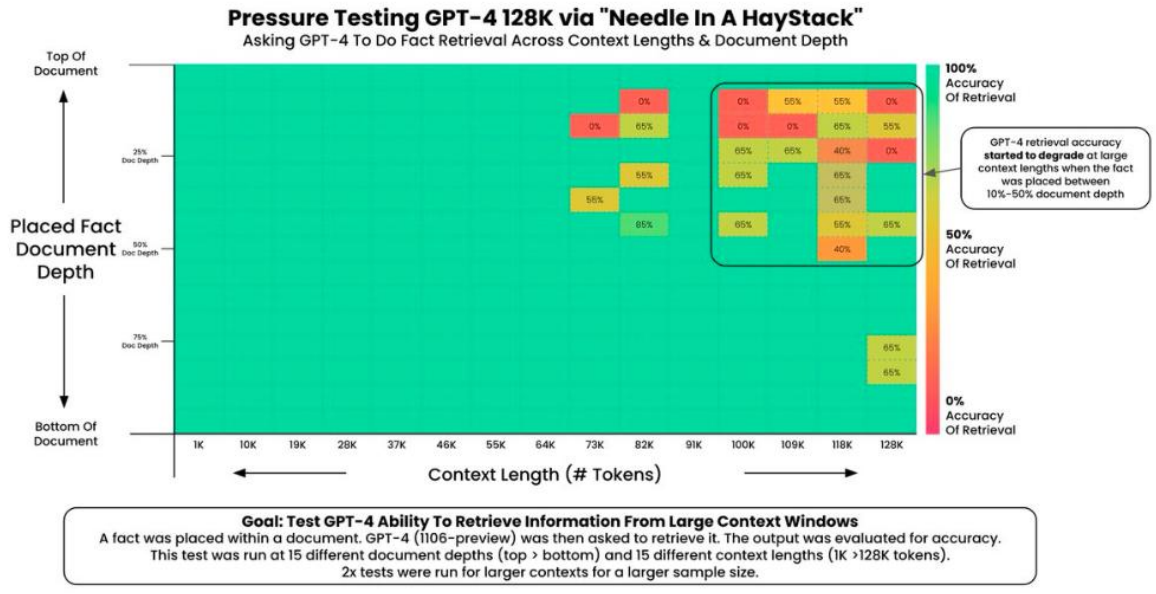
**Figure 3: The context windows of leading LLMs**




However, simply extending the context window is not the whole solution and being selective about which words it pays attention to is one of the key design elements that have allowed LLMs to perform so well. Therefore, while hundreds of thousands of words may now be available to a model, it is precisely by regarding only a part of those data at any one time that allows the model to function. This is all very simplistic, but empirical studies have found that the models do indeed struggle to surface relevant information in their context window where the context window is long.

Arize conducted the needle in a haystack test on GPT-4 – inserting a small snippet of text in a larger document and asking GPT-4 to use that snippet to answer a question. In the chart, the colour reflects the retrieval accuracy of the model, given the length of the text (X axis) and where in the text the snippet was placed (Y axis). They found performance degraded after 64,000 tokens (around 50,000 words), particularly when the information was inserted early in the document (i.e. when lots of other text came between the relevant fact and the question being posed).

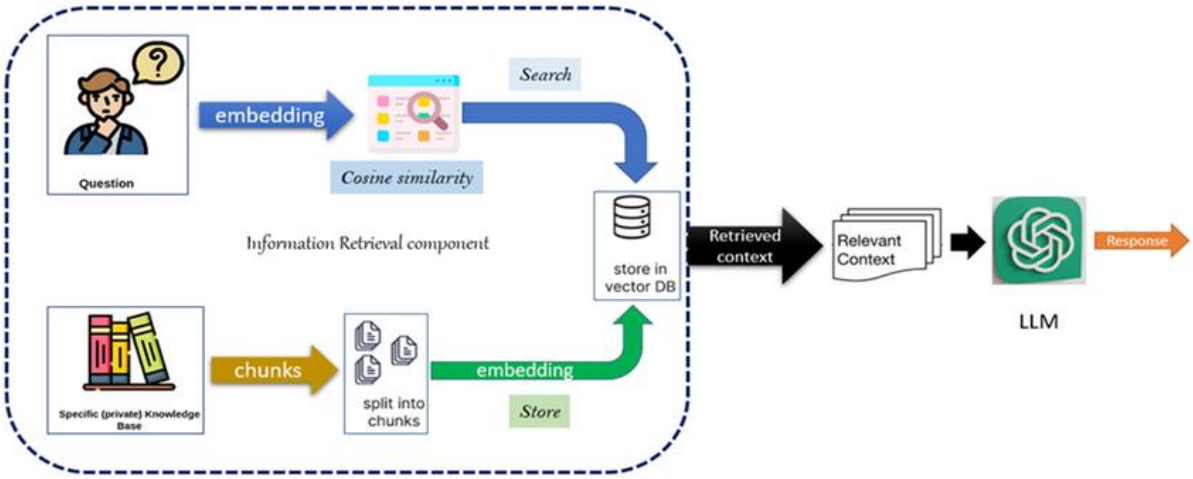
**Figure 4: GPT-4's performance on the needle in a haystack test**



