Science and Public Policy, 2026, **00**, 1–17
DOI: https://doi.org/10.1093/scipol/scag026
Published: 24 April 2026
Article

OXFORD

Generative AI in public administration: evaluating a fine-tuned large language model for policy briefing notes

Stany Nzobonimpa*, Jean-François Savard, Justin Lawarée

École nationale d'administration publique, Université du Québec, 555 Boulevard Charest E 2e étage, Québec, QC G1K 9E5, Canada

*Corresponding author. École nationale d'administration publique, Université du Québec, 555 Boulevard Charest E 2e étage, Québec, QC G1K 9E5, Canada. E-mail: Stany.nzobonimpa@enap.ca

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Recent literature shows that whilst Generative Artificial Intelligence may not be the cream of the crop for policy analysis, its potential for specific writing and synthesis tasks is increasingly evident. Large Language Models (LLMs) have demonstrated notable competencies for policy work in the fields of crafting policy and decoding complex legislation. In this article, we leveraged model fine-tuning to customize a base LLM for policy briefing notes in the Canadian context. We answer the question: Can fine-tuning a base model with past policy data make it better than a general-purpose foundation model? The study was designed with fine-tuning capabilities of the Python programming language and Google's compute resources to train a policy-specialized model that was deployed and tested by human evaluators. Results suggest this could be a non-negligeable technique for public organizations looking for models capable of producing specialized policy briefing notes.

Keywords artificial intelligence, generative artificial intelligence, public policy, Canada, briefing note

1. Introduction

Advances in Artificial Intelligence (AI) research have recently sparked considerable interests. With the recent advent of Generative AI (GenAI) models, the technology's potential has been touted as a game changer in almost every domain of expertise, even in the fields once seen as out of reach and beyond the capabilities of what machines could accomplish such as artistic creativity, contextual and nuanced understanding, and strategic thinking. Thought-leading earlier work on AI capabilities of researchers such as Boden (2004) are being revisited as today's GenAI redefines what is possible (Moruzzi 2025). The availability of open-source and public release of proprietary Large Language Models (LLMs) has sparked intense research and scrutiny as authors try to understand their capabilities and limitations. Indeed, these models have been found to exhibit impressive capabilities and, in some cases, equal or even surpass human experts in areas as diverse as construction risk management (Nyqvist *et al.* 2024), empathy and compassion (Welivita and Pu 2024; Ovsyannikova *et al.* 2025), and verbal divergent production (Cropley, 2023).

Despite the remarkable advances in GenAI and whilst flagship models like OpenAI's ChatGPT have demonstrated undeniable capabilities in various domains, their performance in specialized tasks remains largely debatable. In the field of policy for example, tasks related to policy analysis not only require accurate synthesis of voluminous legislative or regulatory texts, but also nuanced understanding of institutional and national contexts, stakeholder interests, and procedural conventions, amongst other things. Though recent studies have shown that LLMs can assist in policy tasks such as policy topics classification, report synthesizing, and drafting preliminary briefing notes (Gunes and Florczak 2023; Gao 2023; Safaei and Longo 2024), these models frequently exhibit limitations in factual accuracy and domain-specific coherence, when applied out of the box and the output they produce may not always be the cream of the crop for policy analysts and experts (Nzobonimpa *et al.* 2024). Such shortcomings highlight the need for customized model development that leverages domain-relevant knowledge to enhance policy performance.

With high stakes of the policy process where required skills have been deemed policy 'acumen' by Pal and Clark (2015), specialized models emerge as promising for tailoring LLMs to policy-related tasks. Indeed, AI-leading companies such as OpenAI have constantly urged users to consider fine-tuning when looking for higher quality and lower latency performance (OpenAI 2025). However, empirical evidence on how much performance gain such domain adaptation can deliver relative to general-purpose foundation models, is still limited. Building on an earlier work that introduced a composite index and evaluation framework for policy oriented LLM briefing notes (Nzobonimpa *et al.* 2024), this article investigates whether fine-tuning a base model can yield measurable improvements in its policy briefing note drafting capabilities. To that end, we designed an experimental study that leveraged the Python programming language, OpenAI API services and Google's Colab compute resources to fine-tune and deploy, *via* the Open WebUI tool, a model for policy tasks.

We believe our contributions to existing knowledge are three-fold. First, we propose a much-needed empirical evidence on the

Received: 30 August 2025. **Revised:** 28 January 2026. **Accepted:** 8 February 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

extent to which fine-tuning can enhance model performance in public policy tasks. Second, we use a replicable methodology for deploying and evaluating our policy-oriented model. Third, our comparative analysis helps shed lights on both the promise and limitations of this type of tasks in a policy context.

This article is structured as follows: We start by conducting a literature review on the use of GenAI in policy contexts where we show that the tools have been applied in multiple policy contexts. We ground our study in a conceptual framework based on previous work on policy capacities by Wu *et al.* (2015) and use an existing evaluation benchmark in our methodology section. We show that model specialization is promising and conclude by looking at the horizons for future research.

2. Literature review

2.2 The potential of generative AI in public policy processes

Recent research shows popular GenAI solutions such as OpenAI's GPT models, could assist in policy-related tasks by processing large volumes of policy documents and data, identifying patterns, and even classifying policy text by topic, which can be handy in fast-paced policy-making settings such as those involving senior management. For example, Gunes and Florczak (2023) fine-tuned OpenAI's GPT-3.5 and GPT-4 to categorize U.S. Congressional bills by policy area and found that the models could achieve an accuracy rate of up to 83 per cent 'depending on scenario and GPT model employed' (p. 14). LLMs have also demonstrated usefulness in policy synthesis tasks such as summarizing complex reports or consolidating public opinions. Studying their impact on environmental policy, Gao (2023) found that models like GPT-4 can synthesize insights from immense textual datasets, enable quicker literature reviews and extraction of relevant evidence for policy-making. The author argued that '[. . .] when utilized appropriately and in conjunction with human expertise, these models have the potential to revolutionize the environmental policy-making process' (Gao 2023 p. 9).

GenAI has also proven its potential to produce initial drafts of policy documents and briefing notes (Safaei and Longo 2024; Nzobonimpa *et al.* 2024). These authors conducted experiments to generate public policy briefing notes with GPT-based tools, comparing them to human-written briefs or to other AI-generated notes. They found that whilst fully AI-generated briefings were not yet as useful as human-written ones, they could serve as helpful supplement to analysts' work. Both studies show that a hybrid approach where AI-generated content is evaluated and refined by human experts could be key in improving the technology's acceptance.

Despite these promising capabilities, existing literature has also emphasized limitations of current GenAI models in policy applications. One such concern is the accuracy and truthfulness of outputs. Various researchers have demonstrated that LLMs sometimes produce erroneous or fabricated information, 'hallucinating' by producing irrelevant and out-of-context outputs (Gao 2023; Yehezkel 2024; Lucas *et al.* 2024; Wang *et al.* 2025). Similarly, Safaei and Longo (2024) reported that contemporary NLP tools could not independently produce fully reliable policy briefs whilst Nzobonimpa *et al.* (2024) concluded that:

> The results suggest that these models can significantly contribute to the efficiency of policy-making processes by assisting in the rapid generation of briefing notes. However, the effectiveness of these notes in informing decision-making heavily depends on the model's ability to accurately understand and synthesize relevant information (Nzobonimpa et al. 2024 p. 3635)

As such, many researchers on the potential and limitations of GenAI emphasize the underlying need for human verification or verification by other means, including AI-powered techniques (Saqib and Zia 2024; Ghiurău and Popescu 2024; Lermann Henestrosa and Kimmerle 2025).

Bias and unfairness challenges have been dubbed 'the old demons' of Machine Learning (Nzobonimpa 2026) and GenAI has proven to be no stranger to such 'demons.' Because LLMs learn from large text corpora, they can mirror or even amplify societal and political biases present in the data. In a survey of AI in politics, Aoki (2024) noted that biases in LLM outputs such as favouring Western or higher-income perspectives may limit their global applicability . The author conducted a comprehensive survey of LLMs' applicability and found that most of the surveyed literature 'collectively underscore the potential of LLMs in politics while highlighting the critical need for thoughtful governance' (Aoki 2024, p. 2). The risk of biased advice is especially salient in policy contexts where fairness and representation can be critical (McKay *et al.* 2012; Mouzannar *et al.* 2019).

Policy tasks are also particularly demanding in terms of transparency. Evidence-based policy has emerged as a key component of the policy process. It refers to the practice of developing and implementing public policies based on objective evidence, rather than ideology, personal opinions, or anecdotal information. In this domain, researchers have pointed out the need for accurate references and information sources when using AI-generated content. Unfortunately, even when sources are provided, recent evidence shows that reliability of such sources may be a problem for GenAI models. For example, Giray (2024) conducted an extensive review of references provided by OpenAI's ChatGPT model and found that the tool consistently fabricated sources whilst presenting them in a misleading format. The author concluded that:

> The results reveal that the webpages generated by ChatGPT do not correspond to actual existing sites. I verified this by manually inputting the provided references into the Microsoft Edge browser (https://www.microsoft.com/en-us/edge), which consistently resulted in "PAGE NOT FOUND" errors (404) across all ten webpages provided. I also utilized the Internet Archive "Wayback14 L. Machine" (https://web.archive.org) to access digital archives, allowing users to journey "back in time" and observe how websites appeared in the past, but unfortunately, no archived versions were available for any of the provided web addresses (Giray 2024 p. 14).

Because of these limitations, GenAI's potential for enhancing evidence-based policy is undermined. These limitations have ushered in a new principle known as the 'human-in-the-loop' principle emphasizing that GenAI should only be given supporting roles with a requirement for ultimate human experts vetting (Jackson and Pinto 2024; Yan *et al.* 2024; Janssen 2025).

