


---

# Beyond Stakeholder Simulation: AI as Document Exploration Tool in Migration Policy Evaluation

Polski Przegląd Ewaluacyjny  
Nr 1(4)/2025  
©Autorzy 2025  
  
ISSN 2956-5332  
pte.org.pl/o-czasopismie

**Igor Lyubashenko**

SWPS University, Poland; ORCID: 0000-0003-0404-5460

**Paweł Kędzia**

RadLab; ORCID: 0000-0003-0544-4586

## Abstract

Evaluating policy interventions for displaced populations requires authentic stakeholder perspectives, yet traditional methods prove inadequate during humanitarian crises. This paper presents a meta-evaluation of an AI-powered tool utilizing Retrieval-Augmented Generation technology to process 521,089 Telegram messages from Ukrainian and Russian-speaking populations. Our central finding challenges initial design assumptions: while intended to simulate stakeholder perspectives, the tool proved most valuable as a document exploration platform for qualitative data analysis. This shift from simulation to exploration represents a significant methodological insight for evaluation practice. The technical architecture successfully implemented multi-stage filtering, hierarchical clustering into 249 thematic groups, and transparent retrieval mechanisms. We argue that AI technologies offer greatest promise not in replacing stakeholder engagement, but in enhancing evaluators' capacity to systematically process qualitative data. This research contributes to debates on responsible AI integration in evaluation methodology.

## Keywords

evaluation methodology, artificial intelligence, stakeholder engagement, retrieval-augmented generation, migration policy, meta-evaluation

## Poza symulacją interesariuszy: AI jako narzędzie eksploracji dokumentów w ewaluacji polityki migracyjnej

### Abstrakt

Ewaluacja interwencji politycznych wobec ludności przesiedlonej wymaga uwzględnienia autentycznych perspektyw interesariuszy, lecz tradycyjne metody okazują się niewystarczające podczas kryzysów humanitarnych. Artykuł prezentuje metaewaluację narzędzia AI wykorzystującego technologię Retrieval-Augmented Generation do przetwarzania 521 089 wiadomości z Telegramu od populacji ukraińsko- i rosyjskojęzycznych. Kluczowe odkrycie podważa wstępne założenia projektowe: narzędzie zaprojektowane do symulowania perspektyw interesariuszy okazało się najbardziej wartościowe jako platforma eksploracji dokumentów do analizy danych jakościowych. Ta zmiana - od symulacji do eksploracji - stanowi istotny wgląd metodologiczny dla praktyki ewaluacyjnej. Architektura techniczna skutecznie zaimplementowała wieloetapowe filtrowanie, hierarchiczne grupowanie w 249 grup tematycznych oraz przejrzyste mechanizmy wyszukiwania. Argumentujemy, że technologie AI oferują największy potencjał nie w zastępowaniu zaangażowania interesariuszy, lecz we wzmacnianiu zdolności ewaluatorów do systematycznego przetwarzania danych jakościowych. Badanie przyczynia się do debaty nad odpowiedzialną integracją AI w metodologii ewaluacji.

### Słowa kluczowe

metodologia ewaluacji, sztuczna inteligencja, zaangażowanie interesariuszy, generowanie wspomagane wyszukiwaniem, polityka migracyjna, metaewaluacja

---

### Corresponding author(s):

Igor Lyubashenko, SWPS University, ul. Chodakowska 19/31, 03-815 Warszawa, Email: ilyubashenko@swps.edu.pl

## **Introduction**

Contemporary evaluation practice faces unprecedented methodological challenges when assessing policies and programs targeting marginalized populations such as migrants and refugees. Traditional evaluation methods - surveys, focus groups, and stakeholder consultations - often prove inadequate for capturing authentic participant perspectives due to language barriers, geographical dispersion, legal concerns, trauma, and institutional distrust. These limitations become particularly acute during humanitarian crises when evaluators must assess program effectiveness while direct engagement with beneficiaries becomes constrained or entirely inaccessible.

The ongoing conflict in Ukraine created one of the largest refugee movements in recent European history, necessitating rapid deployment of support programs whose effectiveness requires systematic evaluation. Yet conventional evaluation approaches struggle to capture program impacts from beneficiary perspectives when populations remain displaced, vulnerable, and difficult to reach. Evaluators require timely, contextually relevant insights to assess program quality and inform adaptive management, but traditional evaluation cycles often cannot match the pace at which programs evolve during humanitarian responses.

This methodological challenge has encouraged exploration of innovative evaluation approaches that can bridge the gap between evaluators' need for stakeholder input and practical constraints of engaging with displaced populations. The emergence of sophisticated artificial intelligence technologies, particularly Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems, presents novel opportunities for enhancing evaluation methodologies when direct engagement is compromised.

