- Perguntas de pesquisa:

- **Que tipos de discursos sobre a mudança climática são mobilizados e circulam em diversos registros na internet, por diferentes atores sociais, contemporaneamente?**

- **Em que medida os discursos sobre a mudança climática identificados em dimensões discursivas nos registros pesquisados distinguem a produção textual humana de textos correspondentes gerados por modelos de linguagem de grande porte?**

- **Que consonâncias (alinhamentos, reforços semânticos ou estabilização de sentidos) e que tensões (divergências, deslocamentos discursivos ou resistências) emergem da comparação entre os discursos produzidos por atores humanos e aqueles gerados por LLMs?**

- Corpus

Subcorpus Humano

1) Discursos Governamentais

- Decisões COP Mudança Climática da ONU (registro: ata)
- Relatórios IPCC (registro: relatório)

2) Discursos da Imprensa

- Now (registro: reportagem)

3) Discursos Não-governamental

- ONGs – WWF, WRI, Greenpeace (registro: stories)

4) Discursos das redes sociais

- Gettr (registro: postagem de rede social)

5) Discursos Acadêmicos

- Antconcgen (registro: resumos de artigos científicos)

Subcorpus LLMs

Textos gerados a partir de prompts:
a) ChatGPT
b) Gemini
c) Grok

Cada um dos LLMs irá produzir textos a partir do espelhamento dos seguintes registros, discursos:

Governamentais

1) Ata (decisões COPs)
2) Relatórios (IPCC)

Imprensa

3) Reportagem (NOW)

Não-governamental

4) Stories (WWF)
5) Stories (WRI)
6) Stories (Greenpeace)*

*Stories aqui apresentado como um híbrido de blogpost e notícia, é um registro que se encaixa tanto em blogpost quanto news

Sociedade

7) Postagem de rede social (Gettr)

Acadêmicos

8) Resumos de artigos científicos (AntCorGen)

Serão, ao todo, 8 registros diferentes. Conversamos sobre a questão da quantidade para uma análise fatorial mais estável, talvez coletarmos:

Subcorpus humano:

201 Ata (COP)
201 Relatório (IPCC)
603 Stories (201 Greenpeace; 201 WRI; 201 WWF)

201 Abstracts (artigos acadêmicos)

201 Postagens rede social (Gettr)

TOTAL: 1407 textos

*múltiplo de 3, já que serão 3 LLMs

Subcorpus LLMs:
201 Relatórios (divididos entre as 3 LLMs: 67 GPT, 67 Grok, 67 Gemini)

603 Stories (201 Greenpeace [67 GPT, 67 Grok, 67 Gemini]; 201 WRI [67 GPT, 67 Grok, 67 Gemini]; 201 [WWF 67 GPT, 67 Grok, 67 Gemini])

201 Abstracts (divididos entre as 3 LLMs: 67 GPT, 67 Grok, 67 Gemini)

201 Postagens rede social (divididos entre as 3 LLMs: 67 GPT, 67 Grok, 67 Gemini)

TOTAL: 1407 (469 GPT, 469 Grok, 469 Gemini)

Para coletarmos o corpus no News, Gettr e Anticoncgen:

Primeira possibilidade:

- Usar lista de palavras mais frequentes retiradas dos The Synthesis Report do IPCC*

*Since the IPCC was created in 1988, there have been 6 Synthesis Reports:

- The Overview of the First Assessment Report (1990)
- The IPCC Second Assessment Report Synthesis of Scientific-technical Information Relevant to Interpreting Article 2 of the UNFCCC (1995)
- The Synthesis Report of the Third Assessment Report (2001)
- The Synthesis Report of the Fourth Assessment Report (2007)
- The Synthesis Report of the Fifth Assessment Report (2014)
- The Synthesis Report of the Sixth Assessment Report (2023)

The AR Synthesis Report is based on the three Working Group contributions to the Assessment Report as well as on the three Special Reports prepared for each assessment cycle. The Synthesis Report Assessment Report provides an overview of the state of knowledge on the science of climate change, emphasizing new results since the others Assessment Report. It is fully based on the reports of the three Working Groups of the IPCC (WG I: bases físicas do clima; WG II: impactos, adaptação e vulnerabilidade; WG III: mitigação), as well as on the three Special Reports. It provides an integrated view of climate change as the final part of the Assessment Report. The Synthesis Report consists of a Summary for Policymakers and a Longer Report. It is published as a stand-alone publication.