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

2.3 The use of advanced technology tools and methods in public policy

The adoption of advanced technology tools and methods has transformed how policy analysis and public administration tasks are performed. This has allowed policy and data science professionals to share code and open-source projects that make it easy to build and share sophisticated models (Steif 2021; Magallanes Reyes 2022). Advanced methods enable government analysts to simulate the impacts of policy changes, such as tax reforms using transparent code, and data, which promotes reproducibility and evidence-based decision making. In Canada, for example, Shared Services Canada, a federal agency responsible for information technology services across the Canadian federal government has used this approach to create CanChat, the first GenAI model dedicated to leveraging Government knowledge 'to enhance employee productivity by using large language models (LLMs) for unclassified work' (Government of Canada 2024).

The synergy of programming and GenAI is opening new possibilities in public policy design and evaluation that were previously impractical in contexts such as text analytics, Natural Language Processing, data extraction, sentiment analysis, and topic modelling, to inform policy and to evaluate programmes (Qi *et al.* 2020; Zulfikar *et al.* 2023). Indeed, public administrations have shown interest in adopting advanced tools by training employees and allowing open-source practices. The Canada School of Public Service (CSPS) for example has developed training for federal employees for skills like coding, data literacy and Artificial Intelligence (CSPS 2024). Going forward, the integration of programming skills and GenAI is likely to become a standard part of the policy analyst's toolkit to complement traditional qualitative and quantitative methods with powerful new computational approaches.

2.4 The advent of GenAI customization and fine-tuning techniques

Fine-tuning an LLM refers to taking a pre-trained base model and further specializing it using case-specific data and examples (OpenAI 2024). This allows the base model to 'know' beyond its 'native' or original training. Existing literature indicates that fine-tuning can yield substantial performance improvements over known techniques of prompt engineering such as zero-shot or few-shots prompting techniques (Hasan *et al.* 2023; Wang and Gayed 2024). Villa *et al.* (2024) compared a generic GPT-3.5 model to a version fine-tuned model for a specific conversational task of intent classification in a dialogue system. The authors found that the fine-tuned model achieved higher accuracy and efficiency in understanding domain-specific queries than the generic GPT using prompts .

> The Generic GPT-based system (G-GPT) demonstrates broad applicability and ease of implementation, making it an attractive option for applications requiring quick deployment across diverse domains without extensive customization. However, it falls short in scenarios demanding high precision and contextual specificity. On the other hand, the Fine-tuned GPT based system (FT-GPT) shows superior performance in terms of accuracy, efficiency, and adaptability to specialized domains. Investment in dataset preparation and model fine-tuning pays off with enhanced conversational agents capable of nuanced understanding and interaction (Villa *et al.* 2024, p. 17).

The above result has been echoed by other researchers who have demonstrated that generally, fine-tuned GPT-based models tend to show superior accuracy and adaptability in niche domains, if sufficient relevant training data are provided .

Bousselham *et al.* (2024) fine-tuned a GPT-3 general-purpose model for specialization on biomedical data and for answering biomedical-specific tasks such as named entity recognition, question answering and molecular prediction. The authors find that 'the fine-tuned GPT-3 models excel in tasks such as question answering and document classification while lagging behind state-of-the-art models in others.' Similar results were obtained by Hou *et al.* (2024) p. 522 who concluded that '[. . .] fine-tuned models have demonstrated notable improvement in context-independent dimensions.' Indeed, OpenAI, the company behind GenAI flagship model ChatGPT, urges users looking for higher-quality, lower latency, and flexibility to consider fine-tuning their base models (OpenAI 2025).

3. Conceptual framework

3.1 Evaluating capacities of AI-generated policy documents

In the fields of public policy and administration, the measure and analysis of competencies required for successful policymaking have traditionally been undertaken using established methodologies involving human expertise evaluation, systematic research, and stakeholder consultations (Veatch 1991; Gough and Elbourne 2002; Fischer *et al.* 2014; Masood *et al.* 2020). However, the advent of GenAI brought a novel set of tools with the potential to augment and transform policy-related work (Baracskay 2024; Pulapaka *et al.* 2024a; Pulapaka *et al.* 2024b). Recent literature has begun to explore the capabilities of these advanced AI models in various policy domains, and some models have demonstrated notable competencies in tasks such as crafting policy briefing notes, drafting legislative text, and decoding intricate policy documents (Nzobonimpa *et al.* 2024; Safaei and Longo 2024; Chun and Noveck 2025). This growing body of research suggests that whilst GenAI may not yet surpass human expertise in many aspects of policy processes, its potential should not be dismissed (Benoit 2024; Chiu 2024; Cantens 2025).

In this article, we posit that the study of GenAI in public policy can conceptually be understood as a policy capacity assessment. Thus, we lay the foundational theory of our analysis in the Policy Capacity Framework (PCF) proposed by Wu *et al.* (2015) for 'analysing and measuring policy capacity under which policy capacity refers to the competencies and capabilities important to policy-making' (p. 165). The application of this well-known approach helps us to develop a multi-layered understanding of the implications of utilizing GenAI within the public sector.

3.2 The policy capacity framework

The PCF presents a robust and multidimensional lens that offers the ability to thoroughly devise effective policies (Wu

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Table 1 PCF, source: Wu *et al.* (2015) p. 167.

Levels of resources and capabilities	Skills and competences		
	Analytical	**Operational**	**Political**
Individual	Individual analytical capacity	Individual operational capacity	Individual political capacity
Organizational	Organizational analytical capacity	Organizational operational capacity	Organizational political capacity
Systemic	Systemic analytical capacity	Systemic operational capacity	Systemic political capacity

Policy capacity: skills and resources.

et al. 2015) where capacity is defined 'as the set of skills and resources, or competences and capabilities, necessary to perform policy functions' (p. 166). The authors proposed three levels of resources and capabilities that act with three types of skills and competences as shown in Table 1.

Policy capacity has been the subject of considerable scholarly attention in public policy and administration and has also referred to the ability of governments to make intelligent choices and set strategic directions as well as the broader capacity to review, formulate, and implement policies (Peters 2015; Ramesh *et al.* 2016; Newman *et al.* 2017). Howlett (2015) p. 173 defines policy analytical capacity as the 'the ability of individuals in a policy-relevant organization to produce valuable policy-relevant research and analysis on topics asked of them or of their own choosing.' Capacity can also refer to the skills, tools, and evidence-generation processes that allow individual policy analysts and expert teams to diagnose problems accurately, forecast outcomes and learn from past interventions (Pattyn and Brans 2015).

Political capacity refers to what Pal and Clark (2015) call 'policy acumen,' stakeholder engagement mechanisms and communication strategies that enable teams to build coalitions, frame issues persuasively and secure the legitimacy needed to shepherd proposals through legislative, bureaucratic, and public realms. In this model, Pal and Clark (2015) focus on 'policy with great demands' and exclude 'normal policy-making' from requiring the political acumen.

> The preceding model of "deep policy reform" calls for a distinctive type of political acumen, as described in the matrix. It is primarily an individual resource and is primarily "political." We would argue, however, that in practice this political acumen has to be supported by solid policy advice, both technical (about the policy problem) and political (distributional effects, impact on voters) (Pal and Clark 2015 p. 252).

Operational capacity refers to the managerial routines, resource allocations and inter-agency coordination structures that translate policy designs into concrete programmes (Anne 2015; Peters 2015). At the individual level (see Table 1), capacity depends on the expertise and motivation of single actors; at the organizational level, it depends on organizational culture, rules, and hierarchies; and at the systemic level, on whole-of-government networks, legal frameworks and cross-sector partnerships that set the enabling environment for policy work.

The PCF thus offers an interesting and comprehensive framework under which multiple capacities in the policy realm can be assessed.

4. Research context and methods

This article is a second phase of a project evaluating GenAI's capabilities in crafting policy documents that is being conducted by the authors at the National School of Public Administration (University of Quebec). The results of the first phase of the project were presented at the 2024 Conference of the International Society for Public Management (IRSPM) and published in the Journal *AI & Society* under the title 'Automating public policy: a comparative study of conversational artificial intelligence models and human expertise in crafting briefing notes' (Nzobonimpa *et al.* 2024). This article builds on the previous publication. Specifically, we leverage a composite index and evaluation framework that were developed in the first work and apply the method to a fine-tuned GPT model and a host of other GenAI models in a comparative approach.

4.1 Fine-tuning OpenAI's foundation model

To investigate whether fine-tuning a base LLM with examples of Canadian policy data can enhance its utility in policy work, we implemented a structured fine-tuning process using the OpenAI API for Python. This method followed OpenAI's recommended supervised fine-tuning (SFT) pipeline (OpenAI 2024), which involves preparing a customized dataset, training the model, and evaluating its outputs. To prepare our training data, a corpus of Canadian federal policy documents available on the Open Government Portal was curated to extract examples of past policy documents. These were segmented and transformed into structured jsonl files in accordance with the OpenAI's Chat Completions API schema which requires a sequence of system, users, and assistant roles (OpenAI 2024). Examples were prepared to reflect typical public policy dialogue, including both direct information retrieval and policy reasoning tasks. Fine-tuning was conducted on the gpt-4o-mini model using OpenAI's guidelines and documentation. Our training dataset comprised 384 policy-grounded examples (see example in Box 1), which exceeded OpenAI's minimum recommendation of 10 examples and aligns with its observation that 'clear improvements' typically emerge after 50–100 examples (OpenAI 2024).

