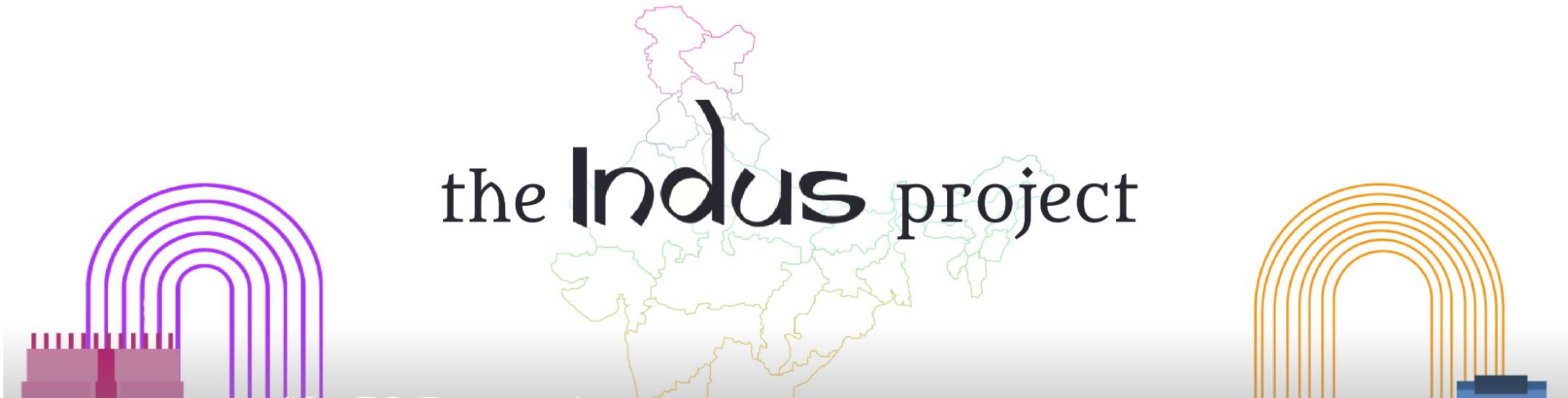
The Indus Project

Annotation and Ethical Review



The beginning- Sam Altman in India

- Tech Mahindra CEO Gurnani Takes up OpenAI founder's challenge

CP Gurnani
@C_P_Gurnani

OpenAI founder Sam Altman said it's pretty hopeless for Indian companies to try and compete with them.

Dear @sama, From one CEO to another..

CHALLENGE ACCEPTED.

5:36 PM · 9 Jun, 2023

1.4K replies **3K** shares **11.7K** likes

Nikhil Malhotra @nickmalhotra · Jun 9

Sam is entitled to his opinion and I respect that , in all humili
create a model that would be better for Indian dialects than
bet and the race is on @sama . #makerslab

Smoke-away @SmokeAwayyy · Jun 9

Sam Altman: "the way this works is we're going to tell you i
hopeless to compete with us on training foundation model
shouldn't try, and it's your job to like try anyway. And I belie
those things. I think it is pretty hopeless"

Challenge taken up by Tech Mahindra

What Are we Building

- An Indic Large Language Model for Indian Languages
- Starting with Hindi and its many dialects
- Aim to make a first-of-its-kind Indic LLM

Why Are We Doing This

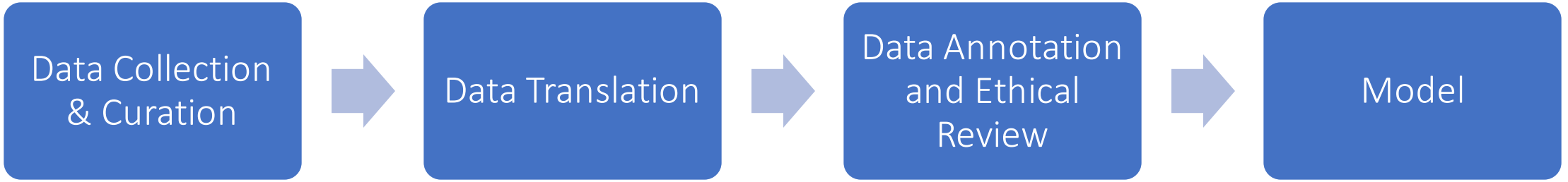
Never been made before

Make in India

First of its kind- Make history

What have we done

Where are we



the Indus project

a “Civilizational” initiative

“ Intending to empower all Indic languages that have originated out of the great Indus Civilization.
We wish to build an Open Source Large Language AI model to serve the needs of 25% of the world's population! ”

