Lacuna
Fund

Request for Proposals:
Natural Language Processing (NLP) - 2024

Questions and Answers

Application	
1.Q	Is there a limitation on the number of applications a single institution can submit?
1.A	There is no limit on the number of proposals that a single institution can submit.
2.Q	Is there a specific format of proposal outline that we should follow?
2.A	According to Section 4 (Proposal Information) of the Lacuna Fund's Natural Language Processing RFP 2024 document, there is a specific format. The proposal narrative should be limited to 10 pages, not including references, with 2.5 cm margins and a minimum of 11-point font. The narrative should address qualifications, problem identification and proposed solution, specifications and deliverables for proposed data and documentation, intended beneficiaries and use cases, methodology, and other relevant details. Fill out information for your proposal, including application questions and budget and deliverable templates on the SurveyMonkey Apply application portal.
3.Q	When and how will the winning proposals be announced? Will it be public, or will only the winning teams be informed?
3. A	The winning proposals will be announced in spring 2025. Winning teams will be notified, and once contracts are executed, grantees will be publicly announced in the Lacuna Fund newsletter and posted on the website. To subscribe to the newsletter, scroll to the bottom of the homepage at lacunafund.org and enter your contact information.
4.Q	Is Portuguese an eligible language?
4.A	We are accepting applications in Portuguese. For those submitting an application in Portuguese, you may apply using either the English, Spanish, or French portals. We do not have a Portuguese portal option available at this time. However, proposals submitted Portuguese in any portal are accepted and will be reviewed.
5.Q	Under section 4 - Proposal Information & subsection "Applicant Information," can you please clarify points 5 & 6. Regarding point 5

	specifically, is it the intention that each participating organization submits their code of conduct/ethics or should each submit the code of conduct established by their corresponding host country ethical committee?
5.A	Point 5 refers to the information about the affiliated institution(s) ethical review processes requested in the “Proposal Narrative” section called “Risks, Including Ethics and Privacy”. The intention is for the applicant to “Identify issues and potential risks, including but not limited to potential privacy and ethical concerns, and describe steps you will take to mitigate them.” You could reference the lead organization’s code of conduct/ethics as part of your response. Point 6 refers to information about the team’s ability to gain national approvals. Please describe your plans to gain approval of your project from the host country ethical committee. If you do not plan to seek national approvals in any of these countries in which research will take place, please explain how your proposed project falls outside the appropriate regulatory authority’s definition of research. For more information, see the application questions available on the Apply page and in the SurveyMonkey Apply application portal.
6.Q	Apart from the current call, can Lacuna Fund support a structure in equipping one or more languages with digital resources? If yes, how to proceed?
6.A	Lacuna Fund only currently supports proposals through a request for proposals process. Future calls for proposals will be available at the Apply page and announced through the newsletter (you can subscribe by scrolling down to the bottom of the home page at https://lacunafund.org/). We also offer General Resources and Language Resources on our website which are available to all. Please see the Datasets page to view datasets in the domains of agriculture, language, and health that have been created with support from Lacuna Fund.

Eligibility

7.Q	I wish to apply for Lacuna fund to develop a dataset in eye disease images. Would I qualify for the funding since it's named NLP but my dataset would mostly be made up of images because this is a challenge I have encountered due to few or non-existent datasets in this area?
7.A	While the RFP focuses on NLP, it does acknowledge the need for datasets in related areas that may support NLP technologies indirectly. The key