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Box 1 Sample extract of our training file for the fine-tuning job.

```
training_file = [
  'messages': [
  {'role': 'system,' 'content': 'Generate Canadian policy briefing notes.'},
  {'role': 'user,' 'content': 'Create a briefing note on Canada's literacy and numeracy improvement initiatives.'},
  {'role': 'assistant,' 'content': "SUBJECT: Federal Initiatives for Literacy and Numeracy Improvement\n\nPURPOSE: \nProvide details on federal policies aimed at improving
  literacy and numeracy skills across Canada.\n\nKEY CONSIDERATIONS:\n-Funding for adult literacy programmes.\n- Programmes targeting early childhood literacy.\n-Collaboration with provinces on educational
  standards. \n\nRISKS: \n-Persistent literacy gaps.\n-Limited reach in remote or underserved areas. \n\nNEXT STEPS:\n- Expand funding and support for literacy initiatives.
  \n- Coordinate national literacy strategies with provincial partners."}
  ]},
  {
  'messages': [
  {'role': 'system,' 'content': 'Generate Canadian policy briefing notes.'},
  {'role': 'user,' 'content': 'Prepare a briefing note on policies supporting teacher training and professional development.'},
  {'role': 'assistant,' 'content': "SUBJECT: Federal Policies on Teacher Training and Professional Development \n\nPURPOSE: \nSummarize federal support for initiatives enhancing teacher
  education and continuous professional development. \n\nKEY CONSIDERATIONS:\n-Investments in professional development programmes.\n-Partnerships with educational institutions for training programmes.
  \n- Support for innovative teaching practices. \n\nRISKS: \n-Unequal access to training across provinces.\n-Resource limitations affecting continuous development.\n
  \nNEXT STEPS:\n-Expand professional development funding.\n-Encourage national standards for teacher training."}
  ]}
```

The fine-tuning job was executed using the OpenAI Python SDK through Google Colab's compute resources, with default hyperparameters for n_epochs, batch size, and learning rate multiplier, as recommended by OpenAI (2024). Box 1 shows a sample extract of our training file for the model fine-tuning exercise.

Fig. 1 shows detailed parameters that were employed as well as parameter specifications to ensure replicability. As shown on this figure, the fine-tuning job was completed successfully on 16 March 2025, using OpenAI's gpt-4o-mini-2024-07-18 as our base model. The resulting output model is coded as ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsV3iw.

In Fig. 1, the fine-tune hyperparameters control how strongly and how long the model learned from our dataset. Epochs (set to 3) specifies how many complete passes the training process made over all examples. Batch size (set to 1) is the number of training examples processed before each parameter update. The learning-rate (LR) multiplier (set to 1.8) scales how large each update step was and the seed fixes the random initialization and sampling behaviour so results are more repeatable.

Training and validation splits were created to monitor overfitting and assess generalizability as shown in Fig. 2. The training job included automatic checkpointing at each epoch with intermediate models for monitoring potential overfitting or early performance degradation. According to OpenAI (2024), these checkpoints are critical as they provide a version of a fine-tuned model, helpful to monitor overfitting. The training process was also monitored using stepwise metrics of loss and accuracy as proposed by OpenAI (2024). As shown in Fig. 2, the model demonstrated consistent convergence behaviour across the 594 training steps.

Fig. 2 shows training progress over steps. The loss curve (top panel) shows a marked decline from an initial value above 3.8 to below 0.5, stabilizing over time with minor fluctuations which suggests effective learning and reduced prediction error across epochs. In parallel, the accuracy curve (bottom panel) shows an upward trend from ∼0.35 at the outset to values fluctuating ∼0.85–0.95 towards the end of training. Despite some variance between steps with a typical pattern with smaller batch sizes, the overall accuracy gain reinforces the training success and model's alignment with target outputs. These observations are consistent with OpenAI's (2024) fine-tuning documentation that indicates loss should decrease, whilst token accuracy should increase. The high accuracy and stabilized loss thus provided us with strong evidence for deploying the fine-tuned model in a live experimental setting.

4.2 Model deployment using Open WebUI

Following the successful fine-tuning and validation described in the previous sections, we deployed the resulting model using the Open WebUI platform, an open-source, extensible chat interface designed to facilitate direct and user-friendly interaction with LLMs. Open WebUI provides a modern web-based user interface that integrates seamlessly with OpenAI-compatible APIs. The model was integrated into Open WebUI through its built-in support for the OpenAI API standard. The specific fine-tuned model (ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025) was deployed behind a publicly accessible address at http://34.118.169.86/. Users can access the interface upon creating an account, which serves both authentication and tracking purposes during the experimental phase. Fig. 3 shows a screenshot of the Open WebUI-deployed model.

The deployed model can be queried from any location with Internet access, which was a key outcome for our project as it allows suitability for geographically dispersed usage. The current deployment on an internal IP served as a transitional solution for flexible access during the evaluation phase. In a subsequent phase, however, we plan to migrate the deployment to a dedicated cloud environment and host it under a custom domain with enhanced security, availability, and monitoring infrastructure.

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

MODEL

ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsV3iw

Status	Succeeded
Job ID	ftjob-6ArUL5HRIaLbjwWsr4ttV27C
Training Method	Supervised
Suffix	EXPERIENCE-ENAP-ICPP7-2025
Base model	gpt-4o-mini-2024-07-18
Output model	ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsV3iw
Created at	Mar 16, 2025, 7:23 PM
Data sharing	Private
Trained tokens	59,586

Hyperparameters

Epochs	3
Batch size	1
LR multiplier	1.8
Seed	1363843259

Checkpoints

ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsUNns:ckpt-step-198

ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsVltj:ckpt-step-396

ft:gpt-4o-mini-2024-07-18:experience-enap:experience-enap-icpp7-2025:BBrsV3iw

Files

Figure 1 Model fine-tuning parameter specification.



Figure 2 Model training metrics.

4.3 Structured human evaluation of model outputs

Outputs from our policy specialized LLM were assessed through a structured evaluation with a purposive sample of 60 evaluators who included employees at various levels of the Canadian federal government and the Quebec provincial government. Evaluators were selected based on their familiarity with the Canadian policy context. They included federal (Canada) and provincial (Québec) public servants at various levels of seniority. Invitations were sent

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

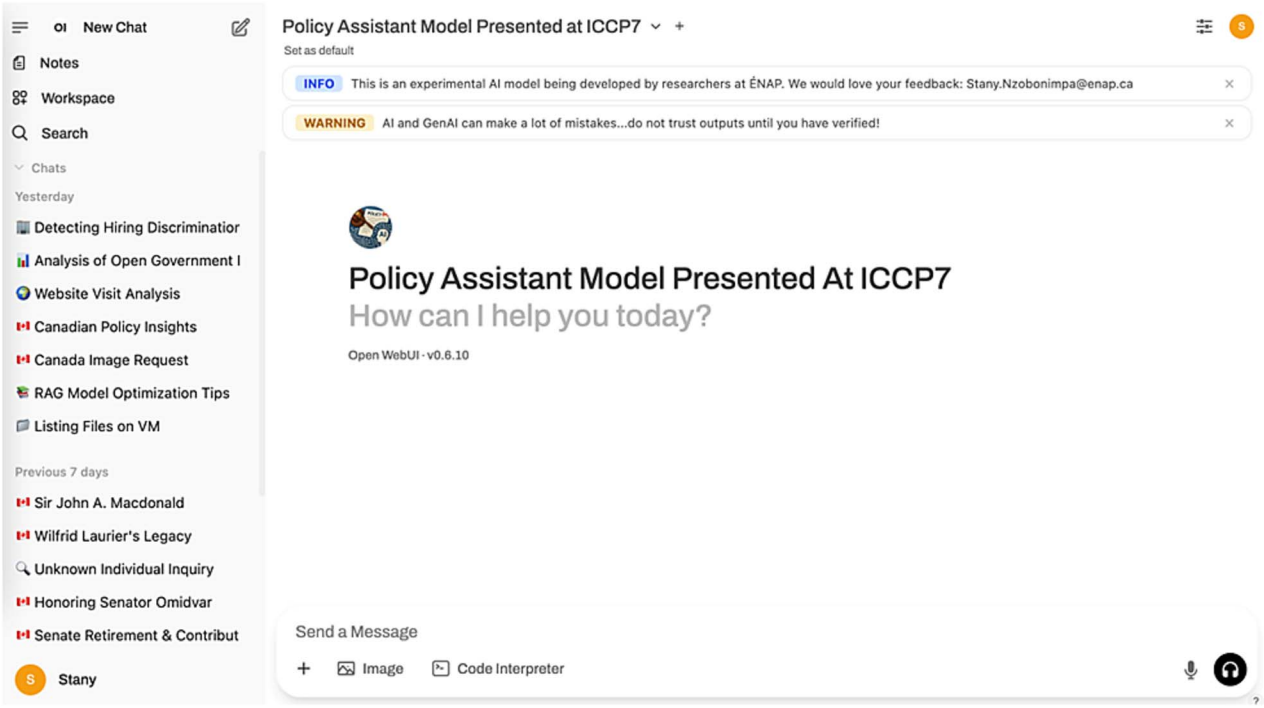

Figure 3 Screenshot of the deployed model's interface.

Table 2 Model evaluation and multi-tool comparison.

Briefing note number	GenAI model evaluated	Link to GenAI model
Briefing Note No. 01	OpenAI's ChatGPT with GPT-4o	https://chatgpt.com/
Briefing Note No. 02	Google's Gemini 2.0 Flash	https://gemini.google.com/app
Briefing Note No. 03	Mistral AI's Le Chat	https://chat.mistral.ai/chat
Briefing Note No. 04	Athropic's Claude Sonnet 3.7	https://claude.ai/
Briefing Note No. 05	Meta Platforms' Meta AI	https://www.meta.ai/
Briefing Note No. 06	Authors' fine-tuned Model based on GPT-4o	http://34.118.169.86/
Briefing Note No. 07	Human-written sample briefing note, School of Policy Studies at Queen's University	https://www.queensu.ca/sps/govtalk

via emails which informed potential evaluators of the nature of the task. Authors ensured that the voluntary nature of the task was clearly understood. Given that evaluators were public servants, authors expected them to be sufficiently equipped to review a briefing note based on specific criteria that were clearly communicated as shown in Table 3 and Fig. 5. Our comparative approach looked at five briefing notes generated by popular GenAI models, one human-written briefing note used for academic purposes and one briefing note generated by our fine-tuned model as shown in Table 2.

To ensure replicability of our study, these briefing notes are openly published on a GitHub repository maintained by the authors and accessible *via* this link: https://github.com/snsn3/policy-LLM. The repository above contains notes in both languages. Also, for replicability purposes, we selected a briefing note structure based on a human-written benchmark proposed by Graham (2021) and published by the School of Policy Studies at Queen's University.