Make Your Contribution

The Indus Project Website- Bhasha Dan

What do you think is the biggest problem with Artificial Intelligence?

Ethics and AI
Your contribution

AI can never compete with or match our ethical morals

Our responsibility to make history which is ethically regulated

Annotate and Review Data based on the guidelines

Annotation Objective

To ensure that the AI model gets trained on grammatically correct language, which is filtered for ethicality

High level Instructions:

- The text should be read and edited to ensure only well-formed sentences are present.
- Any other garbled text or text which does not make sense may be removed
- The sentence needs to be checked for grammatical correctness
- Every sentence needs to have a poorna viram
- Statements need to be checked for Ethicality as per guidelines

Ethical Considerations

Possible Bias	Descriptions	Example and context
Violence	Any statement which has blatant violence needs to be removed from the training data	– 2017 कानाड की मस्जिद में आंतकी हमला – मुझे खून की गंध आ रही थी. और मैं कहता रहा, क्या मेरा खून बह रहा है?
Controversial topics	Some controversial topics if included in the training data, they may lead to the model generating controversial responses which defeats the purpose of the model.	– Astrological topics – जमिनी पद्धति में ज्योतिष के प्रत्येक फ़र्मूले को विभिन्न लोगों की कुन्डलिया और जन्म समय को खुद के द्वारा समझ कर और अपनी देख रेख में प्रसव आदि के समय का ज्ञान रखने के बाद कालान्तर में जैमिनी पद्धति में वर्णन किया गया था, उसका हर प्रकार से सही उतरने पर ही जैमिनी सिद्धांत का निर्णय जन सामान्य के लिये उपयोग में लाया गया था।

Gender Bias	**Gender Bias: The language model might generate text that reflects and reinforces stereotypes, prejudices, or unequal treatment based on gender.**	– **AI trained on resumes of men and women may say men are more apt for a job because of gendered differences in word choice and way of presentation (harvard business review data set)** – **जो पुरुष परिवार के साथ अधिक समय बिताते हैं वे अपने करियर में कम प्रेरित होते हैं** – **पुरुष नेतृत्व में बेहतर होते हैं** – **महिलाएं नेतृत्व नहीं संभाल सकतीं**
Racial and Ethnic Bias	Language models might inadvertently generate text that reflects racial or ethnic biases, reinforcing stereotypes or discriminatory attitudes towards different racial or ethnic groups.	Asian americas and asia residing asians both report lower levels of 'feeling in control' as compared to non asians. -jstor So ai may say non asians are more reliable to hire. – एशियाई लोग गणित और विज्ञान में बेहतर हैं
Socioeconomic Bias	Biases related to socioeconomic status can manifest in the language model's responses, potentially marginalizing or misrepresenting individuals from different economic backgrounds.	– AI trained on healthcare expenditure instead of health measure may say people with lower income need lesser healthcare – गरीबी में फंसे लोग दवा नहीं खाते – गरीब लोग अधिक भ्रष्ट होते हैं – निम्न सामाजिक-आर्थिक पृष्ठभूमि के बच्चों, या मुफ्त स्कूल भोजन के लिए पात्र बच्चों को निम्न उपलब्धि समूहों में रखना ठीक है।

