Resources

1. Website: https://impactchallenge.withgoogle.com/strengtheningdemocracyeurope/
2. Team Review Proposal Form: pre-grant_application_printout.pdf
3. **Full Application: This file**
4. FAQs and Guidance: faqs_google_impact_challenge.pdf

Instructions for Review

1. **Application Questions in Bold.**
2. Final Answers highlighted in Green.
3. Feel free to suggest edits anywhere.

a. **How you will use any profit that your organisation earns from your proposed project.**

The technology is intended to be open-sourced, for public benefit and as such we do not intend to generate a profit from it. If needed we will develop a business model so as to subsidize its use by local newsrooms that cannot afford to invest in such technology.

The technology is intended to be open-sourced, for public benefit and as such we do not intend to generate a profit from it. However we recognize that there is a potential need to develop a business model so as to subsidize its use by local newsrooms and civic activism orgs. that cannot afford to invest in nor pay for access to such technology. We will employ all generated funds arising from this technology through a contract with a nonprofit called SimPPL run by PIs (Board Members) in order to grow a student developer ecosystem and offer long-term consultancies to develop bespoke deployments and conduct pilots of our civic dialogues technology in countries beyond the scope of this grant.

In the event there is any additional profit generated, it will be used to fund student researchers from the global majority and in the countries we focus on to develop features beyond the scope of the grant. This is in order to encourage grassroots civic participation from younger audiences, whom we have a decade-long history of training through our program SimPPL, where PIs Schroeder and Raman are Board Members and Mehta co-founded in 2021.

b. **How does the charitable activity of your project relate to the regular business or commercial activity of your organisation (if any)?**

As a public university, we do not engage in business or commercial activity from this project. The purpose of the grant is to develop public benefit technology and create an ecosystem in

which it can be sustainably deployed for the benefit of local partners beyond the scope of this grant.

so that the long-term support benefits smaller organizations and newsrooms that we anticipate will benefit from it. We partner also with Deutsche Welle, a national newsroom with a global footprint so that we may make the benefits of deploying this technology in pilots accessible to a large audience including their media development staff based in other European countries.

c. What steps will be taken to separate the charitable activities of your project from the general commercial activities of your organisation (if any)?

The technology will be released with an open-source license so that it is easily accessible for all news and media orgs. We will particularly ensure it is delivered to smaller newsrooms and grassroots orgs. that may not have the technical capacity to deploy it themselves.

but are able to partner with us to deploy pilots locally among their audiences. It is with this type of small-scale impact study that they can receive longer-term support to deploy this innovative technology for civic dialogues.

7. What is the mission/overview of your organization?

TU Delft's mission is to be a leading technical university that pursues excellence in education and research, with a focus on making a positive impact on society and the world.

It focuses on a number of specialized educational programs and supports innovative technology and systems applications.

Institution Name: TU Delft
Budget (Euro per year): 987000000
Number of employees: 4647

9. Links:

Website: TU Delft | Technische Universiteit
LinkedIn: (4) Delft University of Technology: Posts | LinkedIn

10. My project is named:

Civitas: Improving Civic Engagement through Personalized Agentic Messaging by Public Media Institutions

11. **Please describe the challenge you wish to solve, the proposed solution, and the users who would benefit. Please directly use the following prompts to begin sentences in your response:**

11ai. **The challenge this project addresses is…**
There is declining trust in independent media and the integrity of online information which is caused by the opaqueness of civic processes and a lack of accessible channels for citizen participation. We work with DW Akademie on a field experiment to mitigate this issue.

11aii. **The challenge is significant because… (please use data to illustrate the problem statement)**

Pew Research highlighted the 20-year reduction of trust in media before AI generated content accelerated it. Depite DSA/OSA regulation improving platform transparency, they're a vector for disinformation to sow discord. To limit partisan polarization, we require reliable auditing mechanisms that advances private self-deliberation by users.

https://www.europarl.europa.eu/at-your-service/en/be-heard/eurobarometer/civic-engagement

11b. **End beneficiaries:**

11bi. **The end beneficiary group(s) that we hope to support are… (please specify demographics such as geography, gender, ethnicity, age, etc.)**
Required, 50 words maximum.

marie.kilg@dw.comI need some help defining this
DW users and adjacent audiences in eastern Europe with a focus on small languages (Bulgarian, Romanian, Ukrainian, Hungarian) and margin user groups, such as the 50+ and 14-18 age demographics. We target the audiences most at-risk of being influenced by Russian disinformation campaigns and fake news on Telegram/Whatsapp.