Methodological challenges of evaluation in crisis contexts are receiving growing attention in evaluation literature. Particularly important is the question of including stakeholder voices in evaluation processes when traditional engagement methods are constrained or impossible to apply. Contemporary literature emphasizes the tension between participation requirements and practical constraints of working in complex and dynamic environments (Cousins and Alborhamy, 2025). At the same time, advancing digitalization is fundamentally transforming the evaluation landscape, offering both unprecedented opportunities and significant challenges for the evaluation community (Potluka et al., 2025).

Digital communication patterns of displaced populations offer an alternative data source for evaluation purposes. Social media platforms and messaging applications generate vast repositories of authentic discourse reflecting program experiences, unmet needs, and policy impacts that traditional evaluation methods might miss or capture only retrospectively. These organic expressions provide evaluators with real-time perspectives that can inform both formative and summative assessment.

This paper presents the development and empirical test of an AI-powered tool designed to address methodological gaps in migration policy evaluation. The tool leverages publicly available communication data from Telegram channels used by Ukrainian and Russian-speaking populations, applying advanced natural language processing to create interfaces capable of supporting evaluation activities through enhanced access to stakeholder discourse.

The rationale for developing this evaluation tool rested on several initial premises: that digital communications contain authentic expressions of program experiences informing evaluation when traditional methods are not viable; that AI technologies can be responsibly employed to support evaluation practice while maintaining methodological rigor and ethical standards; and that evaluators require accessible tools providing contextually relevant insights without compromising vulnerable populations during crisis situations. However, our central finding challenges one key assumption - while the tool was designed to simulate stakeholder perspectives, empirical testing revealed its primary value lies in document exploration and qualitative data analysis rather than perspective simulation. This discovery represents our core contribution to methodological discourse on AI in evaluation.

This work presents the tool's technical architecture and discusses its practical utility for evaluation purposes through the lens of meta-evaluation. Rather than replacing participatory evaluation approaches, we now conceptualize this tool as a methodological complement best suited for processing

existing qualitative data, while maintaining transparency about limitations and the continuing need for direct stakeholder engagement.

The paper contributes to emerging discussions about methodological innovation in evaluation practice. By examining both the potential and limitations of this tool through empirical testing, we advance understanding of how AI technologies might ethically support evaluation research while highlighting critical considerations for maintaining evaluation standards and beneficiary dignity.

## Theoretical Background

Authentic stakeholder engagement stands as a foundational principle in evaluation practice, yet evaluators consistently struggle to capture genuine perspectives from marginalized populations during crisis situations. This challenge has prompted scholars to explore whether artificial intelligence technologies might serve as methodological supports when direct engagement becomes constrained. The question is not whether AI can replace human voices in evaluation - it cannot - but whether these technologies can responsibly augment evaluation approaches when conventional methods prove inadequate.

### *Methodological Innovations in Evaluation Practice*

The presented research aligns with developmental evaluation, an approach formulated by Patton (2010) for contexts characterized by uncertainty, complexity, and continuous adaptation. Developmental evaluation supports innovation development by providing real-time feedback, enabling adaptation to emergent and dynamic realities in complex environments. Unlike traditional formative and summative approaches, developmental evaluation recognizes that in innovative contexts, goals and methods may evolve as interventions develop.

In parallel, evaluation literature documents growing interest in using artificial intelligence and big data in evaluation practice. The systematic review by Bouyousfi and Ouedraogo (2024) identifies three main application areas of AI in evaluation: conceptualization, implementation, and results analysis. However, the authors emphasize unresolved ethical, methodological, and data ownership issues that require attention when integrating new technologies with traditional evaluation approaches.

Potluka and colleagues (2025) point to three interconnected challenges in evaluation digitalization: methodological knowledge gaps, data and technology ownership issues, and ethical considerations. The authors argue that despite the transformation of the evaluation landscape, there are stable principles and standards that evaluators should follow. Particularly important is maintaining algorithmic transparency and ensuring human oversight of automated data processing.

In the context of humanitarian evaluation, real-time evaluation approaches are particularly significant, enabling feedback delivery during ongoing interventions (ALNAP, 2009). However, this approach requires different methods than traditional evaluation due to limited time available for forming evaluative judgments. The presented AI tool can be viewed as an attempt to extend real-time evaluation capabilities through automation of qualitative data analysis.