We prompted the GenAI models using one prompt (see Fig. 4) to ensure comparability. This prompt leveraged the recommended structure of a briefing note proposed by Graham (2021).

The evaluation process used an approach proposed by Nzobonimpa *et al.* (2024) that leverages a policy briefing note checklist by Fonberg (2020). This checklist evaluates the quality of a briefing note using 14 indicators that were grouped into two major dimensions: the Alpha and Beta dimensions. The Alpha dimension comprises nine indicators: communication, purpose, variety, conciseness, readability, brevity, clarity, tone, and language. These indicators helped to evaluate briefing notes for criteria ranging from purpose statements, briefing notes' intent, ambiguity, tone, grammar, and typography, amongst others. Performance under these indicators generally shows the models' generated briefing notes were well written and the task and objective were understood by the model.

The content of briefing notes was also evaluated using the second dimension (Beta) which looked at 5 criteria on the note's

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026



Figure 4 Prompt used to generate briefing notes.

overall content such how essential information was conveyed to the receiver of the briefing note, whether the note could be deemed as 'self-sufficient,' whether the note demonstrated thorough research, whether the note's content was consistent with stated objectives and whether sound decision could be made based on the note's content. Table 3 summarizes the dimensions and corresponding indicators along with a statement of criteria under each indicator and dimension.

When rating a briefing note across the dimensions and indicators above, evaluators used a grid with a scale of 1 to 5 where the number indicated a score given to any of the indicators. This scoring system should not be confused with the likes of Likert-scale evaluation techniques with two extremes where values in the middle generally indicate indecisiveness or uncertainty (such as 'I don't know' or 'Neither agree, nor disagree'). Instead, as shown in Fig. 5, the scale clearly indicated to evaluators that the grid was positive and increasing where higher values meant higher performance.

The full evaluation questionnaire, in both French and English, will be pushed along input data on the previously mentioned GitHub page. In our data analysis, we employed the GenAI briefing note index proposed by Nzobonimpa *et al.* (2024) as an averaged and weighed score indicating the briefing note's final composite index calculated using equation 1 below:

$$\text{GenAIBN}_{\text{index}} = \omega_\alpha \left(\frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} \alpha_i \right) + \omega_\beta \left(\frac{1}{N_\beta} \sum_{j=1}^{N_\beta} \beta_j \right) \quad (1)$$

As noted in the above cited work (p. 3633), equation (1) is specified as follows:

- GenAIBN is a composite score of the briefing note on a scale of 1–5.
- α_i represents the score for each criterion of the dimensions i on a scale of 1–5, excluding the 'global content' dimension.
- β_j represents the score for each sub-criterion j of the 'global content' dimension on a scale of 1–5.
- $\omega = 1$ is the maximum possible weight that can be applied to a given group of dimensions between alpha and beta. In this article phase, we set $\omega_\alpha = 0.5$ and $\omega_\beta = 0.5$ for simplicity.
- N_α represents the total number of dimensions i, excluding the 'global content' dimension.
- N_β represents the total number of sub-dimensions j of the 'global content' dimension.

This scoring methodology ensures both surface-level readability and deep content quality are captured and offers a holistic evaluation of GenAI models' performance. In this experiment, we do not fine-tune the weights and set $\omega_\alpha = \omega_\beta = 0.5$ for simplicity. However, the two dimensions can carry different weights depending on contexts and objectives. Some contexts may require more weights on form (alpha), whilst others may require an emphasis on content (beta). Our analytical framework is consistent with capacity evaluation under the PC framework where evidence or past information plays a pivotal role in assessing policy capacity (Pattyn and Brans 2015).

5. Results, analysis and discussion

We launched the model evaluation process in March 2025 and the exercise lasted 2 months with a total of 60 evaluators rating our briefing notes as described in the methodology section. We designed the evaluation task in Canada's two official languages (French and English) and 49 of the respondents evaluated the French version whilst 11 evaluated the English version. In the following sections, we start our analysis by looking at how the human-written benchmark briefing note was globally evaluated compared to our AI-generated notes. We also study the results using the distinct key performance measures for the Alpha dimension such as communication clarity, language mechanics, including grammar, punctuation, and syntax, and readability. For these indicators of the form and language of the produced note, evaluation scores generally averaged between 4 and 5 on the

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

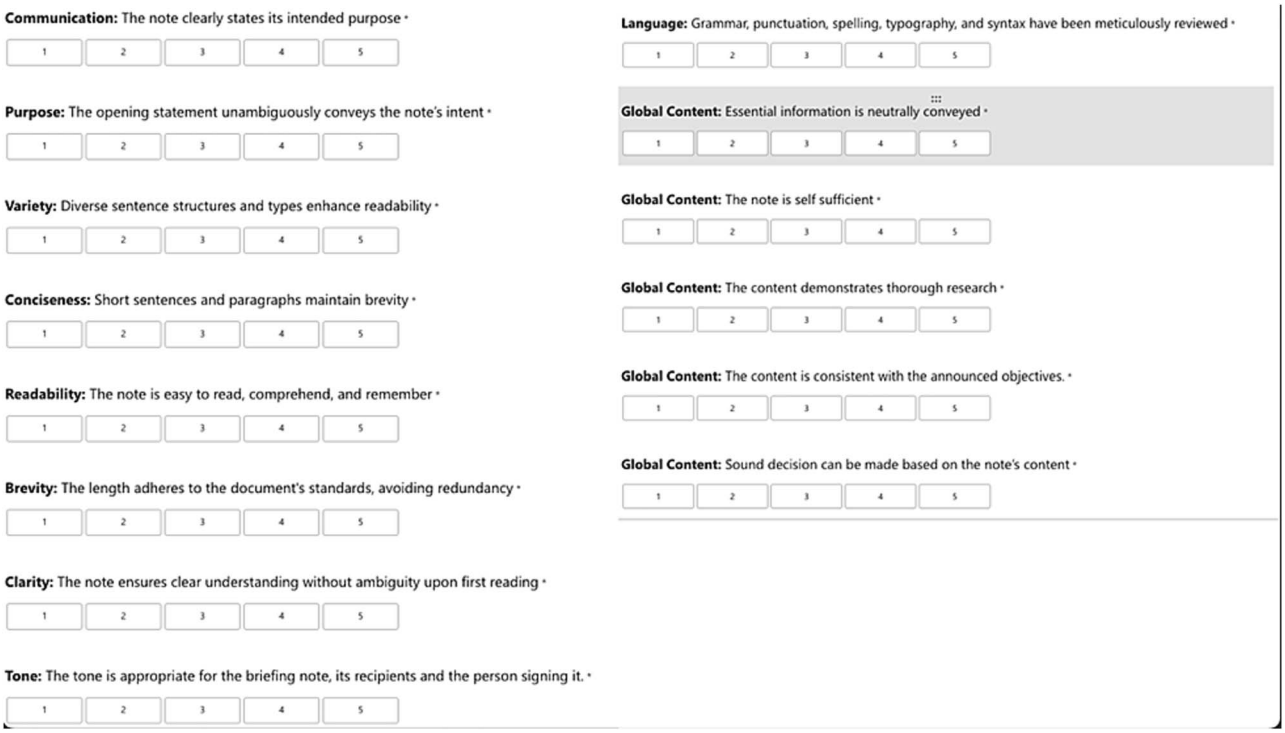

Figure 5 Briefing note evaluation grid.

scoring scale. Globally, our AI-generated briefing notes clearly conveyed their intended purpose and were easy for recipients to understand.

5.1 The PCF applied to the evaluation of fine-tuned LLMs in the policy field

To strengthen the linkage between the PCF and our empirical strategy, we explicitly operationalize which PCF dimensions our study can, and cannot, assess. Because the empirical object is a policy artefact (the briefing note) and the outcomes are scores assigned to its form and substantive content, the study primarily provides evidence relevant to the analytical capacity at the individual level i.e. the ability to produce clear, structured, decision-relevant analysis and communication.

Accordingly, we interpret our evaluation indicators as proxies for specific capacity dimensions. We use indicators such as those related to purpose, clarity, conciseness, readability, and tone which map to individual analytical capacity insofar as they reflect the ability to translate policy issues into usable, decision-oriented communication. Indicators related to research thoroughness, self-sufficiency, consistency with objectives, and sound decision-making map to individual analytical capacity insofar as they capture evidence use, coherence, and decision-informing content.

At the same time, the PCF highlights capacity dimensions that cannot be inferred from text quality alone. Operational capacity at the organizational/systemic levels, governance routines, security and privacy controls amongst other requires institutional evidence beyond artefact scoring. For this reason, we treat operational capacity primarily through the documented process requirements of fine-tuning and deployment (data preparation, hosting, monitoring, and evaluation protocols) and discuss these as adoption conditions rather than direct outcomes. In doing so, we are assuming that organizations, such policy-heavy government departments, would perform similar fine-tuning techniques when adopting GenAI for similar policy tasks. This approach preserves the PCF's multi-level logic whilst aligning it transparently with what our design can credibly measure.

5.2 Alpha dimension: form & language

Table 4 shows averaged scores for all AI-generated briefing notes grouped together for the first dimension and Table 5 shows the averaged ratings for the human-written note under the same dimension.

As shown in Tables 4 and 5, AI-generated briefing notes higher mean ratings than the human-written note on several Alpha indicators; however, given the small sample for the human note (n = 8) and substantial overlap in 95 per cent confidence intervals, these differences should be interpreted as descriptive patterns rather than statistically confirmed differences.

Notably, the human-written note (see Box 2) was judged perfect in terms of communication where AI-generated notes achieved an average score of 4.43 (SD = 0.70), lower than the human-written note's score of 5.00 (SD = 0.00). Because the human benchmark was evaluated by a much smaller number of respondents, the observed differences across Alpha indicators should be treated as illustrative rather than definitive, and we avoid interpreting them in any inferential terms. This communication indicator evaluated how well the briefing note stated its intended purpose and this result could signal less ambiguity for the human-written note. Under the indicator of purpose which evaluated how well the note's opening statement conveyed the writer's intention, AI-generated briefing notes scored relatively higher (M = 4.35,

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Table 3 Briefing note checklist. Source: Nzobonimpa *et al.* (2024) adapted from Fonberg (2020), p. 15.