Regardless of how effective longer context windows are, the consensus remains that they are not the whole solution; the true objective of a longer context window is to permit a longer and more nuanced dialogue, rather than to simply paste all the material of interest in the context window. The more robust solution, which allows even LLMs with relatively short context windows to use a document context of arbitrary size, is a RAG system.

A RAG is essentially a database containing the documents of interest, where they have been pre-processed so that they are already in their embedding form (i.e. numerical representations of words). When prompted, the LLM can then effectively search the database, extract the relevant sections of text from a very large number of documents, and use these to formulate its final response.

**Figure 5:[3] Anatomy of a RAG system**




## Using foundation models

Most of the major LLMs provide some sort of built-in RAG functionality, albeit usually only for paying users. ChatGPT has Custom GPTs, Claude calls them Projects, and Google provides NotebookLM for free (at the time of writing). In all these cases, the user is required to put together their own document set by creating a new instance and uploading the documents of interest to it directly. Plug-ins like Microsoft's Copilot or Google's Gemini promise to turn an organization's entire document library into a searchable RAG-style system.

Our experience with these systems has been mixed and does not match the hype surrounding them. We found it easy to build demonstration RAGs with a handful of documents and, when asked some open, general questions about them, the LLM can produce typically impressive-looking results. However, simply directing the LLM's attention to a set of documents does not in itself prevent hallucination, nor will the LLM be able to answer questions based solely on the contents of the documents, even if prompted to do so. This is because the LLM still has all of the information embedded in it via training at its disposal. Conversely, even with these relatively small RAG systems, important omissions quickly become apparent. Simple tasks like, for example, listing all the documents in the store were often not answered fully without repeated prompting. When retrieving substantive information from them we found that, while a lot of relevant content was surfaced, a great

[3] Figure sourced from: Abdelazim, H., Tharwat Waheed, M., and Ammar, M. (2023) 'Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)'. *International Journal of Advanced Computer Science and Applications*: 14. 10.14569/IJACSA.2023.01411135.

deal was missed and some was misattributed or muddled (e.g. a project was correctly identified but attributed to the wrong country of operation).

There may be several reasons for the disappointing performance. One is simply that the limitations of LLMs flow from their architecture and therefore bolting on a RAG may mitigate them to some extent but does not solve them. Another is that, since these systems are so expensive to run – and access to the LLMs themselves is therefore explicitly rationed – perhaps providers are not truly executing a comprehensive search but rather terminate the process after a small number of hits. This could be tested by an organization building and hosting its own RAG system, but the process is extremely involved and would require a team of data scientists and engineers.

All this does not mean there is no value in RAG systems or that they do not work at all, but they undoubtedly require careful tailoring of inputs and prompts, close supervision of the process of generating answers, and extensive manual verification of the results. In common with our other advice, we recommend limiting their use to discrete tasks within the evaluation workflow where you have a clear idea of the expected output.

We should also mention in this section LLM-enhanced search tools like Perplexity or Copilot's Bing web search. In a sense, these are 'half a RAG' systems. They do not maintain a vector representation of the entire internet, but rather use LLMs to expand upon the search term entered by the user, which is then executed in a conventional way. They can be very useful to find information, particularly when you can describe but not name the thing you are looking for.