### *AI and Perspective Simulation*

Recent research demonstrates growing potential of LLMs to simulate human perspectives and opinions in policy-relevant contexts. Zhang et al. (2024) developed a "Focus Agent" framework that uses LLMs to simulate focus group discussions, generating opinions comparable to human participants, though with noted limitations in capturing individual variance. Park et al. (2024) created generative agents informed by qualitative interview data that achieved high accuracy in replicating individual attitudes on standardized surveys. The concept of "Synthetic Interlocutors" introduced by Soltoft et al. (2024) extends this approach through RAG, creating chatbots that can re-contextualize ethnographic data and generate new analytical insights. These developments suggest that AI systems, when properly designed, might help evaluators access stakeholder perspectives embedded in existing data sources.

Yet these technological possibilities encounter particular complications when applied to migrant and refugee populations. Traditional stakeholder engagement methods already face substantial obstacles in

these contexts. Allen and Slotterback (2017) documented how conventional public engagement approaches failed with Somali refugees in the Twin Cities, highlighting the need for culturally sensitive design that builds trust through community relationships and partnerships with community-led organizations. The digital divide compounds these challenges for AI-based approaches. While platforms like Telegram offer access to community discourse, they inevitably exclude the most vulnerable populations - those lacking digital literacy, device access, or internet connectivity. This creates a representational bias toward more politically engaged, technologically connected individuals within migrant communities (Nikolopoulou et al., 2023). Any evaluation tool relying on digital communications must acknowledge these systematic exclusions.

### *Ethical Considerations in AI Deployment*

The ethical stakes of AI deployment in migration contexts demand careful attention. The literature consistently emphasizes that AI systems are not neutral but reflect biases embedded in their training data and design choices. Hall and Clapton (2021) argue that AI systems used in border control are inherently gendered, racialized, and sexualized, potentially undermining the rights of vulnerable groups. Bircan and Korkmaz (2021) criticize existing AI applications in migration governance for their lack of transparency and migrant involvement in system design, often treating migrants as experimental subjects without adequate consent. These concerns underscore the importance of participatory design approaches that involve affected communities in technology development and deployment. Aoki (2024) stresses the necessity for responsible development, ethical oversight, and governance frameworks to ensure AI integration in politics aligns with democratic values. The power asymmetries inherent in migration contexts make ethical safeguards not optional enhancements but fundamental requirements.

Given these ethical imperatives, methodological innovations must prioritize transparency and validation. The integration of RAG techniques appears crucial for grounding AI responses in authentic source material and reducing hallucination risks. However, technological sophistication evident in recent AI research reveals important limitations alongside capabilities. Studies of LLM-based perspective simulation reveal tendencies toward more common opinions, challenges in maintaining natural dialogue, and risks of inaccurate knowledge attribution (Soltoft et al., 2024; Zhang et al., 2024). Validation of AI-generated perspectives presents ongoing challenges, particularly when working with hard-to-reach populations where traditional validation methods may be inaccessible. The literature suggests that AI tools are most appropriately positioned as complements to rather than replacements for direct stakeholder engagement. McIntosh et al. (2023) demonstrate how AI can support policy development through game-theoretic validation, while maintaining human oversight and decision-making authority.

Current research reveals several critical gaps that this work addresses. Limited empirical studies evaluate AI tool effectiveness in real-world migrant engagement contexts, leaving uncertainty about practical utility beyond controlled research settings. Insufficient development of bias mitigation techniques for vulnerable populations means that ethical concerns remain largely theoretical rather than operationalized in working systems. Inadequate frameworks for validating AI-generated perspectives against authentic community voices create methodological ambiguity about assessing these tools' reliability and appropriateness. The literature converges on the need for more robust methodologies that integrate AI capabilities with traditional research methods, develop participatory design approaches involving affected communities, and create comprehensive ethical frameworks for responsible AI deployment in policy contexts.

## **Architecture of the AI-Powered Tool**

The discussed tool represents an integration of natural language processing, machine learning, and Retrieval-Augmented Generation technologies designed initially to simulate stakeholder perspectives in migration policy contexts. This section outlines the technical architecture with emphasis on methodological choices and their implications for evaluation practice.

### Data Foundation

The tool's knowledge base was constructed from publicly available communications data sourced from Telegram, a messaging platform widely adopted by Ukrainian and Russian-speaking populations during the ongoing conflict. The initial dataset comprised 521,089 messages collected from 12 manually curated

thematic channels, spanning from December 29, 2016, to August 22, 2024. This temporal scope captured communications before, during, and throughout the escalation of the Ukrainian conflict, providing a longitudinal perspective on evolving concerns and experiences.

The selection of Telegram as a data source was driven by its accessibility and widespread adoption among the target demographic. However, as subsequent testing revealed, the prevalence of war-related content in the dataset presented significant challenges for extracting policy-relevant insights beyond immediate conflict concerns - a limitation that proved fundamental to understanding the tool's actual utility.