	consideration would be to explain how your proposed dataset in eye disease images could potentially contribute to or complement NLP research or applications, especially in the context of low-resource settings. To determine your eligibility and the relevance of your proposal, we encourage you to carefully review Section 3.2.1 and consider how your project aligns with the objectives and criteria outlined in the RFP document. You may also want to explore potential partnerships or interdisciplinary approaches that could strengthen your proposal.
8.Q	We are considering counterparts in other countries in the region. However, our organization cannot make payments outside Argentina. Would it be possible to have two Primary Institutions, or would it be possible that more than one institution received funds directly from you? In the same line, would it be possible to have two Project Leaders, or a Co-Leader?
8.A	As it states in the RFP, partnerships are strongly encouraged as a way to strengthen collaboration and maximize the benefits derived from the use of the datasets, but only the lead applicant will receive funds. It is not possible for more than one institution on a team to receive funds directly from Lacuna Fund. If you are able to identify a fiscal sponsor or partner that is able to make payments outside the country, they may apply as the lead applicant or primary institution as long as they are headquartered in the geographic focus of this call (i.e. Africa or Latin America).
9.Q	Can you confirm that a university in Colombia could be an eligible entity as an executor, being a Higher Education Institution?
9.A	As states in the RFP, in the "Organizational Eligibility" section, "To be eligible for funding, organizations must: Be either a non-profit entity, research institution, for-profit social enterprise, or a team of such organizations [...] Have a mission supporting societal good, broadly defined." Universities are eligible to apply for this call for proposals, as well as other organizations with a mission supporting social good, broadly defined (including non-profit entities, research institutions, for-profit social enterprises, or a team of such organizations).
10.Q	Our organization is the Primary Institution, but there will also be Universities involved. Is any kind of formal agreement documentation required to submit the proposal?

10.A	No formal documentation agreement is required to submit the proposal except for responding to the application questions and certifications found within the application portal. Letters of support are optional. However, if your project is selected, there would be formal due diligence processes and other contractual documentation required.
Partners	
11.Q	Do you encourage partnership between academic institutions and startups?
11.A	As it states in the RFP, partnerships are strongly encouraged as a way to strengthen collaboration and maximize the benefits derived from the use of the datasets. Both academic institutions and startups are eligible to apply, if they have a mission supporting societal good, broadly defined.
12.Q	Are commercial partners explicitly forbidden? Could one include part of the budget for purchasing professional services if needed? On the webinar it was emphasized that only non-profit organizations can apply. Does this explicitly exclude commercial organizations as partners? I am particularly thinking of companies who could help support finding/paying a lot of microtasking. Can some budget be put towards purchasing professional services if needed?
12.A	As states in the RFP, in the "Organizational Eligibility" section, "To be eligible for funding, organizations must: Be either a non-profit entity, research institution, for-profit social enterprise, or a team of such organizations... Have a mission supporting societal good, broadly defined." Commercial partners are welcome. For this RFP, commercial enterprises/start-ups with a mission for social good may apply. Part of the budget could be used for purchasing professional services.
13.Q	If an organization outside the eligible region applies in partnership with an in-region institution, who is considered the lead applicant and the receiver of funds?
13.A	The lead applicant must be in-region and would be the receiver of funds (and then could distribute the funds). Other partners are allowed outside of the target region. According to Section 3.1 (Eligibility Criteria) of the Lacuna Fund's Natural Language Processing RFP 2024 document, proposals must be led by an institution headquartered in Africa or Latin America. If an organization outside these regions applies in partnership with an eligible in-region institution, the in-region institution would be considered the lead applicant and receiver of funds. The lead applicant must be from an eligible institution in Africa or Latin America to ensure alignment with the

	fund's geographic focus and objectives. Partnerships with organizations outside these regions can strengthen proposals but do not alter the lead applicant status.