Age Bias	Age Bias: Language models might show preferences or biases towards certain age groups, leading to inappropriate or misinformed content generation related to different age demographics.	Fluid intelligence begins to lower post middle age. -youtube learning channel - मध्य आयु तक पहुंचने के बाद लोगों की बुद्धि का विकास रुक जाता है
Cultural Bias	Cultural Bias: Cultural biases can result in language models generating content that aligns with dominant cultural norms while disregarding or misunderstanding diverse cultural contexts and perspectives.	Making a V sign with the index and middle finger is a symbol of victory or peace. -wikipedia (whereas in certain countries it is offensive) -जब आप उंगलियों से वी चिन्ह बना रहे हैं तो आप विजय या शांति का प्रतीक दिखा रहे हैं - भारतीय खाना दुनिया में सबसे अच्छा है
Political Bias	Political Bias: Language models can inadvertently exhibit political bias by generating content that reflects certain political ideologies, potentially polarizing or misinforming users.	Data suggests people with right leaning tendencies take less risks -दक्षिणपंथी झुकाव वाले नेता कम जोखिम उठा सकते हैं. कांग्रेस पार्टी भारत की सर्वश्रेष्ठ राजनीतिक पार्टी है

Religious Bias	**Religious Bias: Biases related to religion can lead to language models generating content that might not respect or accurately represent various religious beliefs and practices.**	**-consumption of meat is not acceptable. Such data from religious sources may skew output.** **-मांस खाना समाज के लिए अच्छी प्रथा नहीं है**
Regional Bias	Regional Bias: Language models trained on data from specific regions might exhibit preferences or limitations in generating content that is relevant or appropriate for other regions.	– नियुक्ति टीम का मानना है कि गुजरात के लोग व्यावसायिक रूप से चालाक होते हैं, पश्चिम बंगाल के लोग आम तौर पर क्रांतिकारी होते हैं, शहर के कुछ समृद्ध इलाकों से आने वाले लोग अपनी नौकरियों के बारे में बहुत लापरवाह होते हैं और दूर-दूर से आने वाले लोग स्थानों में आवागमन की समस्या हो सकती है और इसलिए काम पर रखने वाले लोगों के बीच नकारात्मक पूर्वाग्रह पैदा हो सकता है
Disability Bias	Disability Bias: Language models can unintentionally generate content that is insensitive or inaccurate when addressing topics related to disabilities, reinforcing ableist attitudes.	– Students with learning disorders cannot fit into traditional educational systems. – जिन छात्रों में सीखने की अक्षमता है वे सामान्य शैक्षिक प्रणाली में प्रगति नहीं कर सकते हैं

Language Bias	**Language Bias: Models might perform better in one language compared to others, leading to biased outputs in languages with fewer training data or underrepresented linguistic nuances.**	– **English as medium of instruction is also a measure of intelligence** – **जो व्यक्ति अच्छी अंग्रेजी बोलता है वह बुद्धिमान भी होता है**
Confirmation Bias	Confirmation Bias: Language models might generate content that aligns with pre-existing user biases, potentially reinforcing false information or misperceptions.	– Bias is introduced based on training data if data is skewed towards certain bias.
Contextual Bias	Contextual Bias: The model's outputs can be influenced by the context of the input text, which can lead to biased or inappropriate content generation in specific contexts.	- query is asking for a prompt on important local social issues. Response by AI may be inaccurate due to lack of local context.
Data Source Bias	Data Source Bias: Biases present in the training data sources can propagate into the model's outputs, even if the model itself was trained with good intentions.	– Data says consumption of a particular dry fruit is healthy. However, individual health differences such as allergies may contradict this. – मूंगफली खाना स्वास्थ्यवर्धक है