11bii. **We are best placed to support them because…**

Required, 50 words maximum.

Eurobarometer states public news is trusted 3x more than private media. Leveraging this trust, we empower a public media institution with GenAI agents to engage in

personalized dialogues with their audiences. This messaging is informed by breaking news given the unique position of Deutsche Welle as a reputable publisher.

11c. **Proposed solution:**

11ci. **The solution we are proposing is…**

Required, 50 words maximum.

Civic literacy interventions by deploying socially intelligent agents to communicate with users in order to engage them in nuanced dialogue informed by verified news. The GenAI agents will interact with public readers of news in conversations over private messaging applications to push verified recent information while encouraging deliberation.

11cii. It will effectively address the problem described above by.. (please provide data, examples, results compared to alternative approaches)
Required, 50 words maximum.

We enable the public to converse and audit online claims through a trusted partner, offering verified information through online retrieval augmented generation. Uniquely, agents engage in long-form dialogue informed by breaking the verified news, rebuilding consumer trust in online information. Personalization and the ability to debate increases messaging persuasiveness (Salvi++; Costello++ 2024).

11d. The outcomes:
Please answer this question.

11di. The quantifiable, tangible change we aim to see in the immediate term of 12-18 months is…
Required, 50 words maximum.
Please answer this question.

Increase in audience engagement with pushed civic messages, user-initiated conversations, submission of claims to audit, volume of conversations pursued, and diversity in conversational themes. Reduction in partisan media sharing, inter-session intervals, and knowledge gaps on local and national governance.

11dii. The quantifiable, tangible change we aim to see in the longer term of 24-36+ months is…
Required, 50 words maximum.

Between baseline and endline surveys we expect participants to exhibit a higher belief in accurate information than misinformation, demonstrate an openness to engage in debate about civic concerns, and increase their participation in civic events (DW organizes some).

11e. Alignment:
Please answer this question.

11ei. My project aligns with the following Google.org Impact Challenge: Strengthening Democracy in Europe focus areas:
Required, select all that apply.
Improving Civic Dialogue

Please review the website for examples of each focus area.

12. **Please provide an overview of the role of technology in your proposed solution and explain how it specifically addresses the problem. Please include the tools, methods, or techniques you are planning to use and explain how users will interact with the technical solution.**
Required, 150 words maximum.

The falling trust in media creates conditions ripe for mis and disinformation campaigns in times of civic uncertainty. We run an experiment to deliver verified information from newsrooms to audiences and allow them to engage in a deliberative process relating to it. We would like to test whether the delivery of personalized information from a trusted public institution can improve civic knowledge, resilience to disinformation, and participation in civic dialogue.

Users are encouraged to engage in debates over the messaging system on WhatsApp and Telegram, and arrive at their own conclusions, supported by verified information presented through a GenAI agent employing online retrieval augmented generation. The agent is aligned with user preferences using direct preference optimization with data (in accordance with GDPR restrictions) obtained from local interviews, focus groups, and supplemented optionally by data on the expression of their preferences as obtained from DW website analytics for the target populations.

13. Why is the technology-based solution needed, how does it augment or is more effective than alternative approaches? If possible, please conceptualise the efficacy of the technology, such as being able to scale to 10x end beneficiaries, reduce the cost of the program by €5 per use, improve the accuracy by 60%, etc. Required, 100 words maximum.

Our work advances a number of other civic chatbots deployed at the city-level such as Boti (Buenos Aires) by improving alignment based on actual expression data for news preferences, personalization based on localized needs, human-centric design applied to define low-friction conversational flows, and free-form debates, cumulatively resulting in increased user knowledge and ultimately improving persuasiveness. We expect to reduce the cost of the conversations to $0.3 per conversation or lower, improve civic knowledge by at least 15% among active users, and conservatively, scale to at least 5x the number of users that in-person media literacy programs cater to.