Budget

14.Q	What is the maximum amount one can apply for?
14.A	According to the Budget section of the RFP document (Section 4) The total pool available is approximately $1 million USD. We would like to fund projects in each of the target regions (Africa, Latin America) and anticipate supporting 6-8 smaller projects with budgets up to $100k USD and 2-3 larger, more complex projects with budgets ranging from $100-250k USD. The Technical Advisory Panel will assess the feasibility and suitability of the budget as well as the linkage between the budget and grant narrative as part of the selection criteria. The RFP does not list a maximum amount of funding that one project can receive. Please apply for what you believe that you would need to complete your proposed dataset; however, we encourage you to keep these ranges in mind to increase your likelihood of receiving funding.
15.Q	Is there an official template for preparing the budget?
15.A	Yes, you can find links to the official budget template and deliverables template on the website and in the SurveyMonkey Apply application portal (NLP 2024): • Budget Template (English) • Deliverables Template (English) Translations of these templates are available in the SMA application portal. To access these templates, toggle the portal to the language of your choice. *Note:* We are accepting applications in Portuguese. For those submitting an application in Portuguese, you may apply using either the English, Spanish, or French portals. We do not have a Portuguese portal option available at this time. However, proposals submitted Portuguese in any portal are accepted and will be reviewed.
16.Q	What are the requirements and the methods supported for transferring funds?

16.A	Payments are paid via electronic fund transfer (EFT) or wire transfer with a direct deposit agreement form and an international wire transfer form. For teams that are selected as finalists to receive funds, the lead organization must submit their due diligence documents for review before being approved to award funds. We require you to conduct due diligence on any partner organizations that will be receiving grant funds.

Dataset Specificity and Scope Requirements

Languages

17.Q	How many languages should be the minimum or maximum consideration per project? What if our proposal is specific to only one African language? Is it still eligible?
17.A	According to the Lacuna Fund's NLP RFP 2024 document, there is no specific mention of a minimum or maximum number of languages that must be considered per project. However, proposals should focus on addressing challenges in low-resource languages and cultures in Africa and Latin America. While there is no strict requirement on the number of languages (and a proposal specific to only one language could be eligible), proposers should justify the selection of language variety(ies) and their impact on underrepresented cultures based on the potential impact and relevance to the goals of the RFP. The project can focus on a single language which includes different language varieties due to the task at hand. For example, if the project focuses on creating culturally relevant hate speech datasets, it might include different in-group language varieties used by different social groups
18.Q	Since Google Translation Dataset has 785 million records spanning across 548 languages, I think I need more clarify of what specific African and Latin America language are underrepresented. Is Portuguese an eligible language? What are the characteristics of a language eligible for a project?
18.A	This RFP focuses on “low-resource languages and cultures in Africa and Latin America.” There are thousands of African languages, all of which could be considered underrepresented. Similarly, Latin American

	languages, especially indigenous ones, can also be viewed in this context. We are particularly focused on under-resourced languages. However, higher-resource languages, such as Portuguese, could be included if the dataset is supposed to have an impact on a culture underrepresented in NLP technology. For example, this could be in a context such as code-switching with another language/speech community from Latin America, such as Piamontese poor immigrants, or Portuguese Creoles (i.e. Cape Verdean Creole, Guinea-Bissau Creole, Papiamento) Another example is when the language variety includes slang or derogatory vocabulary that is region specific, as usual in alignment datasets. We understand that you know your communities best and what languages have resources or need resources. Consider that a supposedly high-resource language might be low-resource in a specific domain, such as question answering over images for blind people for example in a Latin American country. In this context, blind individuals take pictures and ask questions about products they use daily, which might not be recognized by state-of-the-art multimodal language models because the brands, packaging, and other details do not exist in the northern hemisphere. While French, Spanish, and Portuguese have greater representation, there may still be instances (e.g. domains such as digital violence and health) of under-resourcing for language technologies, depending on the situation. For instance, a language can be deemed under-resourced even if it is spoken by hundreds of millions of people. For more information, see these Language Resources on the Lacuna Fund website. https://lacunafund.org/language-resources/
19.Q	Could Holy book and Language book be classified as existing dataset in Africa and Latin America?
19.A	According to the Natural Language Processing RFP 2024 document, existing datasets are defined as those that are openly accessible and can support the development of machine learning models for under-resourced languages and cultures in Africa and Latin America. Holy books and Language books can indeed be classified as existing datasets if they meet these criteria. These texts can be considered existing datasets for some languages. Additionally, there are low-resource languages with communities attempting to revitalize them that only have records made by figures such as priests for evangelization purposes. However, this type of documentation requires rigorous interpretative methodologies to repurpose these records and analyses for NLP and other applications