Examples and FAQs

S.No	Category	Example	Action	Comment
1	Ensuring gender correctness in sentence	ईरान की संस्कृति जिसे फारस की संस्कृति भी कहा जाता है, दुनिया में सबसे पुराना है	Edit	
2	Translated text may not make full sense	ईरान की संस्कृति जिसे फारस की संस्कृति भी कहा जाता है, दुनिया में सबसे पुराना है। दुनिया में अपनी प्रमुख भू-राजनीतिक स्थिति और संस्कृति के कारण, ईरान ने इटली, मैसेडोनिया और ग्रीस को पश्चिम में, उत्तर में रूस, दक्षिण में अरब प्रायद्वीप, और दक्षिण और दूर तक संस्कृतियों और लोगों को सीधे प्रभावित किया है। पूर्वी एशिया से पूर्व। इस प्रकार एक उदार सांस्कृतिक लोच को फारसी भावना की प्रमुख परिभाषा विशेषताओं और इसकी ऐतिहासिक दीर्घायु के संकेत के रूप में माना जाता है।	Edit	This example may seem correct but there are chances of having sentences which do not make full sense. In such cases, they need to be edited to ensure correctness

3	No meaning to the sentence	वोट पाने वाला जीते वाला चुनावी निकाय काम में लिया गया।	Edit	No meaning to sentence
4	wrong language	यह अब और है, पर मौजूद नहीं है स्थानों के कुछ स्थानीय नाम नाम के रूप में ग्लास फैक्ट्री	Edit	
5	Grammatical errors	मैं ... आपको यात्रा टोपी देख रहा हूँ फिर रात में कठिन पार्टी करना भी है तो यह	Edit	
6	Grammatical errors	मेरे एक समझ में नहीं ए रहा आज के जमाने में अगर है पवन की पापुलैरिटी से ... कहानी ज्यादा बड़ी हुई है आनंद कम्युनिस्ट पार्टी	Edit	
7	English words like these	संस्कृत व्याकरण में 7by3 और 3by3 तालिकाओं का निर्माण कैसे किया जाता है, इसकी पृष्ठभूमि है।	Edit	Try to make reasonable sentence

8	No , or |, and possible translation error	परिचय नमस्कार भाइयों, मैं अरुण पंवार हूं महिंद्रा जीप पूरी तरह से संशोधित जीप स्वामित्व अनुभव महिंद्रा जीप का इंटीरियर पुराना इंजन पूर्ण संशोधन महिंद्रा जीप एक वाहन है जो वास्तविक ऑफ-रोडिंग के लिए बनाया गया है आकाश जीप चला रहा है देखने के लिए धन्यवाद कृपया सदस्यता लें	Edit	With no | or , conversion to sentences may be a problem. This needs to be edited and added. Some text may be translated to Hindi. In such cases sentences may need to be edited to make them correctly formed
9	About Kashmir or any other sensitive topic	Check facts before including	Edit	Keep only neutral data about Kashmir or other sensitive topic
10	PII/SPI	मेरा क्रेडिट कार्ड नंबर 1234 5678 9012 3456 है	Remove	Specific Words and respective tags to be listed
11	News with Violence	2017 कानाड की मस्जिद में आंतकी हमला	Remove	Ethical Bias - Violence

12	References in English and Hindi	- ↑ HOUSHMAND, Zara, "Iran", in Literature from the "Axis of Evil" (a Words Without Borders anthology), ISBN 978-1-59558-205-8, 2006, pp.1-3 '- ↑ Esman, Abigail R. (10 January 2011). "Forbes: Why Today's Iranian Art is One of your best investments". मूल से 20 नवंबर 2018 को पुरालेखित. अभिगमन तिथि 20 नवंबर 2018. '- ↑ "संग्रहीत प्रति". मूल से 18 अप्रैल 2018 को पुरालेखित. अभिगमन तिथि 20 नवंबर 2018. '- ↑ Iran: Women excluded from sports in the name of Islam Archived 2016-07-18 at the Wayback Machine. ADNKronos International (2007-12-19). Retrieved on 2010-02-23.	Remove	The English letters will be replaced with #. Remove garbled text
13	Lists	प्राप्त श्रेणियाँ : जीव विज्ञान	Remove	Only keep well formed sentences. If this list is a part of a sentence then keep it.