14. Which of the following best describes the current state of the technical implementation of this project?
Required, select one.
Concept: an idea that does not yet exist and will need to be built

15a. Share additional information about how your idea, approach, or tool has evolved and improved, as well as performance metrics (if available).
Optional, 50 words maximum.

We've built general-purpose personalized messaging systems on Whatsapp and Telegram. We improved health literacy and reduced messaging costs while delivering verified healthcare information to expectant mothers. We used LLMs for semantic understanding and are launching two paid pilots (5,000 users) in Maharashtra, India.

15b. Consider sharing a link to a simple, non-confidential visual representation of the technical components of your project's solution (a diagram or process flow is fine). Please ensure the link is accessible to reviewers.
Optional, link to PDF file preferred.

Technical System Architecture.pdf

16. **Does your project leverage machine learning and/or artificial intelligence? If yes, please answer the following questions. If not, you will be automatically redirected to Q26.**
Yes.

AI Projects
This section is required for projects leveraging machine learning and/or artificial intelligence.

17. In 1-2 sentences, briefly explain the purpose, methodology, input/output data, and task of the proposed AI model. We want to get a quick sense of what your project is planning to achieve. You will have the opportunity to elaborate further in this section.
Required, 75 words maximum.

For example: "For our election information Q&A website, we will use the off-the-shelf ElectionQA LLM that takes user questions in any of 50 supported languages as input, and outputs answers to that question in the user's language. The tool will extract the user's key details (e.g. locale, question details), pull facts from a dataset of local government websites using retrieval-augmented generation (RAG), quote the source website, and direct users to the site for more details."

Socially intelligent agents will be constructed by instruction-tuning large language models on user preference data gathered through surveys, interviews, and website analytics. It is used in direct preference optimization, making the model adapt to and reflect preferences and behaviors of the target audience–tested via "win rates" for agent output against human outputs. This approach enables the agents to facilitate personalized dialogues informed by verified breaking news, enhancing civic engagement and encouraging deliberation on civic topics.

19. Do you plan to develop your own model from scratch, or use an existing publicly available AI model? If you are developing your own model, explain why this is necessary. If you are using an existing model, explain why it is a good fit for your needs.
Required, 100 words maximum.

Ideally we would like to host an instruction tuned variant of Llama 3.1 - 405B to serve responses (multiple adapters). Practically we may rely on aya-101 12.9B (ForAI) and a mixture-of-experts approach with a backstop of the GPT-4 API for use cases where we identify performance gaps from the ensemble (during pilot testing). Both models are natively multilingual and can chat in our target languages, Hungarian, Romanian,

Ukrainian, and Bulgarian off-the-shelf. We will create a serverless infrastructure to serve an inference endpoint from the designated ML model at scale.

20. Please describe any significant datasets you have (or would need) to implement your idea. Please consider sharing information on:
Datatype (e.g., images, text, videos)
Size (e.g., # images or rows)
Attributes (e.g., images, image metadata, image labels)
How frequently data is refreshed
Required, 100 words maximum.

We use out-of-the-box models with retrieval augmented generation over custom hand-translated article data that DW produces for local audiences. In addition, awesome-ukrainian-nlp, romanian-nlp-datasets, awesome-hungarian-nlp, and corresponding bulgarian-NLP tools list datasets in each of these languages on github for generating instruction tuning data where necessary to boost performance, for instance claim-rebuttal pairs. We will focus on constitutional AI approaches to limit toxicity, adversarial attacks and LLM poisoning, and dealing with data privacy issues. We measure robustness and sensitivity in order to minimize the deception of the model by end users.

21. Do you currently have access to this data? Is it public or private? Do you have consent for the proposed use case or sharing? If not, how do you plan to collect or access them?
Required, 100 words maximum.

We possess private data from users through the DW website, and public, free-to-use data from a variety of multilingual datasets listed above. For the website, we intend to utilize views, engagement, and comments with DW stories to inform the priority and weighting scheme of that article in our dataset. Popular articles are likely to feature higher in the user queries as well, and we would benefit from engineering optimizations such as caching and semantic similarity matching over sentences. Our studies will only be conducted with IRB approval from TU Delft following GDPR compliance at each step of the way.

22. What information or decisions will be produced from your data and model, why does it matter, and why would your project's end beneficiaries adopt, use, and integrate the solution to meet their needs?
Required, 150 words maximum.