	beyond their original religious context. Furthermore, it is important to note that religious texts should be used with extreme care, as they may involve significant ethical risks. For more information, see these Language Resources on the Lacuna Fund website. https://lacunafund.org/language-resources/
Technical Considerations	
20.Q	For speech datasets, is there a minimum number of hours of recording?
20.A	According to the Lacuna Fund's NLP RFP 2024 document, there is no specific mention of a minimum number of hours of recording required for speech datasets. However, proposals should provide sufficient justification for the dataset's adequacy to support natural language processing (NLP) technologies for low-resource languages and cultures in Africa and Latin America. Proposals are expected to demonstrate the dataset's potential impact on improving machine learning models and addressing challenges specific to speech processing. Feel free to check out previous datasets supported by Lacuna Fund here: https://lacunafund.org/datasets/
21.Q	Is there specific software for labeling data?
21.A	According to the Natural Language Processing RFP 2024 document, the RFP does not mandate the use of specific software for labeling data. However, it emphasizes the importance of using robust and reliable tools to ensure the quality and accuracy of the labeled datasets. The choice of software should align with the project's goals and the specific requirements of the dataset being developed and preserve fair data principles. If you use LLMs or other foundational models in any stage of your annotation this information needs to be disclosed and made clear in your methodology. If you use crowdworker remote platforms, you need to follow ethical practices of paying a fair wage and requesting informed consent as you would do in in-person collection methodologies.
22.Q	Does hosting the dataset produced also take into consideration public platforms like HuggingFace, Kaggle etc?
22.A	According to the RFP section “Proposal Narrative,” Lacuna Fund emphasizes the importance of making datasets openly accessible to the broader research community to maximize impact. Hosting platforms must assign a digital object identifier (DOI) to the dataset, quantify downloads of the dataset, and collect contact information for dataset downloads. Please

	see more guidance for suggested hosting platforms in Lacuna Fund's Dataset Hosting and Documentation Guidance. While the RFP does not specifically mention platforms like HuggingFace or Kaggle, it encourages the use of reputable and widely-used platforms to ensure broad accessibility and usability. Hosting in HuggingFace makes sense if the dataset does not contain private or sensitive information. An explanation of how informed consent will be collected is crucial for all data. It is especially important for data that might be culturally sensitive, and the application should discuss steps, such as deidentification, that might be taken to mitigate risk. It is important to notice that what is not sensitive information in a culture might be sensitive in another culture.
Data Quality	
23.Q	Can you please clarify the question of providing a sample model and data before submission, should we build a sample before the beginning of the contract? Is there going to be a pilot stage in the project?
23.A	The project can include baseline experiments as well as sample models for evaluation. This might involve testing (State of the Art)SOTA models on your dataset and tasks as is routinely done in NLP. You don't need to provide an improved model, just a demonstration of the usefulness and usability with your data and task. The methodology used and costs associated with this small pilot phase should be included in your proposal, there will not be another call for this.
24.Q	What common mistakes have you seen people make on RFPs that hurt their chances of approval?
24.A	Ensure that your project meets each of the eligibility requirements and fill out all sections of the proposal, addressing each requirement of the narrative (see the "Proposal Narrative" section of the RFP), so that your proposal will not be marked ineligible. Make sure to include clear use cases and sustainability plans. Examples of issues include having incorrect licenses, lacking a clear methodology for preserving data quality, or not having well-defined hosting policies (i.e., specifying where the data will be stored).
Use Cases	
25.Q	What level of specificity is required in articulating how the dataset can help solve the use cases? Is this something that you'd want to see quantified? If so, can they be high level estimates, or do you require thorough quantification?