### *Processing Pipeline and Quality Control*

The architecture implements a multi-stage data processing pipeline transforming raw communication data into a structured knowledge base. The process begins with HTML parsing and chronological organization of messages, followed by storage in a relational database maintaining essential metadata including authorship, timestamps, and channel origins.

Automated translation of all content into Polish used the radlab/pLLama3.1-8B-content generative model, selected for its content filtering capabilities that automatically flag inappropriate content with standardized markers rather than processing potentially harmful material.

Recognizing that raw communication data contains significant volumes of content directly related to military operations and combat activities, the architecture incorporates a filtering mechanism to identify policy-relevant information. The system defines "relevant content" as communications addressing political developments, emigrant experiences, humanitarian assistance, and general conflict impacts while excluding tactical military discussions and frontline operational details.

The filtering process employs a hybrid approach combining unsupervised machine learning with human oversight. The HDBSCAN clustering algorithm (Stewart and Al-Khassaweneh, 2022) segments the entire dataset into 670 thematically coherent clusters, each labelled through automated analysis of representative message samples. A generative model provides initial relevance classifications based on predefined criteria, followed by expert validation and correction to ensure accuracy. This expert-validated subset serves as training data for a dedicated classification model distinguishing relevant from non-relevant content across the full dataset. The application of this classifier reduced the working dataset from 521,089 to 287,127 messages, representing substantial improvement in content quality and policy relevance.

### *Semantic Organization*

The filtered dataset undergoes semantic analysis creating a hierarchical knowledge structure enabling precise information retrieval. Messages are converted into high-dimensional semantic vectors (embeddings) of 1,024 dimensions, capturing nuanced meaning and thematic relationships beyond simple keyword matching.

Dimensional reduction through t-SNE pre-processing (Melit Devassy et al., 2020) followed by HDBSCAN clustering creates 249 primary thematic clusters forming the backbone of the tool's knowledge organization. These clusters are arranged in a hierarchical structure derived from the clustering algorithm's natural tree organization, with each level receiving automated descriptive labels generated through LLM analysis. This hierarchical approach enables the tool to operate at multiple levels of specificity, from broad thematic areas to highly specific subtopics.

### *RAG Implementation*

The core functionality relies on a RAG architecture (Liu, 2025) combining semantic search capabilities with controlled text generation. Individual messages are segmented into manageable chunks of up to 400 tokens and indexed in both semantic (embedding-based) and relational (text and metadata) databases.

This dual indexing approach enables nuanced semantic searches while maintaining access to structured metadata for filtering and validation purposes. The search mechanism can be constrained to specific

thematic clusters or hierarchical levels, significantly reducing hallucination risk by ensuring generated responses are grounded in relevant source material.

The RAG implementation includes prompt engineering allowing users to define personas for the AI assistant, influencing response style and perspective while maintaining factual grounding in source data. This feature was intended to enable policy makers to explore different stakeholder viewpoints - though as our empirical evaluation reveals, this capability proved more limited than anticipated.

### *User Interface and Transparency Features*

The tool provides an interface allowing policy makers to engage in natural language conversations with the AI assistant while maintaining transparency about the system's capabilities and limitations. Users can specify particular topics of interest, define desired persona characteristics, and adjust the scope of knowledge base consultation.

Critically, the interface includes transparency features allowing users to examine source materials underlying any generated response, ensuring accountability and enabling verification of the AI assistant's reasoning. This design choice reflects the tool's positioning as a decision support system rather than an autonomous policy recommendation engine, maintaining human agency in the policy development process.

The architecture incorporates multiple layers of quality assurance and ethical safeguards designed to prevent misuse while maximizing utility for legitimate policy research purposes. Content filtering mechanisms operate at multiple stages, from initial data collection through final response generation. The system maintains detailed logging of all interactions and decision paths, enabling post-hoc analysis and continuous improvement.

## **Empirical Evaluation**

The empirical assessment of the tool aligns with the tradition of meta-evaluation - systematic examination of the quality of evaluation processes and products. The concept of meta-evaluation, introduced by Scriven (1969), assumes that evaluations themselves should be subject to assessment to ensure their reliability and usefulness. Stufflebeam (2001) developed this concept, arguing that meta-evaluations are essential for ensuring that evaluations provide sound findings and conclusions and that evaluation practice continues to improve. Contemporary literature emphasizes that meta-evaluation criteria must be adapted to context - standards used in one environment may require adaptation in another (Ayoo et al., 2023). For AI tools supporting evaluation, this requires developing new quality assessment frameworks that account for both traditional evaluation criteria (utility, feasibility, propriety, accuracy) and technology-specific challenges.