### STORM

STORM is an open-source application designed to create an automatic literature review by performing an LLM-enhanced web search and then synthesizing the results. It defines a series of different 'agents' to get the job done, but they are ultimately powered by GPT-4o or earlier models. It quickly produces impressive-looking results, as in our example on micro-credit. It follows the structure you might expect, highlights the right kinds of issues, considers a range of sources, and draws conclusions on a series of sub-questions.

However, it returns only a limited number of sources drawn from all across the web – i.e. not just academic journals – so at least in its current form is unlikely to be a suitable substitute for the exhaustive journal database searches normally performed for literature reviews. Moreover, while the article generated reads well and seems to have surfaced many relevant facts, it has also surfaced some irrelevant ones and it is unclear if this is a representative selection of the literature or a fair representation of the balance of the articles included. It also lacks narrative coherence: the paper seems to conclude that micro-credit has no positive economic impacts, but still concludes by considering policy recommendations to promote micro-credit. Having said that, such criticisms could be levelled at many literature reviews written by humans, and we believe STORM could ultimately provide a very helpful starting point for a human researcher to build on.

### PaperGuide

PaperGuide is a commercial tool very similar to STORM, which is designed for finding papers and writing literature reviews with as little as a single prompt. It quickly generates reasonable-looking results, but the resulting document was much more cursory than the one produced by STORM. Although it claimed to have found thousands of relevant articles

online, the output was based on only the top 10 papers (although we acknowledge this may be a limitation of the free demo). PaperGuide also has options to include various standard sections of an academic essay (methodology, findings, limitations, etc.), which could usefully give more control back to the user. In the end, however, we should flag that the automatic literature review generated by PaperGuide reached the opposite conclusion to STORM on the same topic. This is a salutary reminder of how unwise it would be to simply pick one of these tools, hit a button, and make uncritical use the output.

### Humata

Humata is an application where you can upload documents and ask a chatbot questions about them, ultimately generating summaries or reports based on them. It is aimed at academic/technical users, who can focus on a single document or a set of documents together. Our experience with the tool was typical – the chatbot quickly surfaced much relevant information, but it also missed some relevant content and hallucinated (or at least exaggerated) other content. What is really appealing about Humata, however, is that the interface shows the generative AI outputs on one side of the screen, while showing the source document on the other, highlighting the cited passages. This interface invites the human researcher to work in tandem with the AI, checking its claims and using its outputs selectively.

### Summary

All these tools promise one application – to substantially automate the literature review process. Our views are similar to those for the various stages of deriving insights from individual documents: this does not have any obvious implications either way for the usefulness of the evaluation or the fairness with which it was conducted. There is potentially a huge gain in feasibility, cutting out time and cost, so we rate this green. It could also plausibly expand the scope of literature reviews, making them more comprehensive and representative and thus enhancing accuracy; however, the ever-present risk of hallucination presents a significant threat to accuracy, so we rate this amber overall. Finally, accountability we rate red, since there is a risk that human evaluators surrender their judgment and control to these black box processes.

**Table 4: Synthesizing findings: traffic light ratings**

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| **Literature review** | | | | | |

## 3.2.5 Disseminating learnings

Once the substantive evaluation process is concluded and the results written up, there is still the question of how to best communicate those results to a range of audiences. It is commonplace to have several versions of the final report: a technical report, a policy brief, and a public-facing blog that cover the same content but to a different level of detail and perhaps emphasizing different elements (research process vs. findings). This is the sort of task LLMs are eminently suited to doing well: taking one sequence of words and converting it into another with the same meaning is essentially a form of translation. At the most basic level, simply asking an LLMs to rewrite a text will correct spelling and grammar mistakes since they never make them in their output (unless prompted to do so). Moreover, these copyediting and redrafting tasks are a low-risk application, involving relatively small amounts of text that can be readily verified by the human researcher.

## Foundation models

We found this to be a task that the standard chat interfaces could do a good job of even with minimal prompting. Taking a research abstract and simply asking 'Rewrite this as a brief aimed at policymakers' or 'Rewrite this in a form suitable for young students to understand' immediately produced results in the correct style. We still observed a tendency to invent content – particularly detail not present in the original text – but this could be mitigated by more extensive prompting or, ultimately, by manual review.