For example, the model may predict a future value that provides information that may help with a decision by plugging into an existing platform that your project's end beneficiaries already use.

Usage will be incentivized within our experiment for a fixed period of time and we expect a dropoff at incentive withdrawal, but we expect users that followed DW's reporting to continue using it due to easier and faster access to the same information. The system will provide verified information and facilitate deliberative dialogue to reduce partisan biases and increase civic interest, helping strengthen democratic processes. The persuasiveness of LLMs has been able to successfully steer human conversations and mitigate entrenched conspiracy theory beliefs. Our claims verification mechanisms and personalized messaging intervention should help improve civic dialogue for the user.

23. How will you measure and evaluate the performance of your AI solution? How will you know whether the system has succeeded or failed?
Required, 100 words maximum.

Beyond held out test set performance with standard metrics, the socially intelligent agent's effect will be evaluated using a multi-armed randomized controlled trial. The trial will track user engagement via increases in audience interaction with civic messages, user-initiated conversations, and claims submitted for verification. It will monitor behavioral changes, such as reductions in partisan media engagement and inter-session intervals, alongside improvements in civic knowledge. Baseline and endline surveys will assess changes in participants' beliefs in accurate information, openness to civic debate, and participation in civic events. Personalization effects will be isolated by comparing groups with and without personalized messaging. Success is defined by increased civic engagement and knowledge, with stretch goals of reduced partisan bias and improved trust in verified sources.

24. How has your organisation used AI/ML in the past? If this is your organisation's first AI/ML project what has been the greatest deterrent?
Required, 100 words maximum.

For example, data availability and quality, talent availability, computing resources, tool accessibility, competing strategic priorities.

Our organization (TU Delft) has utilized machine learning (ML) in various innovative ways, with PI Raman focusing on personalization and human-centered AI. His research primarily involves developing computational techniques to enhance digital agents' social

intelligence, enabling them to interact with humans more seamlessly in a lifelike manner. This includes understanding and deploying human-like social cues, both verbal and non-verbal, in applications ranging from distance learning to e-healthcare and immersive games. Additionally, he has worked on projects that use ML to analyze and predict social human behaviors, aiming to create interactive systems that better understand and engage with users.

25. What practices does your organisation have in place to ensure your AI solution is developed responsibly? Refer to Google's AI Principles and Responsible AI Practices. Required, 100 words maximum

In this project, responsible AI practices are applied by ensuring ethical and effective development and deployment of the AI solution. The project aligns with Google's AI Principles by creating socially beneficial technology that enhances civic engagement and dialogue. AI agents provide verified information and facilitate discussions to address misinformation and build media trust. Transparency and user privacy are emphasized through adherence to GDPR regulations and IRB approval for data use. The technology is open-source, ensuring accessibility and accountability, benefiting smaller newsrooms and grassroots organizations. Personalization is carefully managed to respect privacy and enhance civic messaging effectiveness.

	Activities	Success Metrics	
H1 2025	Conduct user research to understand target audience needs and preferences. Demonstrate initial prototype of personalized messaging.	Completion of 100 user interviews and surveys. Identifying key product requirements and quantified value proposition. Deployment of initial messaging agent on WhatsApp and Telegram.	
H2 2025	Instruction tune model on user preference data and test performance of social agents. Hold workshops with users	Model alignment increases the 'win rate' of the LLM. 3 feedback workshops conducted with at least 25	

	for testing and commence pilot testing with smaller groups of audiences.	participants each. 90% of trained users demonstrate proficiency at using the system.	
H1 2026	Recruitment of candidates for the multi-wave field experiment. Deploy the system to the target sample of audiences as per experimental design. Monitor the system and gather user feedback.	Onboard 3000 candidates across all languages and regions. Achieve an 80% compliance rate for response to initial messages. Monitor engagement metrics continuously.	
H2 2026	Gather the data from the study and clean the dataset + check for discrepancies. Conduct a thorough analysis and present rigorous results to the scientific community.	Collect feedback from at least 200 users for continuous improvement. Publication submitted. Up to 20% increase in civic awareness and up to 30% increase in civic dialogues.	