### *Pilot Study*

The tool underwent preliminary evaluation through a pilot study with three users: two public administration workers with experience in migration policy implementation, and one researcher specializing in qualitative methods. Participants were recruited through professional networks and selected based on their potential use of such tools in their work. Each participant received a standardized introduction to the tool's functionality and access for a testing period of approximately four weeks.

This pilot study aimed to identify initial usability patterns and technical issues rather than provide generalizable findings. The small sample size (N=3) and limited testing period constrain our ability to draw definitive conclusions; however, the qualitative feedback offers valuable exploratory insights into the tool's current capabilities and limitations that can inform future development and larger-scale evaluation.

### *Key Findings*

The pilot test revealed a fundamental mismatch between the tool's intended function and its actual performance. While designed to simulate migrant stakeholder perspectives for policy consultations, the tool appeared to function more effectively as a document exploration system. Users primarily engaged

with it to explore qualitative data and understand thematic patterns within the Telegram corpus, with engagement levels varying from single use to weekly interaction.

The most consistent feedback concerned technical performance. All participants rated the chatbot's response accuracy positively, suggesting the underlying RAG architecture functions effectively when generating responses. However, system performance emerged as a critical barrier, with response times affecting user experience.

The tool's limited ability to effectively simulate migrant perspectives became apparent during testing. This limitation appears to stem from the Telegram dataset's characteristics - heavily focused on crisis response and war-related information rather than the everyday experiences and personal narratives necessary for authentic stakeholder representation.

Public administration workers identified specific potential applications, particularly for quick access to information and recognizing key needs in analysed topics. They also noted that the possibility to run the tool on premise constitutes particular value from the perspective of data protection. They valued having a "Polish product with responsible persons here on site," suggesting that local development and support represent important considerations for public sector adoption.

The ability to customize response granularity (four levels of detail) and system prompts received positive feedback, with participants viewing these features as valuable for adapting the tool to specific analytical needs. Users expressed interest in expanding the tool's capabilities to include image modelling and data analytics from various sources, and emphasized the importance of regular knowledge base updates.

When asked whether they would recommend the tool to colleagues, responses ranged from "rather yes" to "hard to say," reflecting its transitional development stage. This mixed reception, combined with performance issues and the gap between intended and actual functionality, indicates that while the underlying approach shows promise for document analysis applications, significant refinement is needed before the tool can meaningfully support migration policy development through stakeholder perspective simulation.

### *Limitations*

Several significant limitations constrain interpretation of these findings and the tool's broader applicability.

The small sample size (N=3) and limited testing period preclude generalizable conclusions about the tool's utility. Participants were not randomly selected, and their professional contexts may not represent the full range of potential users. Future research should include larger, more diverse samples with extended testing periods.

The Telegram corpus, while extensive, proved heavily skewed toward crisis-related and war-focused content. This emphasis on immediate conflict concerns meant the dataset lacked the everyday experiences, personal narratives, and nuanced policy feedback necessary for comprehensive stakeholder perspective simulation. The temporal concentration of messages during acute crisis periods further limited the tool's ability to capture longer-term integration experiences.

Telegram users represent a self-selected population with digital access and literacy, systematically excluding the most vulnerable migrants - those lacking devices, connectivity, or digital skills. The perspectives captured thus over-represent more connected, possibly younger and more educated segments of the migrant population, while under-representing elderly individuals, those in rural areas, or those with limited resources.

All content was translated from Ukrainian and Russian to Polish using automated translation. While the translation model was selected for quality, automated translation inevitably introduces some meaning loss or distortion, potentially affecting the authenticity of captured perspectives.

The absence of direct comparison with traditional stakeholder engagement methods (such as focus groups or interviews with the same populations) means we cannot definitively assess how well the tool captures authentic perspectives. Future research should include such comparative validation.

### *Reconceptualizing Tool Function*

The findings highlight critical constraints in using social media platforms as primary data sources for stakeholder simulation. Telegram channels, while offering authentic community discourse, proved insufficient for capturing the breadth of migrant experiences necessary for effective policy support. This suggests that future implementations would benefit from access to more diverse data sources, including other social media platforms, though such access faces significant practical constraints following privacy-focused policy changes.

More promising appears to be the tool's potential application to large corpora of qualitative interviews with target populations. Traditional ethnographic and interview data, when processed through the same architectural framework, could provide the rich, experiential content that social media communications often lack. This approach would combine the depth and nuance of traditional qualitative research with the accessibility and analytical power of AI-assisted exploration.

Despite limitations in persona simulation, testing revealed that the tool functions effectively as a sophisticated document exploration and analysis system. The conversational interface enables researchers and policy makers to efficiently navigate large corpora of qualitative data, identifying patterns, themes, and specific perspectives that might be difficult to locate through traditional analytical methods.