## SciFocus

SciFocus is a product to facilitate academic essay writing. It has some of the functionality discussed above to brainstorm, review literature, and summarize findings, but it emphasizes the writing-up stage of the process and has a large number of modules for redrafting various types of input texts (articles, abstracts, emails, etc.) in various ways (make wordier, formalize, convert into bullets, etc.). While it claims to harness recent LLMs, the small sample of results we generated was not impressive, with the software making very few changes to the text and not really improving it. There were very few options to tailor the results.

## ProDream

ProDream is another tool aimed at academic authors – primarily undergraduate students – and offering very similar functionality: outlining, editing, citation, and proofreading. We cannot comment on how effective it is since the tool is currently available on an invitation-only basis.

## Summary

In this section we identify one basic application of these tools: copyediting. More specifically, we set aside the substantive job of writing up, which we believe the analysis and synthesis tools cover, and focus here on redrafting the final write-up for different purposes and audiences.

In this case we do not believe the criterion of propriety is relevant since this does not relate to the execution of the evaluation itself. We rate green for utility and accountability, because producing a wider range of research communication products tailored to different stakeholders' needs should make the evaluation more useful and allow a wider range of people to engage with it. As usual we rate green for feasibility because of the potential economies. We rate only amber for accuracy: there are always risks of hallucination, but these are limited when effectively transcribing one finished text from one style to another without making substantive judgements.

### Table 5: Disseminating learnings: Risk ratings

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| **Copyediting and redrafting** | | | | | |

## 3.3 Horizon scan / emerging applications: what is to come on AI use in evaluations?

### 3.3.1 The coming year

Since we began work on this report, major new advances in LLM technology have been announced. First and foremost is a new generation of 'reasoning models', led by OpenAI's o1, which are essentially LLMs with an extra loop whereby they prompt themselves. The modelling technique builds on the success of prompt chaining and similar approaches where the LLM is directed to think through the problem step by step and talk out its entire thought process (which, after all, is the only thought process it is capable of). That process of having the model produce a large amount of intermediate 'reasoning' text before arriving at a final response has essentially been embedded into the product: these are not a fundamentally new type of model (they are essentially another software engineering trick), but can be very effective and users report noticeably better results from reasoning models.

Their main aim is to improve LLMs' problem-solving ability, i.e. on problems requiring reasoning and logic. Despite the extravagant claims being made about the effectiveness of these new tools, it is not immediately clear why better logical reasoning would improve performance on tasks to extract and synthesize information, which should largely depend on the model's language understanding alone.

OpenAI's new models have proved extremely costly to train and run, and quite apart from concerns around energy and water use resulted in the product being released to a new ChatGPT user tier at a near-prohibitive US$ 200 per month. Since then, DeepSeek, an AI start-up from China, made the second recent breakthrough, effectively demonstrating that it was possible to massively reduce the training costs (although this might be partly because they were largely replicating OpenAI's original research). Whatever may be the case, perhaps prompted by this, OpenAI has very recently granted even free users of ChatGPT some access to o1.

Off the back of such models, new research-oriented AI tools are being released. In February 2025, OpenAI announced the release of deep research, a tool similar to those reviewed in the synthesizing findings section, which scours the internet for relevant information and produces a research report. For the moment it is only available to top-tier users so we cannot speak to its efficacy directly. The tool has received positive reviews alongside Google's identically named new deep research, which offers similar functionality but, we suspect, without employing a reasoning model to get the job done.

### 3.3.2 The far future

Tools based on reasoning models are already becoming available and we can expect them to become more widespread over the coming months. What about the further future? The honest answer is, this is anyone's guess.

It has been widely reported that LLM development is running into diminishing returns, as adding yet more data and building ever-more complex models does not lead to better 'next word' predictions (at least not as measured by the common benchmarks, e.g. passing various professional exams).

Indeed, the advent of reasoning models might be seen as an implicit recognition of the fact that the pure language modelling approach has run out of road. Having the model effectively

prompt itself behind the scenes is a way of squeezing better performance out of the same underlying language model, at the cost of greater amounts of processing. Despite this, Sam Altman (like other tech leaders) continues to insist that AGI is on the horizon and the sorts of technologies OpenAI is developing will lead us all there.