27. What are the 1-2 most significant risks or unexpected repercussions you anticipate in this project? Please include any potential negative social or ethical considerations surrounding the use of technology and your plan to minimise these potential risks and negative impacts.
Required, 100 words maximum.

Developing consumer-facing GenAI systems requires addressing risks related to robustness, privacy, and security. Ensuring robustness involves adapting to domain shifts and aligning with user preferences. To prevent misuse, such as DDoS attacks, request rates are capped. Training paradigms like Direct Preference Optimization (DPO), Reinforcement Learning from Human Feedback (RLHF), and Constitutional AI are evaluated for human relevance. Additionally, assessing user engagement is crucial to ensure the system remains 'sticky' and solves meaningful problems without external incentives.

28. What makes your main project team best suited to work on this project? Please directly use the following prompts to begin sentences in your response:
If applicable, your main project team includes team members from your organisation and partner organisations.

Christian builds LLMs, multiagent systems, and safer AI with DeepMind, Adobe, BBC. Chirag leads UX design, personalisation, and safety tuning (preference optimisation, and interface design, in context instructions). Swapneel runs field experiments on platforms given his work with Twitter, Meta, Google. He works with Marie leading DW's AI arm.

28a. The project team's strengths and expertise are…
Required, 50 words maximum.

Fine tuning large language models to increase alignment with human feedback. Building consumer facing tools on private messaging apps (WhatsApp, Truth Social). Disseminating newsroom information to a wider audience. Creating an observability layer to identify user behavioral patterns in this demo.

28c. Some existing gaps of expertise are…
Required, 50 words maximum.

Policymakers that could inform our work from a legislative standpoint. Domain experts in local governments including historians and civics buffs. Platform software engineers who have built consumer facing technology before. Marketing scientists and consumer behavior experts.

28d. We plan to fill these gaps by…
Required, 50 words maximum.

Swapneel's co-advisor co-founded the World Bank's behavioral insights unit and will advise our work. Swapneel's nonprofit, SimPPL (PIs Raman, Schroeder de Witt are Board Directors), has built consumer facing digital literacy systems for India that are deployed in pilots. SimPPL has close connections with policymakers.

29. How will you ensure that your project is inclusive and accounts for varied perspectives and viewpoints, especially from underrepresented populations? How will you collect and use feedback from your project's end beneficiaries?
Required, 100 words maximum.

We reduce language and cultural barriers and support marginalized communities; ethnic minorities, including Hungarians in Ukraine, face challenges related to language rights and cultural identity, affecting their access to education and public services. Groups like the Roma face systemic discrimination, resulting in limited access to education, healthcare, and economic opportunities, perpetuating cycles of poverty and exclusion. We help improve their knowledge of civic processes that they may most benefit from. Feedback is collected at each stage of the process through interviews, surveys, and focus groups. Conversational data is anonymized and utilised with user consent to improve our system.

30. Please list the partner organisations that you are working with or would like to work with who will support this project proposal's execution and success. Please leave this question blank if you are not planning to engage other partners.

Optional, indicate up to 5 partners. We strongly welcome and encourage collaboration - especially between technical and social sector experts in the form of partnerships and coalitions.

Under Status, please type the number that specifies if the partner is (1) Existing partner, (2) In contact with but not a solidified partner, (3), Aspirational partner, not yet contacted. No need to enter ( ), please only enter the number.

Co-PI Schroeder de Witt at Oxford at 10% of his time for 2 years along with overheads and fixed costs at Oxford's rates. Includes travel and hardware budget.	90000	
Cost of field experiment for	135000	

3000 individuals at a recruitment and incentive cost of 30 Euro per individual for a multi-wave survey, and separately includes focus group participation. This recruitment is supported by DW.		
PI Raman at TU Delft will require a Ph.D. student and co-advise a postdoc between himself and Co-PI Schroeder de Witt (split 237000 for student, 165000 for postdoc). Also includes PI Raman's salary at 10%.	585000	
Deutsche Welle costs for program implementation and staff for the program duration. DW runs media development programming in these countries and is able to put us in touch with local support organizations and help with recruitment of individuals.	100000	
Nonprofit SimPPL (subcontractor) will advise the technical deployment and Co-PI Swapneel Mehta will conduct the field experiments	70000	
Travel and conference presentations	20000	

33. How will your project and its impact grow beyond what you have proposed in this application? For example, can it scale directly and/or to other geographies, serve as a model for other efforts, or advance the field?