This “chat with documents” functionality represents a significant practical contribution, allowing users to engage with complex datasets through natural language queries while maintaining transparency about source materials. The tool's ability to provide contextually relevant responses grounded in authentic source material proves valuable for rapid policy research and preliminary stakeholder perspective exploration.

The tool's flexible, data-driven architecture demonstrates potential for adaptation across various domains beyond migration policy. Testing suggests promising applications in contexts including public policy discourse analysis, health communication research, labour market trend identification, educational content development, and community forum analysis.

## **Discussion: Implications for Evaluation Practice**

The empirical findings suggest a reconceptualization of the tool's role in evaluation methodologies. Rather than serving primarily as a stakeholder simulation system, the tool appears most valuable as a sophisticated analytical instrument that can augment traditional qualitative research approaches. This positioning aligns better with ethical considerations around authentic representation while providing concrete practical benefits for evaluation practitioners.

### *From Simulation to Exploration*

The most significant finding from our meta-evaluation concerns the gap between intended and actual functionality. This gap is not merely a technical shortcoming to be overcome through further development, but rather reveals important insights about the appropriate positioning of AI tools in evaluation practice.

The initial design premise - that AI could simulate stakeholder perspectives based on social media discourse - proved overly ambitious given current technological capabilities and data constraints. However, this finding should not be interpreted as failure. Instead, it points toward a more realistic and potentially more valuable application: AI as a tool for enhanced exploration and analysis of existing qualitative data.

This reconceptualization has significant implications for how the evaluation community should approach AI integration. Rather than seeking to replace or simulate human perspectives, AI tools may offer greatest value in helping evaluators more efficiently access, organize, and analyse the qualitative data they already collect through traditional methods. The tool's demonstrated capacity to enable natural language queries across large document corpora, combined with its transparency features showing source materials, aligns well with established evaluation principles of data grounding and analytical rigor.

### *Technical Architecture Implications*

The successful implementation of the filtering, clustering, and retrieval mechanisms validates the chosen technical approach, even where content limitations affected final outputs. The multi-stage quality control pipeline, semantic organization system, and transparency features all performed as designed, suggesting that the architectural framework provides a solid foundation for future applications with more suitable datasets.

The RAG implementation successfully grounded responses in source materials and minimized hallucination risks, demonstrating that the technical approach can support reliable policy research applications when provided with appropriate data inputs.

### *Ethical Dimensions*

The findings reinforce several ethical considerations for AI deployment in evaluation contexts. First, the digital divide documented in our limitations section raises fundamental questions about representation and voice. Any AI tool drawing on digital communications will systematically under-represent certain populations. Evaluators must explicitly acknowledge these exclusions rather than treating AI-derived insights as representative of all stakeholders.

Second, the tool's greater success as document exploration rather than perspective simulation may actually represent an ethical advantage. Simulation of perspectives raises concerns about authenticity and potential misrepresentation - even well-intentioned simulation risks putting words in the mouths of vulnerable populations. Document exploration, by contrast, maintains clearer boundaries between original voices and analytical interpretation, preserving greater fidelity to authentic expression.

Third, the emphasis on transparency features - allowing users to trace any response back to source materials - reflects evaluation principles of accountability and verifiability. This design choice should be considered essential rather than optional for AI tools in evaluation contexts.

Fourth, the importance users placed on local development and on-premise deployment highlights data sovereignty concerns that evaluation practitioners should consider when selecting AI tools. For work with sensitive populations, the ability to maintain data within institutional or national boundaries may be crucial for ethical practice.

### *Future Development Directions*

Based on empirical findings, future development should prioritize expanding data source diversity and developing partnerships with research institutions that maintain extensive qualitative datasets. Integration with traditional interview corpora appears particularly promising, potentially creating hybrid methodologies that combine human-centred research depth with AI-assisted analytical capabilities.

Additional testing with diverse datasets across different policy domains would help establish best practices for data selection, quality assessment, and application-specific customization. The tool's demonstrated flexibility suggests potential for scaling across various research contexts while maintaining ethical standards and methodological rigor.

### *Implications for Democratic Participation*

The findings reinforce the importance of positioning AI tools as enhancements to rather than replacements for authentic stakeholder engagement. While the tool provides valuable analytical capabilities, the limitations observed in persona simulation underscore the irreplaceable value of direct community participation in policy development processes.

The most appropriate role for such technology appears to be supporting more effective utilization of existing qualitative data and enabling more efficient preliminary analysis that can inform the design of authentic engagement strategies. This approach preserves space for genuine participation while leveraging technological capabilities to enhance research efficiency and accessibility.

