**Ethical Complaince**

This document describes the ethical considerations and compliance measures for a project that aims to develop a Hinglish language model by processing Hindi subtitles extracted from YouTube videos. The project uses machine learning techniques to analyze and learn  linguistic patterns present in  subtitles, contributing to advances in natural language processing of Hinglish, a mixed language of Hindi and English widely spoken  in India.

**Ethical Considerations Copyright Compliance:**

Captions are taken from publicly available YouTube content and are subject to YouTube's Data API Terms of Service. Subtitle extraction  is done in a manner that respects the copyright  of the video creator. The Project uses  data for research and educational purposes in accordance with the principles of the Fair Use Doctrine, but recognizes that fair use is jurisdiction dependent and subject to legal interpretation.

**Data protection and GDPR compliance:**

The extracted data  (subtitles) is not personal, it does not contain any identifiable personal data and therefore complies with GDPR regulations. Precautions will be taken to ensure that  personal data within the meaning of  GDPR is not inadvertently collected, stored or processed.

**Transparency and Accountability:** Data collection, processing, and model development methodologies are well documented  and  transparent in all project reports and publications. This project includes a contact point  for questions regarding data use, data protection, and ethical considerations.

**Data Use and Storage:** The data collected  is used only to build language models and improve Hinglish's natural language processing capabilities. Data is stored securely and access is restricted to  project team members and protected against unauthorized access, security breaches, and data loss.

**Ethical AI and fairness:** Models are designed to be unbiased and unbiased to avoid maintaining or reinforcing harmful stereotypes and biases  in the training data. Continuous monitoring is implemented to detect and mitigate  unintended consequences and misuse of  language models.

**Compliance Measures Regular Audits:** We conduct regular audits of how we collect, process, and store the data  to ensure continued compliance with ethical standards and legal requirements.

**Data Minimization:** Follow the data minimization principle to limit the data collected to  what is needed to build the language model.

**Impact Assessment:** Conduct a  impact assessment to identify potential adverse effects on individuals or groups and take steps to reduce these risks.

**Consent and Rights:** Although the data used is publicly available and not private, the project recognizes the intellectual property rights of the original content creators and uses the data in a manner that respects these rights.

**Conclusion**

This project is committed to maintaining high ethical standards in the development and use of the Hinglish language model, by adhering to these guidelines, this project aims to make a positive contribution to the field of natural language processing, while respecting the rights of individuals and the broader community.