This project will serve as a model for other newsrooms to take on the role of governance in systems that deliver local information to audiences. Provision of accurate and verified information can be scaled with safe and socially intelligent AI agents. Moreover, newsrooms are opening up to the idea of redefining their role as curators rather than creators of information and we know this because we have sold products to newsrooms including the New York Public Radio, VTDigger, and presented to numerous others including Yonhap in Korea or Times of India. We aim to create a viable model to continue funding this messaging system with a paid version that might allow access to highly engaged users on a continuing basis beyond the experiment duration.

34. How would you sustain funding for this work beyond Google.org's funding?

Within the duration of the project we aim to identify the value proposition we are delivering to end users of this system so that we are able to identify a relevant business model that sustains the project beyond the initial grant. At SimPPL, the team has built and scaled technology like https://parrot.report to detect threat actors and built https://sakhi.simppl.org beyond the scope of the smaller grants, into its own organization to advance social good. We aim to do the same with Civitas.

36. Please provide links to any relevant previous public speaking engagements or media features where your team members have discussed the project or their expertise in related areas.

Building whatsapp based systems for digital health literacy: Swapneel Mehta, Rest of World Feature - https://restofworld.org/2024/3-minutes-with-swapneel-mehta-simppl

33. **How will your project and its impact grow beyond what you have proposed in this application? For example, can it scale directly and/or to other geographies, serve as a model for other efforts, or advance the field?**

This project will be open-sourced to serve as a model for other newsrooms to take on the role of governance in systems that deliver local information to audiences. Provision of accurate and verified information can be scaled with safe and socially intelligent AI agents. Moreover, newsrooms may need to redefine their role as curators rather than creators of information. At SimPPL, we have also sold products to newsrooms including

the New York Public Radio, VTDigger, and presented to numerous others including Yonhap in Korea or Times of India so we understand their new role in the information ecosystem.

34. How would you sustain funding for this work beyond Google.org's funding?

Within the duration of the project we aim to identify the value proposition we are delivering to end users of this system so that we are able to identify a relevant business model that sustains the project beyond the initial grant. At SimPPL, the team has built and scaled technology like https://parrot.report to detect threat actors and built https://sakhi.simppl.org beyond the scope of the smaller grants, into its own organization to advance social good. We aim to do the same with Civitas.

35. How will you share your best practices and lessons with other stakeholders? Please share a light-touch communication plan to disseminate learnings. In your response you may want to reference:
Your preferred audience (e.g., community organisations, policymakers)
Ways of communicating (e.g., workshops, social media, newsletters, conferences)
Content you'd like to share (e.g., case studies, reports)
Required, 100 words maximum.

To share best practices and lessons, the project will focus on engaging community organizations, policymakers, and media stakeholders. Communication will occur through workshops, conferences, and social media platforms. Content shared will include case studies, reports, and insights from field experiments. This approach ensures broad dissemination and encourages collaboration and feedback. Additionally, newsletters will be used to regularly update stakeholders on progress and findings. By leveraging these channels, the project aims to foster a community of practice that supports the adoption of successful strategies and enhances civic engagement across diverse populations.

Beyond a channel to 'push' updates on preferred topics and engage in a deliberative process, the user is incentivized to submit queries soliciting information on topics of civic importance such as legislation, local governments, civic events, and elections and submit uncertain claims for review. The system complements media development efforts locally that DW has been engaged in for a number of years. It offers a natural distribution mechanism from a reputed and trusted source, increasing its adoption and ensuring higher rates of compliance among users.

Among participants in our field experiment with access to this messaging system, we expect to see a reduced partisan lean, and increased interest in civic concerns including local elections, governance, and legislation. We will test whether the provision of a personalized messaging system increases deliberative dialogue and improves civic engagement and isolate the differential effect of personalization in such a context. We will compare those with access to personalized messaging systems with another group without access to personalization, and the control group without access to the system itself, creating a multi-armed randomized controlled trial. We will account for demographic, temporal, usage-based, literacy and socio-economic status-based confounders. Between baseline and endline surveys we expect increased belief in the democratic process, confidence in civic knowledge, and openness for debate.