## Conclusion

This paper presents a meta-evaluation of an AI tool designed to address methodological challenges in migration policy evaluation. Meta-evaluation, understood as the evaluation of an evaluation or evaluation system (Scriven, 1969; Stufflebeam, 2001), enables systematic assessment of both technical architecture and practical utility of innovative solutions supporting evaluation practice. Through empirical assessment of the tool, this research contributes to evaluation methodology in several key areas.

First, our findings demonstrate that rigorous evaluation of AI tools reveals important gaps between intended functions and actual performance. While designed to simulate stakeholder perspectives for policy evaluation, empirical testing showed the tool functions most effectively as a sophisticated instrument for exploring and analysing qualitative evaluation data. This meta-evaluative insight has significant implications for evaluation practice: AI technologies may offer greatest value not in replacing traditional stakeholder engagement, but in enhancing evaluators' capacity to process and interpret large volumes of qualitative data systematically.

Second, the technical architecture validation confirms that multi-stage filtering, semantic clustering, and Retrieval-Augmented Generation can maintain methodological rigor when processing evaluation-relevant discourse. The successful implementation of quality control mechanisms - reducing 521,089 messages to 287,127 policy-relevant communications organized into 249 thematic clusters - demonstrates that AI approaches can support systematic data management in evaluation contexts. This addresses core evaluation concerns about data quality and analytical transparency.

Third, our empirical evaluation reveals critical considerations for assessing the utility of AI tools in evaluation practice. User testing, while limited in scale, identified data source quality as a fundamental determinant of tool effectiveness: Telegram communications focused on crisis response proved insufficient for capturing everyday program experiences necessary for comprehensive evaluation. This finding suggests that AI evaluation tools demonstrate greatest promise when applied to rich qualitative datasets such as beneficiary interviews, program participant narratives, or systematic stakeholder consultations - precisely the types of data that evaluation research generates but struggles to analyse at scale.

Fourth, the tool's demonstrated adaptability across domains positions it as a methodological innovation with broad evaluation applications. Beyond migration policy, the architecture shows promise for program evaluation across sectors including health interventions, educational programs, labour market initiatives, and community development projects. This flexibility addresses evaluation practice needs for efficient analysis of diverse stakeholder feedback while maintaining transparency about data sources and analytical processes.

From a meta-evaluation perspective, this research establishes frameworks for assessing quality and ethical dimensions of AI applications in evaluation contexts. The emphasis on transparent retrieval mechanisms, source verification capabilities, and acknowledgment of limitations reflects evaluation standards for methodological rigor and stakeholder protection. The identification of performance constraints and data quality requirements provides guidance for future development of AI-assisted evaluation approaches.

The findings underscore evaluation practice principles: methodological innovation must serve evaluation purposes of credible assessment, meaningful stakeholder engagement, and actionable learning. AI tools should enhance rather than replace participatory evaluation approaches, supporting more effective utilization of qualitative data while preserving authentic voice and agency of program participants and beneficiaries.

**Declaration of conflicting interests:** The authors declare no conflicts of interest in relation to this research. The development of this tool was conducted as part of academic research without commercial funding or influence.

**Use of Generative-AI tools declaration:** The authors used the Anthropic Claude Sonnet 4,5 model to assist with language polishing, as English is not the authors' native language. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for accuracy and interpretation.