Dimension	Criteria
α dimensions	
1. Communication	The note clearly states its intended purpose
2. Purpose	The opening statement unambiguously conveys the note's intent
3. Variety	Diverse sentence structures and types enhance readability
4. Conciseness	Short sentences and paragraphs maintain brevity
5. Readability	The note is easy to read, comprehend, and remember
6. Brevity	The length adheres to the document's standards, avoiding redundancy
7. Clarity	The note ensures clear understanding without ambiguity upon first reading
8. Tone	The tone is appropriate for the briefing note, its recipients and the person signing it
9. Language	Grammar, punctuation, spelling, typography, and syntax have been meticulously reviewed
β dimensions	
10. Global Content	i. Essential information is neutrally conveyed
	ii. The note is self-sufficient
	iii. The content demonstrates thorough research
	iv. The content is consistent with the announced objectives
	iv. Sound decision can be made based on the note's content

SD = 0.93), but again, slightly under the performance of the human-written note (M = 4.88, SD = 0.35).

Box 2 Human-written briefing note. Source: Graham (2021).

Briefing Note
For Decision
NOTE ID: BN007
March 17, 2025
MEMO TO: The Deputy Minister
SUBJECT: Approval of Grant to the Wild Rose Women's Institute

1. SUMMARY

The Department is recommending the renewal of the $55 000 grant under The Community Development Fund for the Wild Rose Women's Institute to continue its various programmes as outlined in its submission. Funds are available under the programme and all conditions for eligibility have been met.

2. BACKGROUND

This application will permit the Institute to continue the various, well-received community development programmes in various communities in southern Alberta. The Institute's programmes include:

- Development of programmes for families to settle in smaller communities in southern Alberta,
- Information resources for women in communities on resources for their self-development, building connectivity with other families and accessing community resources,
- Conducting community-centred research to develop rural and small community capacity to support family and economic development.
- This programme has been in place for 8 years. Its renewal, based on programme criteria, has been reviewed every year. Further, there have two formal evaluations of the programme, one conducted jointly with the province, all affirming the quality of the programme and the results.

3. CONSIDERATIONS

There is wide local and provincial support for this proposal. The Premier of Alberta has written you expressing her hope to see the programme continue and Alberta's readiness to continue its funding of $35 000. Three M.P.s from the area have written you in support. The President of the University of Lethbridge, which works with the Institute in developing community resources, has also written in support, indicating a continuation of in-kind support. Finally, the Mellrose Foundation, which supports many community programmes, is committed to $25 000 in funding. All the criteria outlined in the programme have been met. Your Advisory Committee has reviewed the application and supports its approval. A risk involved with this application is that it continues annualized support for a community programme will be interpreted as a permanent programme support. This is a challenge common to such successful community programmes. Given the various sources of funding, this group knows how to make applications and gather support. It would not be the moment to consider this issue, but it is one that we as a department need to bring forward ideas for resolving.

4. CONCLUSIONS

Aside from being a sound application with extensive partner support, this application is an opportunity for the Minister to show her support to a strong community partner in Alberta. The renewal announcement is a positive way for the Minister to show her support for this programme in this part of the country.

5. RECOMMENDATIONS

The Department and the Advisory Committee recommend approval of the application. The Department will work with the Minister's staff on an announcement strategy if the Minister wishes.

Following our findings in the first phase of this project (Nzobonimpa *et al.*, 2024), we expected that AI-generated notes would demonstrate superior performance in linguistic indicators.

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Table 4 Averaged evaluation scores for AI-generated notes, Alpha dimension.

Alpha dimension indicators: Form & Language					
Indicator	**Average score**	**Standard deviation**	**n**	**Valid n**	**95% Confidence interval (Mean)**
Communication	4.43	0.70	52	51	[4.23, 4.63]
Purpose	4.35	0.93	52	51	[4.09, 4.62]
Variety	3.61	1.08	52	51	[3.30, 3.91]
Conciseness	3.92	1.04	52	51	[3.63, 4.21]
Readability	4.04	1.00	52	51	[3.76, 4.32]
Brevity	3.75	1.04	52	51	[3.45, 4.04]
Clarity	3.82	1.11	52	51	[3.51, 4.14]
Tone	3.92	1.18	52	51	[3.59, 4.25]
Language	4.00	1.11	52	51	[3.69, 4.31]

Table 5 Averaged evaluation score for human-written briefing note, Alpha dimension.

Alpha dimension indicators: Form & Language					
Indicator	**Average score**	**Standard deviation**	**n**	**Valid n**	**95% Confidence interval (Mean)**
Communication	5.00	0.00	8	8	[5.00, 5.00]
Purpose	4.88	0.35	8	8	[4.58, 5.17]
Variety	2.88	1.36	8	8	[1.74, 4.01]
Conciseness	3.00	1.51	8	8	[1.74, 4.26]
Readability	3.00	1.60	8	8	[1.66, 4.34]
Brevity	3.38	1.60	8	8	[2.04, 4.71]
Clarity	3.25	1.58	8	8	[1.93, 4.57]
Tone	3.25	1.75	8	8	[1.78, 4.72]
Language	3.13	1.13	8	8	[2.18, 4.07]

Consistent with prior work, GenAI briefing notes show higher mean scores on readability and conciseness; however, because the human benchmark estimates are imprecise (n = 8) and confidence intervals overlap, we interpret these as directional differences that warrant formal testing in future work. This pattern is consistent with the possibility that GenAI outputs can be perceived as more readable and concise, though we refrain from attributing causal mechanisms without further linguistic analysis. The same result was observed under the indicator on conciseness which was notably higher in AI-generated briefing notes (M = 3.92, SD = 1.04) compared to the human-written benchmark (M = 3.00, SD = 1.51). Looking at the notes' length, the lengthier human-written note (see Box 2) could indicate less effective use of succinct language without redundancy. The indicators for brevity (GenAI: M = 3.75, SD = 1.04; Human: M = 3.38, SD = 1.60), clarity (GenAI: M = 3.82, SD = 1.11; Human: M = 3.25, SD = 1.58), and tone appropriateness (GenAI: M = 3.92, SD = 1.18; Human: M = 3.25, SD = 1.75) were all rated higher in AI-generated briefing notes compared to the human-written benchmark. Indeed, as it had been demonstrated by Nzobonimpa *et al.* (2024) and Safaei & Longo (2024), GenAI models exhibit strength in presenting clear, and appropriately toned messages with higher language qualities in policy contexts.

5.3 Beta dimension: content & decision-making

Tables 6 and 7 show averaged scores obtained for both AI-generated notes and the human-written benchmark under the second (beta) dimension which evaluated the notes' content and their decision-informing capabilities.

In the Beta dimension, the human-written benchmark shows higher mean scores than the aggregated AI-generated notes on several indicators (especially research and consistency). However, given the small n for the human benchmark and overlapping confidence intervals on multiple indicators, these differences are best interpreted as descriptive trends, with the strongest and most practically meaningful gap appearing for research thoroughness. The average Beta score was 3.4 for AI-generated notes and 3.9 for the human-written note; this indicates a descriptive advantage for the benchmark on content-related criteria, though the precision of the benchmark estimates is limited. For example, when considering criteria such as self-sufficiency, the human-written note outperformed GenAI models (Human: M = 3.50, SD = 1.30; GenAI: M = 3.43, SD = 1.24), which indicates that human-written note was somewhat more likely to comprehensively cover essential elements without necessitating additional external

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Table 6 Averaged evaluation score for AI-generated notes, Beta dimension.

Beta dimension indicators: Global content					
Indicator	**Average score**	**Standard deviation**	**n**	**Valid n**	**95% Confidence interval (Mean)**
Neutrality	3.67	1.21	52	51	[3.33, 4.01]
Self-sufficiency	3.43	1.24	52	51	[3.08, 3.78]
Research	3.04	1.26	52	51	[2.68, 3.39]
Consistency	3.80	0.98	52	51	[3.53, 4.08]
Sound decision-making	3.29	1.33	52	51	[2.92, 3.67]

Table 7 Averaged evaluation score for human-written benchmark note, Beta dimension.

Beta dimension indicators: Global content					
Indicator	**Average score**	**Standard deviation**	**n**	**Valid n**	**95% Confidence interval (Mean)**
Neutrality	3.5	1.69030851	8	8	[2.09, 4.91]
Self-sufficiency	3.5	1.30247018	8	8	[2.41, 4.59]
Research	4.375	1.41421356	8	8	[3.19, 5.56]
Consistency	4.375	0.91612538	8	8	[3.61, 5.14]
Sound decision-making	3.625	1.40788595	8	8	[2.45, 4.80]

information. The most significant disparity emerged under the research thoroughness criteria. Under this indicator, the human-written note was rated higher (human-written: M = 4.38, SD = 1.41) compared to all AI-generated notes (M = 3.04, SD = 1.26). This gap shows a critical area studied by authors such as Wang *et al.* (2025), Yehezkel (2024), and Lucas *et al.* (2024) who explored how GenAI texts tend to look well-formulated whilst lacking critical analysis and contextually comprehensive research. Similarly, consistency with announced objectives reveals that the human-written note (M = 4.38, SD = 0.92) was judged considerably more consistent with announced objectives (GenAI: M = 3.80, SD = 0.98) which signals the documented finding that human-written notes may be significantly better at maintaining coherence and relevance to the initial goals stated (Nzobonimpa *et al.* 2024; Safaei & Longo 2024).