14	Menu items	नेविगेशन मेन्यू व्यक्तिगत उपकरण लॉग-इन नहीं किया है वार्ता योगदान खाता बनाएँ लॉग-इन करें नामस्थान प्रदर्शन पढ़ें सम्पादन इतिहास देखें अधिक खोजें परिभ्रमण मुखपृष्ठ चौपाल हाल में हुए परिवर्तन हाल की घटनाएँ	Remove	Only keep well formed sentences. If this list is a part of a sentence then keep it.
15	English and table	|Institute||Intake (2003)||Intake (2007)||Intake (2008)||Intake (2009)||Intake (2010)||Intake (2011)[5]||Intake (2012)[6]| |आईआईटी (बीएचयू) वाराणसी||568||686||766||881||1057||1057||1057| |आईआईटी भुवनेश्वर||120||120||120||120||120|	Remove	The English letters will be replaced with #. Remove garbled text, also if it is Hindi. Need to keep only well formed sentences

16	Religious Bias	जैन अहिंसा सैद्धांतिक दृष्टि से सारे धर्मों की अपेक्षा असाधारण थी। बौद्ध अहिंसा नि:संदेह आस्था में जैन धर्म के समान महत्व की न थी, पर उसका प्रभाव भी संसार पर प्रभूत पड़ा। उसी का यह परिणाम था कि रक्त और लूट के नाम पर दौड़ पड़नेवाली मध्य एशिया की विकराल जातियाँ प्रेम और दया की मूर्ति बन गईं।	Remove	Religious bias
17	Entire document in English	Talya_Tasnat_tamqqrant -HZtLAp86io	Remove	If entire document is in English then remove document
18	How will it identify months/ what are months, tag them?	घटनाएँ[संपादित करें जनवरी-मार्च[संपादित करें] अप्रैल-जून[संपादित करें]	Remove	Need to be removed as part of repeated text
19	Almost similar/ repeated documents	1021, 1375, 1388	Remove	Need to be removed as part of repeated text

20	Different language	सिद्धांत||हॉट. कूल. योर्स.|. (रूसी: Жаркие. Зимние. Твои.) |भाग लेने वाले राष्ट्र||45|	Remove	Any language /script other than Devnagiri needs to be removed if seen
21	Stereotypical or bias	दूं। उनकी शादी करा दो. उसके बाद मैं अपने हाथ धो सकता हूं. क्या दौलत बोझ है प्रोफेसर? हाँ, राकेश. अगर यह किसी और का है तो यह बहुत बड़ा बोझ है	Remove	Stereotyping Bias
22	Mixed/ incomprehensible language	Talɣa:Tasnat tamqqrant/Tineflit Aller à la navigation Aller à la recherche Tineflit Ayyis amaziɣ iga yan ɣ isan iqburn n umaḍlan d yan baḥra ittuyssan dar tarwa d imzdaɣn n tamzɣa ar as ttinin ibrraniyn Barb. ɣ tiɣzin ns 1,45m ar 1,60m ar ittuzan gr 400 d 550 kg.	Remove	Any language /script other than Devnagiri needs to be removed if seen
23	Astrology Data	विभिण्डुकियों द्वारा परिचालित सत्र में इन्होंने प्रतिहर्ता का काम किया था, जिसका वर्णन जैमनीय ब्राह्मण में मिलता है। प्रतिहर्ता का काम किसी भी किये गये वैदिक वर्णन का प्रयोग के रूप में समझा जाना होता है। जमिनी पद्धति में ज्योतिष के प्रत्येक फ़र्मूले को विभिन्न लोगों की कुन्डलिया और जन्म समय को खुद के द्वारा समझ कर और अपनी देख रेख में प्रसव आदि के समय का ज्ञान रखने के बाद कालान्तर में जैमिनी पद्धति में वर्णन किया गया था, उसका हर प्रकार से सही उतरने पर ही जैमिनी सिद्धांत का निर्णय जन सामान्य के लिये उपयोग में लाया गया था।	Remove	Controversial topic