**Funding:** This work was supported by the National Science Centre, Poland, grant number 2022/45/B/HS5/00933.

## References

- ALNAP (2009) Real-time evaluations of humanitarian action: An ALNAP Guide. Pilot Version. London: ALNAP/ODI. Available at: <https://alnap.org/help-library/resources/real-time-evaluations-of-humanitarian-action-an-alnap-guide/>
- Allen, R. and Slotterback, C.S. (2017) 'Building immigrant engagement practice in urban planning: The case of Somali refugees in the Twin Cities', *Journal of Urban Affairs*, 43(6), pp. 740-755. Available at: <https://doi.org/10.1080/07352166.2017.1360745>.
- Aoki, G. (2024) Large Language Models in Politics and Democracy: A Comprehensive Survey. Version 2. arXiv. Available at: <https://doi.org/10.48550/ARXIV.2412.04498>.
- Ayoo, S., Leeming, M. and Huff, S.R. (2023) 'Meta-evaluation: Validating program evaluation standards through the United Nations Evaluation Quality Assessment', *Evaluation Journal of Australasia*, 24(1), pp. 14-39. Available at: <https://doi.org/10.1177/1035719X231220979>.
- Bircan, T. and Korkmaz, E.E. (2021) 'Big data for whose sake? Governing migration through artificial intelligence', *Humanities and Social Sciences Communications*, 8(1). Available at: <https://doi.org/10.1057/s41599-021-00910-x>.
- Bouyousfi, S.E. and Ouedraogo, M. (2024) 'Artificial intelligence and big data-driven evaluation research and practices: A systematic literature review', *Evaluation*, 31(3), pp. 303-330. Available at: <https://doi.org/10.1177/13563890241289937>.
- Cousins, J.B. and Alborhamy, Y. (2025) 'Theory-practice connections in collaborative approaches to evaluation: A systematic review of practice', *American Journal of Evaluation* 46(4). Available at: <https://doi.org/10.1177/10982140251355140>.
- Hall, L. and Clapton, W. (2021) 'Programming the machine: Gender, race, sexuality, AI, and the construction of credibility and deceit at the border', *Internet Policy Review*, 10(4). Available at: <https://doi.org/10.14763/2021.4.1601>.
- Liu, Y. (2025) 'Retrieval-Augmented Generation: Methods, Applications and Challenges', *Applied and Computational Engineering*, 142(1), pp. 99-108. Available at: <https://doi.org/10.54254/2755-2721/2025.KL22312>.
- McIntosh, T., Liu, T., Susnjak, T., Alavizadeh, H., Ng, A., Nowrozy, R. and Watters, P. (2023) 'Harnessing GPT-4 for generation of cybersecurity GRC policies: A focus on ransomware attack mitigation', *Computers and Security*, 134. Available at: <https://doi.org/10.1016/j.cose.2023.103424>.
- Melit Devassy, B., George, S. and Nussbaum, P. (2020) 'Unsupervised Clustering of Hyperspectral Paper Data Using t-SNE', *Journal of Imaging*, 6(5), 29. Available at: <https://doi.org/10.3390/jimaging6050029>.
- Nikolopoulou, K., Kehagia, O. and Gavrilut, L. (2023) 'EasyRights: Information Technology Could Facilitate Migrant Access to Human Rights in a Greek Refugee Camp', *Journal of Human Rights and Social Work*, 8(1), pp.22-28. Available at: <https://doi.org/10.1007/s41134-022-00233-0>.
- Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P. and Bernstein, M.S. (2024) Generative Agent Simulations of 1,000 People. arXiv:2411.10109. Available at: <http://arxiv.org/abs/2411.10109>.
- Patton, M.Q. (2010) *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: Guilford Press.
- Potluka, O., Harten, S., Kocks, A. and Dvorak, J. (2025) 'Digitalization in evaluations and evaluations of digitalization: The changing landscape of evaluations', *Evaluation*, 31(3). Available at: <https://doi.org/10.1177/13563890251357650>.
- Scriven, M. (1969) 'An introduction to meta-evaluation', *Educational Product Report*, 2, pp. 36-38.
- Soltoft, J.I., Kocksch, L. and Munk, A.K. (2024) Synthetic Interlocutors. Experiments with Generative AI to Prolong Ethnographic Encounters. Version 1. arXiv. Available at: <https://doi.org/10.48550/ARXIV.2410.11395>.
- Stewart, G. and Al-Khassaweneh, M. (2022) 'An Implementation of the HDBSCAN\* Clustering Algorithm', *Applied Sciences*, 12(5), 2405. Available at: <https://doi.org/10.3390/app12052405>.
- Stufflebeam, D.L. (2001) 'The metaevaluation imperative', *American Journal of Evaluation*, 22(2), pp. 183-209. Available at: [https://doi.org/10.1016/S1098-2140\(01\)00127-8](https://doi.org/10.1016/S1098-2140(01)00127-8).

Zhang, T., Zhang, X., Cools, R. and Simeone, A.L. (2024) 'Focus Agent: LLM-Powered Virtual Focus Group', in Proceedings of the 2024 ACM Conference. Available at: <https://doi.org/10.1145/3652988.3673918>.

## Appendix: Tool Functioning Diagram

This appendix presents a visual representation of how the AI-powered decision support tool operates, designed to illustrate the core mechanics for general audiences without requiring technical expertise.

The tool operates as an intelligent interface between policy makers and authentic community discourse. When a policy maker poses a question such as “What are the main concerns of Ukrainian families regarding children’s education?”, the system:

1. analyses the question - the AI interprets the query to identify relevant themes (education, family concerns, Ukrainian context);
2. searches the knowledge base - the system retrieves messages from the organized collection that relate to education concerns, filtering by Ukrainian speakers and family-related discussions;
3. contextualizes responses - using the RAG architecture, the AI synthesizes information from multiple relevant messages to provide a comprehensive response that reflects authentic community voices;
4. maintains transparency - every response includes links to the original source messages, allowing policy makers to verify authenticity and explore additional context;
5. enables follow-up - users can ask clarifying questions, request examples, or explore related topics through continued conversation.