We believe that LLMs will be components in increasingly useful tools. However, we note that the overwhelmingly more successful approach to adoption so far has not been to use an extremely advanced model and let it do all the work. Instead, success has come by taking targeted inputs from the LLM into a traditionally engineered software product, retaining a large amount of hard-coded, explicitly programmed behaviours.

Reasoning models might appear to casual users to operate more autonomously, but ELIZA reminds us that human users of technology are credulous and easily fooled. Moreover, while a reasoning model might produce better responses (effectively by making more attempts that are hidden from the user) the name is a misnomer. In reality, the model still does not have any independent frame of reference or a concept of truth and falsehood: all it can ever do is generate sequences of text that seem to best fit the context of the proceeding words. Thus, as Gary Marcus says, 'There is no principled solution to hallucinations in systems that traffic only in the statistics of language without explicit representation of facts and explicit tools to reason over those facts'.

We may be wrong. Perhaps our 'general intelligence' *is* an emergent property of a much cruder system that on the face of it just juggles words. Perhaps all human intelligence is ultimately mediated via language and, as Sam Altman claims, 'we are all stochastic parrots'. These problems have troubled philosophers for centuries and may prove to become more practical concerns over the coming years. The fact that Geoffrey Hinton, one of the founding fathers of neural networks – the foundational technology underlying LLMs – has serious concerns regarding the rapid development of AGI gives us pause for thought. But our expectation is that LLMs will find their place as one among a range of useful data science approaches, rather than ultimately supplanting humanity itself as the dominant form of intelligence on the planet.

# 4 How can SSC reap the benefits of AI innovations?

## 4.1 What new innovations should SSC prioritize adopting?

In line with our advice throughout section 3, we believe that the most appropriate early applications of LLMs should be things are basically linguistic tasks, which essentially can all be conceptualized as translation jobs:

- Between natural (human) languages;
- From natural language to programming language;
- From audio to text; and
- From formal voice to informal or from technical style to a lay audience.

These are tasks that LLMs are readily suitable for and are easy to quality control and verify.

The next category of uses would be for ideation, where they can be employed for any application where the LLM is used to enrich the thought process of the user, who retains control of the final output. Some applications we reviewed were:

- Survey design;
- Stakeholder gaming;
- *First draft* literature review; and
- *First draft* meeting minutes.

The third category would be using LLMs to generate substantive evaluation outputs that are narrow, targeted, and easy to verify, for example:

- Coding an interview; and
- Extracting citations.

SSC should be most cautious about adopting applications where the LLM attempts to carry out large parts of the evaluation process with no opportunity for humans to scrutinize the individual steps being taken. These would include:

- Document summarization;
- Literature search; and
- Thematic summarization.

In general, our advice for adopting these tools for evaluation purposes very much echoes this advice from data.org on the use of generative AI in general: these can be very useful tools but it is essential that human evaluators remain in charge. People should only adopt and use these tools to the extent that they understand them, their limitations, and the ways in which they are likely to go wrong.

## 4.2 What does this imply in terms of training?

While it goes without saying that evaluation staff will need operational training in how to use any new tools effectively, this will not be sufficient and provided alone could even be counterproductive. We have found that widely touted 'good practices' for using LLMs turn out to have limited effectiveness, depending on the context. Moreover, in many ways the technical community of practice (including the developers of the models themselves) does not understand how they work well enough to produce practical guidance that will be consistently useful. In a worst-case scenario, evaluators might learn how to push the right buttons to generate an output that then turns out to be not only of lower quality than they would have produced via a traditional evaluation workflow but that draws misleading conclusions.

Therefore, we would recommend that all staff, including managers and senior staff, receive some high-level training on the foundational technologies underlying LLMs, so that they can come to their own view on when LLMs are likely to be an effective tool.

## 4.3 What are the risks and how can SSC avoid causing harm?

The ethical debate around LLMs has broadly considered two categories of risk: harm to individuals and harm to society. Many of the harms to society are not intrinsic to the technology and rather relate to how this technology is being developed in our current socioeconomic context. For example, profit-maximizing development by large corporations leads to the externalization of many costs (e.g. water consumption and carbon emissions resulting from energy use) and to the appropriation of vast amounts of data at no cost (i.e. without remunerating the original creators). These risks are all valid but are in a sense ephemeral: after all, we could still develop LLMs and avoid all these issues by organizing our society in a different way. Without putting too fine a point on it, we put all this in the class of political debate, and your view will depend on your beliefs and values.