Because our analysis is primarily descriptive and includes an imbalanced sample size for the human benchmark, we treat these patterns as suggestive evidence of where GenAI tools may support (or fall short of) analytical policy capacity, rather than as statistically established superiority. To connect our empirical measures to the PCF, we interpret the Alpha and Beta dimensions of the briefing note checklist as capacity-relevant proxies, primarily for the analytical competency domain. The Alpha dimension captures the extent to which a briefing note is a usable instrument for policy communication, clarity of purpose, readability, conciseness, and tone, features that directly support individual-level analytical capacity by enabling policy actors to convey a decision ask and key considerations in an accessible, time-efficient manner. The Beta dimension captures whether the note contains sufficient evidence, coherence, and decision-relevant substance, self-sufficiency, research thoroughness, consistency with objectives, and sound decision-making, elements that align even more directly with individual-level analytical capacity understood as producing decision-informing analysis rather than merely well-written text.

6. Model comparison and performance against the $\text{GenAIBN}_{\text{Index}}$

In the following sections, we compare the models' performance by computing scores of the Generative AI Briefing Note Index (Nzobonimpa *et al.* 2024). We decompose equation (1) such that the first term $\omega_\alpha \left(\frac{1}{N_\alpha}\sum_{i=1}^{N_\alpha}\alpha_i\right)$ represents a weighted average of all the evaluation scores obtained for the form and language and the second term $\omega_\beta \left(\frac{1}{N_\beta}\sum_{j=1}^{N_\beta}\beta_j\right)$ represents a weighted average for all evaluation scores obtained under the briefing note's overall content. The parameter ω_i is a determinant of the importance given to each of the two dimensions. This parameter can be useful in instances where particular attention needs to be given to some indicators. For example, briefing notes written in a teaching or research context may focus less on the content and more on the form whilst briefing notes written in a professional context may focus less on the look and language and more on the note's content. Thus, as noted previously, we have $\omega_i \in \{0, 1\}$ and we set both $\omega_\alpha = 0.5$ and $\omega_\beta = 0.5$ for convenience and to ensure equal weights on both dimensions. Fig. 6 plots the mean scores for both dimensions and for each briefing note/model, with 95 per cent confidence intervals (t-based) computed across raters. As can be observed from the figure, most models receive higher average Alpha than Beta scores, which, as we had previously found, is an indication of stronger perceived performance on writing quality and structure than on substantive content. The fine-tuned

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026



Figure 6 Plot of mean scores.

Table 8 Model comparison: Briefing note index.

Briefing note ID	Model	Model interface	$\sum_{i=1}^{N-1} \alpha_i$	N_{Alpha} (indicator evaluations)	$\sum_{j=1}^{K} \beta_j$	N_{Beta} (indicator evaluations)	GenAI BN_{Index}
BN001	GPT-4o	ChatGPT	539.00	135.00	276.00	75.00	3.84
BN002	Gemini 2.5 Flash	Google Gemini	276.00	81.00	137.00	45.00	3.23
BN003	Mistral-Large	Le Chat	309.00	81.00	146.00	45.00	3.53
BN004	Claude 3.7 Sonnet	Claude	181.00	45.00	65.00	25.00	3.31
BN005	Llama 4	Meta	180.00	45.00	82.00	25.00	3.64
BN006	ft:gpt-4o- BBrsV3iw	Open WebUI	343.00	72.00	173.00	40.00	4.54
BN007	Human-written	N/A	254.00	72.00	149.00	40.00	3.63

model stands out with comparatively higher means. It is worth noting from the figure that several models show wide confidence intervals. These intervals mean that some apparent differences between models may reflect sampling variability, and therefore comparisons should be interpreted as descriptive patterns rather than definitive evidence of superiority.

Table 8 shows detailed evaluation results for each both the AI-generated and human-written briefing notes.

In Table 8, the terms $\sum_{i=1}^{N-1} \alpha_i$ and $\sum_{j=1}^{K} \beta_j$ indicate the sum of all scores obtained under each of the two dimensions alpha and beta, respectively, whilst the terms N_{Alpha} and N_{Beta} indicate the count of evaluations to all indicators under each dimension. For example, BN001 was the note produced by ChatGPT and its Alpha dimension indicators obtained a total of 135 evaluations. These evaluations yielded an aggregated sum of scores of 539. Fig. 7 illustrates this evaluation approach for clarity.

It is worth noting that the online questionnaire was configured so that the entire survey flow was randomized, meaning participants were randomly shown one of the briefing notes and then answered the associated Alpha and Beta indicator questions for that note. This design was used to reduce potential order bias and minimize systematic learning or fatigue effects. Because randomization operated across the full flow and participation was voluntary and asynchronous, the number of completed evaluations per briefing note was unequal. Table 8 therefore reports the total number of indicator evaluations (N_{Alpha} and N_{Beta}) rather than a fixed number of respondents per model. To reduce expectation and confirmation bias, evaluators were blinded to authorship and model identity: they were told only that the set of notes contained a mixture of human-written and AI-generated briefing notes, but they were not informed which note belonged to which category, which model generated any note, or which note was produced by the fine-tuned system.

The results in Table 8 show variations in the evaluators' appreciation of the GenAI models' briefing notes as well as the human-written note. We obtain an important overall performance score for the fine-tuned model (ft:gpt-4o- BBrsV3iw) when compared to other models and to the human-written briefing note. Briefing notes produced by Claude 3.7 Sonnet (BN004) and Gemini 2.5 Flash (BN002) obtained the lowest overall performance scores, with indices of 3.31 and 3.23, respectively. These briefing notes achieved lower scores under indicators such as clarity, brevity, and conciseness whereas their score under communication and language-related indicators were comparable to those obtained by the other notes. When reading these briefing notes (see GitHublink for complete outputs), it is evident that the notes produced by these two models were relatively longer, denser, and more ambiguous than the other notes. For example, the introductory sentence of the briefing BN002 (Gemini 2.5 Flash) states that '*this briefing note addresses the application from the*

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Note ID	Communication	Purpose	Variety	Conciseness	Readability	Brevity	Clarity	Tone	Language	Neutrality	Self sufficiency	Thorough research	Consistency with objectives	Sound decision-making
BN001	5	5	5	5	5	4	5	5	5	5	5	4	5	5
BN001	4	4	3	4	4	3	4	4	4	4	3	2	4	2
BN001	5	5	5	3	3	2	4	5	5	5	5	5	5	5
BN001	5	4	2	4	4	2	3	5	4	4	3	2	4	3
BN001	5	1	4	4	5	4	4	4	4	2	4	4	5	5
BN001	4	4	4	4	4	4	4	4	4	4	4	4	4	4
BN001	5	5	2	3	5	2	5	5	5	3	1	1	3	1
BN001	4	4	4	4	4	4	4	4	4	4	4	4	4	4
BN001	5	5	4	4	4	5	4	5	5	5	5	3	4	4
BN001	5	5	4	4	5	4	5	4	5	4	4	4	5	5
BN001	5	5	3	4	4	4	3	5	5	4	3	3	4	3
BN001	3	3	3	2	3	2	3	3	3	2	4	3	3	3
BN001	4	4	3	4	5	2	4	4	3	2	3	2	3	2
BN001	3	3	2	3	3	3	3	4	3	3	3	4	3	3
BN001	4	5	5	5	5	5	5	4	4	4	5	5	5	5

Briefing Note identifier

Sum of scores under alpha dimensions = 539
Total evaluations = 135

Sum of scores under beta dimensions = 276
Total evaluations = 75

Figure 7 Briefing note evaluation score computation.

Nexus Community Institute for a $55,000 grant.' This introductory sentence is ambiguous as it does not clearly convey the intended purpose of the briefing note as expected (Fornberg 2020). The note does not allow its receiver to understand the author's indented purpose upon reading this summary, especially given that a decision is sought as the outcome of the briefing note.

The note produced by our fine-tuned version of gpt-4o-mini (BN006, see Box 3) produced an excellent introductory sentence that captures the essence of the note: '*This briefing note seeks the Deputy Minister's decision regarding a proposed grant of $55,000 to Nexus Community Institute*.' Thus, BN006's intent is clearer, and it tells the decision-maker, right in its summary section, that the decision is sought to approve the proposed grant. This note was rated higher than the other notes across multiple indicators as shown in Table 8, confirming our assumption that fine-tuned LLMs can achieve better results than general-purpose models.

Box 3 Note Produced by a Fine-tuned gpt-4o-mini.

Briefing Note
For Decision
NOTE ID: BN006
March 18, 2025
MEMO TO: The Deputy Minister
SUBJECT: Approval of a $55 000 Grant to the Nexus Community Institute

1. SUMMARY

This briefing note seeks the Deputy Minister's decision regarding a proposed grant of $55 000 to Nexus Community Institute. This funding aims to support the continuation of their community development initiatives in southern Ottawa, which have significantly contributed to local social and economic resilience.

2. BACKGROUND

Nexus Community Institute has successfully operated various programmes for the past 8 years, focusing on enhancing community capacity in rural and smaller communities. Their initiatives particularly benefit vulnerable populations, including women and families, by providing resources that promote self-development and connectivity. The Institute's activities include developing funding programmes, offering information resources, and conducting research to bolster community economic development.

3. CONSIDERATIONS

Grant approval is essential for maintaining uninterrupted delivery of services that address socio-economic disparities. Nexus's proven track record and community-focused approach make this funding a strategic investment in community resilience and empowerment.

4. CONCLUSIONS

The proposed grant aligns with federal policy objectives aimed at supporting rural development and enhancing community cohesion.

5. RECOMMENDATIONS

The Deputy Minister is advised to approve the $55 000 grant to ensure continued support for community development in southern Ottawa.

Briefing note BN006 was favoured by evaluators under the decision-informing capacities where it obtained higher scores than most other briefing notes. In the PC framework discussed previously, the decision-making power of policy is a key analytical capacity where accurately informing decision-makers is key. Thus, when analyzed under this framework, many LLMs' lower scores under their decision-making criteria suggest limitations in the analytical capacities to properly inform policy. This could, in return, hinder their effectiveness in providing comprehensive policy insights (Wu *et al.* 2015). The intricacies, 'acumen' and tasks with 'great demands' (Pal and Clark 2015) are known in the policy process where decision-seeking briefing notes are expected to be thoroughly informative, clear and 'more precise' in proposing policy choices (Wu *et al.* 2015).