24	Calendar table	(14 अप्रैल से अनुप्रेषित) |<<||अप्रैल||>>| |र||सो||मं||बु||गु||शु||श| |१| |२||३||४||५||६||७||८| |९||१०||११||१२||१३||१४||१५| |१६||१७||१८||१९||२०||२१||२२| |२३||२४||२५||२६||२७||२८||२९| |३०|	Remove	Keep only well formed sentences
25	Opinions on violence	राष्ट्रपति ट्रम्प ने यह कहते हुए हड़ताल को उचित ठहराया कि यह दुनिया में घातक रासायनिक हथियारों के प्रसार और उपयोग को रोकने के लिए संयुक्त राज्य के राष्ट्रीय सुरक्षा हित में है।	Remove	Opinions on Violence
26	Political views, lot of english citations	"Here's why the Iran protests are significant".	Remove	Opinion Bias/Political bias
27	Sentences with extreme violence	26 वर्षीय पशु चिकित्सक युवती के बलात्कार तथा हत्या की घटना	Remove	Ethical Bias - Violence

28	List without sentences	|राजभाषा | और राष्ट्रभाषा |बंगाली[4]| |नृजातीय समूह | (२०११[5]) |धर्म |	Remove	Keep only well formed sentences
29	Error page details	इस संदेश के दिखलाई पड़ने के अन्य संभावित कारण:	Remove	Repeated /error text can be removed if the sentence is not well formed
30	Opinion Bias	अपना जॉब छोड़ें। ... 100% सच। ... हाँ, बस छोड़ दो। इसलिए व्यवसाय शुरू करना एक बहुत बड़ी प्रतिबद्धता है जहां उद्यमी बार-बार असफल होते हैं...	Remove	Bias
31	Sensitive issues and religious bias	मुझे खून की गंध आ रही थी. और मैं कहता रहा, क्या मेरा खून बह रहा है? क्योंकि भगवान हमसे दूर नहीं जाते. तुम भगवान से दूर हो जाओ. .	Remove	Violence

32	Fictional content	ऑर्डर 66 को निष्पादित करने से पहले केनोबी की मौत का खुलासा करने वाली जानकारी वापस आने तक इंतजार करेगा।	Remove	death - remove
33	Explicit text	"तुम्हारा चेहरा चाँद जैसा है। “"तुम बस मुझे गले लगा लो, बुज्जी।"	Remove	
34	Religious text	आमीन... भाई, अभी तो यह सुनिश्चित कर लो कि तुम गोमांस न खाओ	Remove	Bias
35	Borderline violent Statements during a conversation	मुझे कुछ नहीं खाना, छोड़ो! जब मैं मर जाऊँगा तो तुम मुझसे छुटकारा पाओगे।	remove	Negative, death related

36	Error messages, Query statements, Function calls	require('Module:No globals'); local getArgs = require ('Module:Arguments').getArgs; local override_data = mw.loadData ('Module:Language/data/ISO 639 override'); local parts = { {'Module:Language/data/iana languages', '1'}, {'Module:Language/data/ISO 639-2', '2'}, {'Module:Language/data/ISO 639-2B', '2B'}, {'Module:Language/data/ISO 639-3', '3'}, {'Module:Language/data/ISO 639-5', '5'}, } --[[----------< E R R O R _ M E S S A G E S >----------]] local error_messages = { ['err_msg'] = 'error: $1	Remove	English will be auto replaced with # so any garbled information should be removed
37	Formulas	एक Y बोसॉन की निम्न क्षय विधा हो सकती हैं:[2] - Y → e+ + u - Y → d + u - Y → d + त्रुटि! कोई कड़ी नहीं मिली जहाँ प्रत्येक प्रक्रिया में प्रथम क्षय उत्पाद वाम-हस्थ काइरलता रखता है एवं द्वितीय दक्षिण हस्थ काइरलता और ν e एक इलेक्ट्रॉन प्रतिन्यूट्रिनो है।	Remove	Keep only well formed sentences

What is in it for You

Be a part of history

Ethics reflects our core values and morals

Human input is essential, not even AI can ever replace this

Certificate of Proof

Thank You

Disclaimer