A second category of social harms is more specific to LLMs themselves: the risk of perpetuating harmful biases in society, be that racism, sexism, homophobia, or a myriad other prejudices. Again, any individual's assessment of an individual risk and where to draw the line on what is and is not an acceptable viewpoint will depend on their views on these issues. The broader point is that LLMs, like other forms of machine learning, can only make predictions for the future based on historical precedent. If we think the historical precedent is problematic, feeding it into our models for them to learn from is a bad idea. There are practical ways to address these problems, and the tech companies developing these tools have actually done far more to address such problems than they have on the economic and environmental issues mentioned first. This has included vetting and editing training data, as well as conducting reinforcement learning and other techniques to 'train out' the bias.

Much of the debate around harms to individuals concerns how the human user perceives and responds to the LLM they are interacting with. There is a philosophical discourse around the extent to which LLMs can be said to have understanding and how they can create meaning in dialogue with a human interlocutor, some of which is relevant to the debate above about what tasks we can expect an LLM to perform well. More prosaically, there is a risk that humans attribute human intelligence and emotion to the LLM: it is very conceivable that some people could fall in love, or become dependent, or be driven to self-harm or death by an LLM producing harmful content.

Evaluators at SSC should be fully aware of the range of risks associated with LLMs and in particular the risk that they perpetuate harmful beliefs and stereotypes. However, in all candour, we believe the ethical risks surrounding LLMs are relatively limited in this context. SSC is a professional working environment and SSC staff are professionals using a tool to complete their work with a specific goal in mind; they are not naïve end users who could end up interacting with countless tools in numerous ways. Nor is the subject matter of SSC evaluations or the process of conducting the evaluation likely to centre largely on sensitive or difficult issues, although these can arise in any context.

By far the biggest practical risk that SSC should bear in mind is not that these tools will cause direct harm to a staff member or produce results that damage broader society. It is simply that they will be ineffective and produce outputs that appear to get the job done but are actually wrong.

The optimal way to manage that risk is to adopt LLM tools incrementally, use them to perform discrete parts of the evaluation process and not skip steps, and always have close oversight and manual correction of their outputs. While much of the ambition of the tech sector seems to be to completely replace human labour in the execution of certain tasks, we believe that human evaluators will be in the driving seat of the evaluation process for the foreseeable future, and probably forever.

# 5 Recommendations

## 5.1 How efficient are each of the steps in the current SSC evaluation process?

While the current process is structured, there are opportunities to enhance efficiency, particularly in time-intensive tasks such as document review, data analysis, and report writing. Our investigation did not include a quantitative assessment of efficiency gains, such as measuring the time saved by using AI compared to performing tasks manually. Our findings nonetheless revealed significant opportunities for improvement in the planning, conducting, and reporting phases of SSC evaluations. During the planning phase, evaluators dedicate substantial time to identifying and reviewing extensive volumes of information. The conducting phase presents challenges in the form of time-intensive analysis of both qualitative and quantitative data. Furthermore, in the reporting phase, evaluators noted the considerable time required to comprehensively incorporate all relevant information and findings into evaluation reports. These findings highlight key areas where efficiency could be enhanced in the evaluation process.

SSC evaluators have already begun integrating AI and LLM tools, primarily CanChat, into their workflow for tasks involving unclassified data. These tools show promise in streamlining activities across all evaluation phases, from generating evaluation questions to tailoring communication products.

## 5.2 Where are the opportunities to leverage LLM tools?

Based on the findings from interviews with SSC evaluators, there are significant opportunities to leverage LLM tools across all phases of the evaluation process. Interviews with evaluators revealed several promising opportunities for leveraging AI in the evaluation process. For the planning phase, some evaluators reported that AI tools have assisted in streamlining document reviews, enabling more efficient identification of information relevant to specific evaluations. In the conducting phase, those who have experimented with AI noted a reduction in time spent on tasks such as extracting themes from transcripts and identifying pertinent indicators for evaluation questions. The reporting phase also showed potential for AI-driven improvements, with evaluators indicating using AI in drafting initial text for report sections, tailoring dissemination materials for diverse audiences, and enhancing the overall writing process. These findings suggest that AI technologies could offer significant efficiencies across various stages of the evaluation lifecycle.