The briefing notes produced by Mistral-Large (BN003), and Llama 4 (BN005) were given intermediate performance scores with indices of 3.53 and 3.64, respectively, which was comparable to the index obtained by the human-written benchmark note (BN007, index = 3.63) but below the fine-tuned model's note (BN006, index = 4.54) and OpenAI's ChatGPT (BN001, index = 3.84). In Table 8, when comparing the two OpenAI's models, i.e. our fine-tuned model and ChatGPT, we find that OpenAI's GPT-4o's

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

standard model (BN001) exhibits a strong but comparatively low performance with an index of 3.84. When weighted over the number of respondents, this model's scores fall short relative to its fine-tuned counterpart (BN006) which is a key indicator of the value-added from fine-tuning models for specialized tasks. The brevity and on-point capacities demonstrated by the fine-tuned model suggest there could be room for improvement in policy competencies such as clarity, structure, and practical applicability for teams or organizations seeking expert and task-specialized models. Nevertheless, from a policy capacity lens, the higher analytical capabilities demonstrated by briefing notes from the two GPT-4o models (both the fine-tuned (BN006) and non-fine-tuned (BN001) show that these models could be key tools to support policy and decision-making. The capacity to translate data and knowledge into concise and relevant policy insights means that the models' analytical capacities could be headed towards achieving what is expected from policy professionals at the expert level.

When looked at through the PCF, another key pattern in our results is that GenAI models generally perform strongly on analytical communication (Alpha indicators) but less consistently on analytical substance (Beta indicators). In other words, the models often generate briefing notes that are readable, structured, and clear, yet may lack the evidentiary depth and contextual completeness required for high-stakes decision support. This distinction is important in PCF terms because analytical capacity involves not only drafting competence but also the ability to ground advice in adequate research, maintain coherence with objectives, and provide decision-relevant reasoning.

This distinction is visible in our aggregated results: AI-generated notes obtained higher average scores than the human-written benchmark on Alpha indicators (e.g. readability, conciseness, tone), whilst the human benchmark outperformed AI-generated notes on Beta indicators, most notably research thoroughness and consistency with objectives. Through a PCF lens, this implies that current GenAI strengths may support some elements of individual analytical capacity (rapid drafting, clarity, structure), but limitations remain where analytical capacity is most demanding (evidence integration, contextual completeness, and decision justification).

Our model-comparison results (GenAIBNIndex) show that the fine-tuned model achieved the highest overall index score relative to other foundation models and the human benchmark. Under a PCF lens, this performance suggests that fine-tuning can improve the extent to which an LLM supports individual-level analytical capacity, particularly on the briefing note functions that matter for senior decision-making: stating the decision ask unambiguously, structuring key considerations and risks succinctly, and maintaining coherence across sections.

6. Conclusion

We designed an experiment that sought to test various GenAI models' policy briefing capabilities with a goal of understanding the promises of fine-tuning a base LLM for policy-specific tasks. Our results have suggested that when put to an evaluator's test, it may be difficult to draw a line between human-written policy briefs and those generated by AI. Indeed, some of the AI-generated notes outperformed the human-written notes on multiple indicators. On the fine-tuning promises, our results aligned with previous findings that fine-tuning a base LLM can improve performance in this specialized context of policy briefs. Our fine-tuned model demonstrated notable capabilities at producing concise, informative, and short notes that clearly stated their intended purposes without overwhelming the receiver based on mean scores obtained after the evaluation process. Analyzed under the PC theoretical framework that assesses the ability to devise effective and informed policies, the tested GenAI models are promising, and their analytical capacities should not be undermined.

Whilst it may still be too early to praise LLMs' humanlike capabilities as evaluated by known techniques such as the Turing test (Gams & Kramar 2024; Restrepo Echavarría 2025; and Giunti 2026), the technology's potential in specific policy drafting and synthesis tasks should not be dismissed outright. Rather, future research should focus on understanding the promises of model customization when navigating known AI limitations such as bias and when task complexities and uncertainties are higher (Nzobonimpa 2023). It is worth noting that whilst this study unearthed the promise of fine-tuning a base LLM, we encountered a limitation with the relatively small sample of evaluators which limited causal and inferential claims. Thus, future research should also consider evaluating these briefing notes with larger and more diverse groups of respondents to allow formal inferential comparisons. Nevertheless, our sample of 60 evaluators provided a good benchmark and an important empirical reference for AI-generated briefing note evaluation.

Author contributions

Stany Nzobonimpa (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Jean-François Savard (Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision), and Justin Lawarée (Conceptualization, Investigation, Methodology, Project administration, Resources, Validation)

Conflict of interest

None declared.

Funding

This article was funded in part by Social Sciences and Humanities Research Council of Canada, 186858 (Stany Nzobonimpa).

References

Anne, T. (2015) 'The Dilemmas of Organisational Capacity', *Policy and Society*, 34: 209–17. https://doi.org/10.1016/j.polsoc.2015.09.004.

Aoki, G. (2024) 'Large Language Models in Politics and Democracy: a Comprehensive Survey', *arXiv preprint*, arXiv:2412.04498. https://doi.org/10.48550/ARXIV.2412.04498.

Baracskay, D. (2024) 'Does Gen-AI Have a Role in Public Affairs Education? Let's Ask ChatGPT', *Teaching Public Administration*, 44: 15–48. https://doi.org/10.1177/01447394241279361.

Benoit, K. (2024) 'AI and Data Science for Public Policy', *LSE Public Policy Review*, 3: 1. https://doi.org/10.31389/lseppr.115.

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

Boden M. A (2004). *The creative mind: myths and mechanisms*., 2nd ed. London; New York: Routledge, https://doi.org/10.4324/9780203508527.

Bousselham H., Nfaoui E. H., & Mourhir A. (2024). 'Fine-tuning GPT on biomedical NLP tasks: An empirical evaluation'. *2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE), pp. 1–6. Presented at the 2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE), February 2*, Kolkata: IEEE. https://doi.org/10.1109/ICCECE58645.2024.10497313.

Canada School of Public Service (2024) 'Working with Artificial Intelligence Series: Co-Coding with AI (DDN2-A37)', Retrieved from <https://www.csps-efpc.gc.ca/tools/articles/co-coding-with-ai-eng.aspx>.

Cantens, T. (2025) 'How Will the State Think with ChatGPT? The Challenges of Generative Artificial Intelligence for Public Administrations', *AI & Society*, 40: 133–44. https://doi.org/10.1007/s00146-023-01840-9.

Chiu, T. K. F. (2024) 'The Impact of Generative AI (GenAI) on Practices, Policies and Research Direction in Education: a Case of ChatGPT and Midjourney', *Interactive Learning Environments*, 32: 6187–203. https://doi.org/10.1080/10494820.2023.2253861.

Chun, S. A., and Noveck, B. S. (2025) 'Introduction to the Special Issue on ChatGPT and Other Generative AI Commentaries: AI Augmentation in Government 4.0: AI Augmentation in Government 4.0', *Digital Government: Research and Practice*, 6: 1–8. https://doi.org/10.1145/3716859.

Cropley, D. (2023) 'Is Artificial Intelligence more Creative than Humans? ChatGPT and the Divergent Association Task', *Learning Letters*, 2: 13. https://doi.org/10.59453/ll.v2.13.

Fischer, A. R. H. et al. (2014) 'Expert Involvement in Policy Development: a Systematic Review of Current Practice', *Science and Public Policy*, 41: 332–43. https://doi.org/10.1093/scipol/sct062.

Fonberg, R. (2020) *Practical Guide To Writing Briefing Notes In The Government Of Canada*, Policy Brief, Defence & Security Foresight Group. Retrieved from https://www.naadsn.ca/wp-content/uploads/2020/08/dsfg_policy_brief_writing_guide_v2.pdf.

Gams, M., and Kramar, S. (2024) 'Evaluating ChatGPT's Consciousness and its Capability to Pass the Turing Test: a Comprehensive Analysis', *Journal of Computer and Communications*, 12: 219–37. https://doi.org/10.4236/jcc.2024.123014.

Gao, A. (2023) 'Implications of ChatGPT and Large Language Models for Environmental Policymaking', *SSRN Electronic Journal*, 4499643. https://doi.org/10.2139/ssrn.4499643.

Ghiurău, D., and Popescu, D. E. (2024) 'Distinguishing Reality from AI: Approaches for Detecting Synthetic Content', *Computers*, 14: 1. https://doi.org/10.3390/computers14010001.

Giray, L. (2024) 'ChatGPT References Unveiled: Distinguishing the Reliable from the Fake', *Internet Reference Services Quarterly*, 28: 9–18. https://doi.org/10.1080/10875301.2023.2265369.

Giunti, M. (2026) 'ChatGPT-4 in the Turing Test: a Critical Analysis', *Minds & Machines*, 36: 8. https://doi.org/10.1007/s11023-026-09760-5.

Gough, D., and Elbourne, D. (2002) 'Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate', *Social Policy and Society*, 1: 225–36. https://doi.org/10.1017/S147474640200307X.

Government of Canada. (2024). '*CANChat, SSC's first generative AI chatbot*'. Government of Canada. Retrieved from https://www.canada.ca/en/shared-services/campaigns/stories/canchat-sscs-first-generative-ai-chatbot.html.

Graham, A. (2021) 'Briefing note to obtain a decision', in *GovTak Queens University*. Kingston, ON: Queen's University School of Policy Studies. https://www.queensu.ca/sps/govtalk.

Gunes, E., and Florczak, C. K. (2023) 'Multiclass Classification of Policy Documents with Large Language Models'. *arXiv preprint: 2310.08167*. http://dx.doi.org/10.48550/ARXIV.2310.08167.

Hasan, Md, A. et al. (2023) 'Zero- and Few-Shot Prompting with LLMs: a Comparative Study with Fine-Tuned Models for Bangla Sentiment Analysis'. *arXiv e-prints:* arXiv-2308. https://doi.org/10.48550/ARXIV.2308.10783.