25.A	According to Section Proposal Narrative of the Lacuna Fund's NLP RFP 2024 document, it's important to clearly explain how the proposed dataset will tackle specific challenges in natural language processing. Section Selection Process and Evaluation Criteria, mentions that we are looking for proposals that show a solid grasp of the problem and how the dataset will help solve it. We are keen to see both rough estimates and, if possible, detailed quantification of the expected impacts, benefits, and risks.
26.Q	Given that these datasets can help solve an array of problems, what is the ideal number of use cases to highlight as potential targets for the dataset?
26.A	According to the Lacuna Fund's NLP RFP 2024 document, there is no specific mention of an ideal number of use cases to highlight for the dataset. However, proposers are encouraged to focus on a manageable number of use cases that demonstrate the dataset's potential impact and relevance to addressing challenges in natural language processing for low-resource languages and cultures in Africa and Latin America. The RFP emphasizes the importance of quality over quantity in showcasing use cases. Proposals should clearly articulate how the dataset will contribute to solving specific problems in NLP applications, providing detailed insights into the expected benefits and outcomes.
Policies	
27.Q	The RFP seems to say that both the data and any tools used must be made open source. This would seem to exclude the use of tools that have been previously created whose owners do not want to open source them? Again I am thinking of a partner company who supplies the use of a platform for data collection. If partners have tools that they are willing to allow us to use, is it acceptable that they don't share the IP of those tools?
27.A	The dataset itself is what's important. You can use commercial tools to create it, but the dataset itself, associated metadata, and any models or publications developed with grant funds must be made openly available using one of the licenses listed in our IP policy. Check out our Intellectual Property (IP) Policy.
28.Q	Are there any specific potential data sources that shouldn't be used for privacy/ethical reasons?
28.A	Grantees must ensure datasets are curated ethically, e.g avoiding material that is Copyright protected. Furthermore, if you are working with medical data, for example, this could have sensitive or confidential information. You can be cautious about this kind of data and explain your methodology for deanonymization if necessary. In the case of many indigenous

	communities, there are types of knowledge, such as medical or religious knowledge, that may be desired to be archived for preservation purposes but not made freely accessible to the general public. This type of documentation can be useful and even used for NLP, but it must be safeguarded, and such safeguarding should be specified in the project. If sensitive information related to Indigenous knowledge/religion is included, applicants would need to request a license exception to the Lacuna Fund IP policy. Check out our intellectual property policy for more details on licenses when privacy might be a concern. Please also refer to Lacuna Fund principles on Ethics (linked on p. 4 of the RFP) and the evaluation criteria on Ethics on p. 5 of the RFP, as well as the guidance on Data Management and Licensing & Ethics and Privacy on p. 10, Section 4 – Proposal Narrative of the RFP. You can also check out our NLP resources at: https://lacunafund.org/language-resources/
Mentorship	
29.Q	Is it compulsory to get a mentor?
29.A	No, it is not required. It is a complimentary opportunity available for you.
30.Q	What is the process to benefit from mentorship? What would this Masakhane mentorship be?
30.A	The main benefit from mentorship is to have a mentor who will review your draft proposal and discuss avenues for strengthening it. Mentees can request different forms of assistance such as: discussing gaps in low-resource NLP, writing a research proposal and preparing budgets. For more information, see the recordings and slides of the applicant webinar available on the Lacuna Fund website at: • Recording (English): https://vimeo.com/982223215 • Recording (French): https://vimeo.com/982223384 • Recording (Portuguese): https://vimeo.com/982223670 • Recording (Spanish): https://vimeo.com/982223853 Slides: • English • French • Portuguese • Spanish The Masakhane Mentorship Program is available through this Google Form; the original deadline was 15 July 2024.
31.Q	Is Masakhane mentorship in English or another language?

31.A	The mentorship will be available in each language (English, French, Spanish, Portuguese). When you apply for mentorship, feel free to indicate which language you would like mentorship, and the Masakhane team will let you know if they can accommodate your request. You can also reach out to Masakhane on the email address listed with any questions to:| masakhane_leadership@googlegroups.com