Preparatory Ethics Questionnaire

What kind of data do you produce or use in your research?	Numerical data: Data that consists of numbers, which can be either discrete or continuous. This includes data derived from mathematical formulas, statistical calculations, or measurements such as temperature, height, weight, and time. Categorical data: Data that represents distinct categories or groups, without any inherent numerical value. This includes data such as gender, race, marital status, or types of animals. Textual data: Data that consists of text, such as written documents, articles, social media posts, and emails. This can include data generated by humans, such as legal texts or transcriptions of interviews. Synthetic data: Data that is generated by artificial intelligence (AI) algorithms or synthetic data generators for various purposes, such as testing, simulation, or training machine learning models. Qualitative data: Data that is obtained through qualitative research methods, such as interviews, surveys, or questionnaires, and represents subjective opinions, beliefs, or experiences of human participants.
The topic or focus of your research can be classified within the following group of disciplines:	Information Sciences, including ICT
In your opinion, could your research involve, develop, produce or use dual-use items, technology or software?	No
Please select what applies to your research:	My research includes active involvement of human research participants AND/OR gathering new data from human participants Within my research project I design or develop AI-technologies/algorithms, or the project involves the deployment and/or use of AI-technologies/algorithms for practical applications
How many people will participate, who are they, and how will they be selected?	We plan to have 20 to 200 people participate as annotators. The annotators will participants obtained through either 1) Prolific, a paid online crowdsourcing platform specialized in gathering human data for research studies and AI dataset creation, or 2) from the student population at UvA. Screening prerequisites will be that (i) participants must be fluent in English and (ii) they should not have taken part in any initial pilot study.
How will participants be approached and recruited?	Participants will be recruited through either 1) Prolific's specific support for participant selection, or 2) through an contact from one of the INDElab researchers.
What research method(s) do you plan to use, e.g. interview, questionnaire, field observation, audio/video recording, etc.?	The task consists of the following: a human annotator is presented with 1) a summary of a Wikipedia page, and 2) a statement generated from the Wikidata knowledge graph triple that has been aligned with the summary. The annotator must indicate if they disagree with the statement, and if so, whether they disagree on the factuality of the statement or the meaning of any of the terms used in the statement. We intend to have each annotator to provide annotations for 20 summary/statement pairs. Annotators will use a web interface generated using Potato, an open-source annotation tool. Annotators will be first presented with a pre-annotation screen outlining the annotation guidelines, after which they will be asked to annotate 20 items individually.
Where will the project be carried out, e.g. public place, lab space,	The project will be conducted completely online used Web-hosted user interfaces and email.

researcher’s office, private office at organisation?	
Will the study involve engaging participants in the discussion of distressing or sensitive topics (e.g. sexual activity, drug use, ethnicity, political behaviour, potentially illegal activities)?	No
What process(es) will you employ to ensure that participants are freely giving information consent to participate?	Participants selected through the Prolific service are managed and screened by that service; participants from the UvA student population will be asked to provide consent using a web form.
Unless specifically and clearly consented (e.g. a media release form), will it be possible to link personal data back to individual participants in any way?	No
What mechanisms are in place to ensure that participants can withdraw participation and/or their data?	Participants selected through the Prolific service are managed and screened by that service; participants from the UvA student population will be able to email researchers if they wish to withdraw participation and/or their data.
Will deception of any sort be used?	No
Will participants receive any compensation for taking part?	Yes
If you intend to make payments to participants please state this here and explain your justification. Payments may be made to participants for reimbursement of travelling, out-of- pocket expenses and compensation for time. An investigator who wishes to make any other payment must state his/her reasons for wishing to do so.:	We may offer a small gift certficiate for participation (5-10 euros).
Consider whether the data or model used, encodes, contains, or potentially exacerbates bias against people of a certain gender, race, sexuality, or who have other protected characteristics. For instance, does the used data set represent the diversity of the community where the approach is intended to be deployed? Does the used model perform worse for some groups than for others?	likely/possibly

Preparatory Ethics Questionnaire

Please explain:	The models being used in this study to annotate the data are both open-source and proprietary large language models, accessed through publicly-available inference APIs. These models are known to contain certain biases due to their training on large corpora of Internet content. However, they are used in this study to make factuality and topicality classifications of knowledge graph statements, and as such their use should not result in unfair treatment or disadvantage anyone involved.
Is there a non-negligible risk of misapplication of the AI, i.e. a risk that the model will be applied to a dataset which is not representative of the dataset on which it was trained, and thus produce results that will misinform human decision making?	Yes
Please explain:	The models being used in this study to annotate the data are both open-source and proprietary large language models, accessed through publically-avaliable inference APIs. It is remotely conceivable that the models used in this study might be misapplied, but any such use would be out of the context of this project.
Does the AI likely generate decisions (or advice) that would not be explainable, resulting in professional users (e.g., factory operators, medical staff, policy makers, and others) being unable to provide justifications for the decisions informed by the AI model?	Yes, the AI’s decisions (advice) are (potentially) not explainable
Please explain:	Large language models are currently not able to guarantee a factual, coherent explanation for their outputs. The design of our use of LLMs does rely on chain-of-thought techniques where the model will be used to produce a rationale for its decision making; as an extrinsic explainability method, this will allow some insight into plausible reasons for the model's classification decision.
Is there a non-negligible risk that those who will use the AI model to inform their decision making will have unrealistic expectations about its capacities and misinterpret its output?	Yes
Please explain:	Large language models can be misused by humans who are interpreting their output in the context of a decision making process. But this is true regardless of anything specific that this project is doing; this project is focused on evaluating a language model's ability to accurately determine whether a statement from a knowledge base is false, and if so, whether the reason for that determination is because the model disagrees with the factuality of the statement or with the meaning of the terms used in the statement, providing a rationale for its judgement. Therefore we believe the risk cited above is low.

Preparatory Ethics Questionnaire

Is it likely the AI generate decisions (or advice) that negatively impact people or society, including through plausible alternative uses?	Yes
Please explain:	In general, large language model output can lead to harms if misused or misapplied by humans. Again, this is true regardless of anything specific that this project is doing; this project is focused on evaluating a language model's ability to accurately determine whether a statement from a knowledge base is false, and if so, whether the reason for that determination is because the model disagrees with the factuality of the statement or with the meaning of the terms used in the statement, providing a rationale for its judgement. Therefore we believe the risk cited above is low.
Is it likely that in the future the AI could be (a) used in a way that violates people’s privacy and could potentially result in surveillance, or (b) forced on people without their consent, or (c) used in other, ethically questionable, ways?	No
Please explain:	In the case of this study, the large language models are simply being used to generate judgments about the factuality and topicality of knowledge graph statements. There is no aspect of their use in this case that impinges on peoples' privacy.
Do any of the parties involved in overseeing or carrying out the research have a potential conflict of interest?	No
Will the study expose the researcher to any risks (e.g. when collecting data in potentially dangerous environments or through dangerous activities, when dealing with sensitive or distressing topics, or when working in a setting that may pose ‘lone worker’ risks)?	No
Are there any other ethical considerations that have not been covered by the questions above, and on which you would like advice/guidance?	No
The applicant agrees to have answered this form truthfully to the best of their knowledge.	Yes
Would you like to have advice on an issue by the ethical committee of your sector?	I don’t encounter an ethical issue

Preparatory Ethics Questionnaire

Do you have a public sensitive topic in your research project?	No
How do you appreciate the Research Management Services?	#####