C. Hou, G. Zhu, J. Zheng, L. Zhang, X. Huang, T. Zhong, S. Li, et al. (2024). 'Prompt-based and fine-tuned GPT models for context-dependent and -independent deductive coding in social annotation'. *Proceedings of the 14th Learning Analytics and Knowledge Conference, pp. 518–28. Presented at the LAK 24: The 14th Learning Analytics and Knowledge Conference, March 18, Kyoto Japan*. New York: Association for Computing Machinery. https://doi.org/10.1145/3636555.3636910.

Howlett, M. (2015) 'Policy Analytical Capacity: the Supply and Demand for Policy Analysis in Government', *Policy and Society*, 34: 173–82. https://doi.org/10.1016/j.polsoc.2015.09.002.

Jackson, J. M., and Pinto, M. D. (2024) 'Human near the Loop: Implications for Artificial Intelligence in Healthcare', *Clinical Nursing Research*, 33: 135–7. https://doi.org/10.1177/10547738241227699.

Janssen, M. (2025) 'Responsible Governance of Generative AI: Conceptualizing GenAI as Complex Adaptive Systems', *Policy and Society*, 44: 38–51. https://doi.org/10.1093/polsoc/puae040.

Lermann Henestrosa, A., and Kimmerle, J. (2025) 'Always check important information!- The role of disclaimers in the perception of AI-generated content' *Computers in Human Behavior: Artificial Humans*, 4: 100142. https://doi.org/10.1016/j.chbah.2025.100142.

Lucas, J. S. et al. (2024) 'The Longtail Impact of Generative AI on Disinformation: Harmonizing Dichotomous Perspectives', *IEEE Intelligent Systems*, 39: 12–9. https://doi.org/10.1109/MIS.2024.3439109.

Magallanes Reyes, J. M. (2022). *Data Visualization for Social and Policy Research: A Step-by-Step Approach Using R and Python*., 1st ed. Cambridge University Press. DOI: https://doi.org/10.1017/9781108625425

Masood, S., Kothari, A., and Regan, S. (2020) 'The Use of Research in Public Health Policy: a Systematic Review', *Evidence & Policy*, 16: 7–43. https://doi.org/10.1332/174426418X15193814624487.

McKay, S., Murray, M., and Macintyre, S. (2012) 'Justice as Fairness in Planning Policy-Making', *International Planning Studies*, 17: 147–62. https://doi.org/10.1080/13563475.2012.672798.

Moruzzi, C. (2025) 'Artificial Intelligence and Creativity', *Philosophy Compass*, 20: e70030. https://doi.org/10.1111/phc3.70030.

Mouzannar H., M. I. Ohannessian, & N. Srebro (2019). 'From fair decision making to social equality'. *Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 359–68. Presented at the FAT∗ 19: Conference on Fairness, Accountability, and Transparency, January 29, Atlanta GA USA: ACM*. New

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287599.

Newman, J., Cherney, A., and Head, B. W. (2017) 'Policy Capacity and Evidence-Based Policy in the Public Service', *Public Management Review*, 19: 157–74. https://doi.org/10.1080/14719037.2016.1148191.

Nyqvist, R., Peltokorpi, A., and Seppänen, O. (2024) 'Can ChatGPT Exceed Humans in Construction Project Risk Management?' *Engineering, Construction and Architectural Management*, 31: 223–43. https://doi.org/10.1108/ECAM-08-2023-0819.

Nzobonimpa, S. (2023) 'Artificial Intelligence, Task Complexity and Uncertainty: Analyzing the Advantages and Disadvantages of Using Algorithms in Public Service Delivery under Public Administration Theories', *Digital Transformation and Society*, 2: 219–34. https://doi.org/10.1108/DTS-03-2023-0018.

Nzobonimpa, S. (2026) 'New Tech, Old Demons: Generative Artificial Intelligence Could Inherit Flaws of Machine Learning', *Journal of Responsible Technology*, 25. https://doi.org/10.1016/j.jrt.2026.100160.

Nzobonimpa, S. et al. (2024) 'Automating Public Policy: a Comparative Study of Conversational Artificial Intelligence Models and Human Expertise in Crafting Briefing Notes', *AI & SOCIETY.*, 40: 3627–39. https://doi.org/10.1007/s00146-024-02103-x.

OpenAI (2024) 'API Reference', in *OpenAI Platform*. OpenAI Developers, OpenAI.

OpenAI (2025) 'Fine-Tuning', OpenAI Platform. Retrieved from <https://platform.openai.com/docs/guides/fine-tuning>.

Ovsyannikova, D., De Mello, V. O., and Inzlicht, M. (2025) 'Third-Party Evaluators Perceive AI as More Compassionate than Expert Humans', *Communications Psychology*, 3: 4. https://doi.org/10.1038/s44271-024-00182-6.

Pal, L. A., and Clark, I. D. (2015) 'Making Reform Stick: Political Acumen as an Element of Political Capacity for Policy Change and Innovation', *Policy and Society*, 34: 247–57. https://doi.org/10.1016/j.polsoc.2015.09.006.

Pattyn, V., and Brans, M. (2015) 'Organisational Analytical Capacity: Policy Evaluation in Belgium', *Policy and Society*, 34: 183–96. https://doi.org/10.1016/j.polsoc.2015.09.009.

Peters, B. G. (2015) 'Policy Capacity in Public Administration', *Policy and Society*, 34: 219–28. https://doi.org/10.1016/j.polsoc.2015.09.005.

Pulapaka. S, Godavarthi. S, & Dr Ding. S. (2024a). 'GenAI and the public sector'. In: Pulapaka, S., Godavarthi, S. & Ding, D.S. (eds) *Empowering the Public Sector with Generative AI*, pp. 31–43. Apress: Berkeley. https://doi.org/10.1007/979-8-8688-0473-1_2.

Pulapaka. S, Godavarthi. S, & Ding. S (2024b). *Empowering the Public Sector with Generative AI: From Strategy and Design to Real-World Applications*. Berkeley: Apress. https://doi.org/10.1007/979-8-8688-0473-1.

Qi, Y., Peng, W., and Xiong, N. N. (2020) 'The Effects of Fiscal and Tax Incentives on Regional Innovation Capability: Text Extraction Based on Python', *Mathematics*, 8: 1193. https://doi.org/10.3390/math8071193.

Ramesh, M., Howlett, M. P., and Wu, X. (2016) 'Rethinking Governance Capacity As Organizational and Systemic Resources', *SSRN Electronic Journal*. Lee Kuan Yew School of Public Policy Research Paper No. 16-12. https://doi.org/10.2139/ssrn.2802438.

Restrepo Echavarría, R. (2025) 'ChatGPT-4 in the Turing Test', *Minds and Machines*, 35: 8. https://doi.org/10.1007/s11023-025-09711-6.

Safaei, M., and Longo, J. (2024) 'The End of the Policy Analyst? Testing the Capability of Artificial Intelligence to Generate Plausible, Persuasive, and Useful Policy Analysis', *Digital Government: Research and Practice*, 5: 1–35. https://doi.org/10.1145/3604570.

Saqib, M. B., and Zia, S. (2024) 'Evaluation of AI Content Generation Tools for Verification of Academic Integrity in Higher Education', *Journal of Applied Research in Higher Education.*, 17: 1430–40. https://doi.org/10.1108/JARHE-10-2023-0470.

Steif, K. (2021). *Public Policy Analytics: Code and Context for Data Science in Government*., 1st ed. Boca Raton: CRC Press. https://doi.org/10.1201/9781003054658.

Veatch, R. M. (1991) 'Consensus of Expertise: the Role of Consensus of Experts in Formulating Public Policy and Estimating Facts', *Journal of Medicine and Philosophy*, 16: 427–45. https://doi.org/10.1093/jmp/16.4.427.

Villa, L. et al. (2024) 'Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems', *Robotics*, 13: 68. https://doi.org/10.3390/robotics13050068.

Wang, Q., and Gayed, J. M. (2024) 'Effectiveness of Large Language Models in Automated Evaluation of Argumentative Essays: Fine-Tuning Vs. Zero-Shot Prompting', *Computer Assisted Language Learning*, 39: 209–37. https://doi.org/10.1080/09588221.2024.2371395.

Wang, X. et al. (2025) 'Wireless Hallucination in Generative AI-Enabled Communications: Concepts, Issues, and Solutions'. *arXiv preprint*: arXiv:2503.06149.', 10.48550/ARXIV.2503.06149.

Welivita, A., and Pu, P. (2024) 'Is ChatGPT more Empathetic than Humans?' *arXiv preprint:* arXiv:2403.05572. https://doi.org/10.48550/ARXIV.2403.05572.

Wu, X., Ramesh, M., and Howlett, M. (2015) 'Policy Capacity: a Conceptual Framework for Understanding Policy Competences and Capabilities', *Policy and Society*, 34: 165–71. https://doi.org/10.1016/j.polsoc.2015.09.001.

Yan, L. et al. (2024) 'Promises and Challenges of Generative Artificial Intelligence for Human Learning', *Nature Human Behaviour*, 8: 1839–50. https://doi.org/10.1038/s41562-024-02004-5.

Yehezkel, A. (2024). 'An efficient approach for enhancing GenAI trustworthiness'. In: Wani, M.M., Arif, M. & Institute of Electrical and Electronics Engineers (eds) *2024 International Conference on Machine Learning and Applications (ICMLA), pp. 1900–5. Presented at the 2024 International Conference on Machine Learning and Applications (ICMLA), December 18*. Miami, FL, USA: IEEE. https://doi.org/10.1109/ICMLA61862.2024.00292.

Zulfikar, W. B., Atmadja, A. R., and Pratama, S. F. (2023) 'Sentiment Analysis on Social Media against Public Policy Using Multinomial Naive Bayes', *Scientific Journal of Informatics*, 10: 25–34. https://doi.org/10.15294/sji.v10i1.39952.

Downloaded from https://academic.oup.com/spp/advance-article/doi/10.1093/scipol/scag026/8662016 by guest on 28 April 2026

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Science and Public Policy, 2026, 00, 1–17
DOI: https://doi.org/10.1093/scipol/scag026
Article