To maximize these opportunities, we recommend implementing a phased approach to LLM integration, starting with tasks that show immediate efficiency gains. Developing standardized guidelines and policies for using LLM tools in each evaluation phase will be crucial to ensure consistency and quality. This should include guidance on data privacy and security, ethical considerations, and quality control measures.

We advise investing in targeted training programs to improve evaluators' proficiency with key LLM tools that SSC invests in, focusing on prompt engineering, understanding AI capabilities and limitations, and best practices for specific evaluation tasks. Creating a centralized prompt library can streamline LLM tool usage and promote best practices across the evaluation team.

If SSC is to monitor its progress in utilizing LLM tools, it is essential to establish a robust feedback mechanism to continuously assess and improve the effectiveness of LLM tools in the evaluation process. Additionally, addressing data privacy and security concerns by developing clear protocols for using LLM tools with sensitive information is paramount for data safeguarding as well as for evaluators to understand their options and responsibilities.

SSC should promote collaboration and knowledge sharing on AI use in evaluations. This could include establishing communities of practice within and across departments, sharing best practices, maintaining a prompt library, and developing resources for evaluators.

However, challenges persist, including concerns about output accuracy, data privacy, and the need for clear usage guidelines. To fully leverage the potential of AI in evaluations, SSC should prioritize developing a comprehensive AI strategy that includes targeted training programs, improved use case tracking to better monitor and quantify efficiencies, and user-friendly mechanisms for best-practice sharing among evaluators.

## 5.3 What LLM tools are available?

A wide range of LLM tools have been developed and new ones are coming to market all the time; they are far too numerous to enumerate all of them, although we believe we have captured a great portion of those aimed primarily at evaluators. These individual tools are reviewed in section 3.2.

We have also reviewed the relevant functionality in foundation models and find they are capable of many tasks “out of the box” with minimal prompting. Many of the LLM tools on the market could be approximated in-house by a developer, although this often wouldn’t be cost effective.

## 5.4 What tools would we recommend adopting, in what priority order?

Rather than recommending individual products, we have identified a dozen or so discrete tasks from the evaluation process that these tools claim to perform throughout the evaluation process, mainly from SSC’s “conducting” phase. We have made broad recommendations about whether these tasks are ready for introduction into the evaluation workflow, on the basis of whether they are likely to improve or harm an evaluation on the basis of the AEA criteria.

We find the lowest risk applications to be at opposite ends of the evaluation flow: either assisting in evaluation design, the development of survey instruments etc. or the writing up of findings for diverse audiences. In both cases, the LLM is not tasked with making any substantive evaluative judgements and is more used as a device by the human evaluator for ideation – this is a very healthy place to start.

We also see significant upsides and limited downsides in using LLM tools for data collection, although we have some concerns about these tools divorcing the evaluators further from the evaluated and the communities they are both meant to serve.

The most controversial areas are using LLMs to perform the analysis itself and synthesize findings. In a sense these offer the greatest promise and create the greatest excitement, but there are significant risks: who is in control of and accountable for an evaluation where the substantive judgements have been delegated to AI? More fundamentally, can we be

confident those judgements are correct? The risks of hallucination are significant and have yet to be resolved. Moreover, we would question how big the upside is? In many evaluations much of the time and cost involved is in planning and data collection – the work of reviewing and assessing the raw material is relatively limited and seems a suitable last bastion for human ownership and control.

### Table 6: Summary of ratings for all LLM tools

| Evaluation process/criteria | Utility | Feasibility | Accuracy | Propriety | Accountability |
|---|---|---|---|---|---|
| **Evaluation design** | | | | | |
| Survey design | | | | | |
| Stakeholder gaming | | | | | |
| **Data collection** | | | | | |
| Avatar-conducted surveys | | | | | |
| Chatbot-facilitated interviews | | | | | |
| Machine transcription / translation | | | | | |
| **Deriving insights** | | | | | |
| Document summarization | | | | | |
| Interview coding | | | | | |
| Extracting claims | | | | | |
| Thematic summary | | | | | |
| Quantitative analysis | | | | | |
| **Synthesizing findings** | | | | | |
| Literature review | | | | | |
| **Disseminating learnings** | | | | | |
| Copyediting and redrafting | | | | | |