<u>NÃO INCLUIR PARA NA LISTA DE PALAVRAS FREQUENTES (para Rogério)</u>

Capa

Contracapa

Dados de impressão do relatório (ISBN etc)

Sumário

Dedicatória

Tabelas com imagens

Tabelas de gráficos

Tabelas de números, MAS INCLUIR TABELAS COM PALAVRAS

Imagem

Referências bibliográficas

Segunda possibilidade:

Outra possibilidade seria pegarmos o glossário online dos relatórios do IPCC.
Exemplo de palavras:

Ablation (of glaciers, ice sheets, or snow cover)

Abrupt change

Abrupt climate change

Acceptability of policy or system change

Access (to food)

Access to modern energy services

Acclimatisation

Accumulation (of glaciers, ice sheets or snow cover)

Active layer

Acute food insecurity

Adaptation

Adaptation behaviour

Adaptation deficit

Adaptation Fund

Adaptation gap

Adaptation limits

Adaptation needs

Adaptation opportunity

Adaptation options

Adaptation pathways

Adaptive capacity

Adaptive governance

Adaptive management

Added value

Additionality

Adjustments (in relation to effective radiative forcing)

Advection

Adverse side-effect

Aerosol

Aerosol effective radiative forcing (ERFari+aci)

Aerosol optical depth (AOD)

Aerosol–cloud interaction

Aerosol–radiation interaction

Afforestation

Agreement

Agricultural and ecological drought

Agriculture, Forestry and Other Land Use (AFOLU)

Agroecology

Agroforestry

Air mass

Air pollution

Airborne fraction

Albedo

Alkalinity

Altimetry

Annular modes

Anomaly

Antarctic Ice Sheet (AIS)

Anthropocene

Anthropogenic

Anthropogenic emissions

Anthropogenic removals

Anthropogenic subsidence

Apparent hydrological sensitivity (ηa)

Arctic oscillation (AO)

Arid zone

Aridity

Artificial ocean upwelling (AOUpw)

Assets

Atlantic Meridional Mode (AMM)

Atlantic Meridional Overturning Circulation (AMOC)

Atlantic Multi-decadal Oscillation (AMO)

Atlantic Multi-decadal Variability (AMV)

Atlantic Zonal Mode (AZM)

Atmosphere

Atmospheric boundary layer

Atmospheric rivers (ARs)

Attribution

Australian and Maritime Continent monsoon (AusMCM)

Autonomous adaptation

Autotrophic respiration

Avalanche

Avoid, Shift, Improve (ASI)

Dúvidas:

1 – Sobre a lista de palavras

a) Temos a lista gratuita do glossário do IPCC, ou podemos fazer uma lista de frequência pelos Relatórios síntese. Gostaria de sua opinião sobre a questão para poder organizar com o Rogério, caso optemos por lista de palavras dos relatórios irá gerar um custo.

2 – Quanto aos tamanhos

a) Sobre a quantidade de textos: seriam 201 textos para cada Discurso-Registro (gov, ong, imprensa, rede social, academia)? E seria possível em vez de produzirmos 200 textos de cada discurso espelhado para cada LLMs (GPT, Grok, Gemini), dividirmos em 3 (200/3)?
Exemplo:
201 textos de imprensa – 67 do GPT, 67 Grok, 67 Gemini

b) Qual seria uma justificativa plausível, no domínio, para o corte nesse tamanho (201 textos de cada e não 400, por exemplo)*?
*Por que pensamos nesses números? Por uma questão de custo, para diminuir esses custos sem comprometer a estabilidade do estudo para a análise fatorial

c) Ainda no caso do tamanho dos textos, caso a gente decida por 201, no caso dos Discursos das ONGs, seriam 67 para cada ONG ou 201 para cada? Porque se fizermos 67 para cada ONG não será possível uma Anova, por exemplo, mas a comparação seria com o discurso ONG no geral, sem traçar uma comparação com cada uma das ONGs. E ainda assim, teríamos mais textos de ONGs do que dos outros atores sociais

3 – Sobre poucos textos, mas extensos

a) No subcorpus Governamental, tanto as atas de decisão das COPS quanto os relatórios do IPCC são muito grandes para que os LLMs reproduzam um texto espelhado. São páginas e mais páginas. Seria preciso espelhar RESUMOS feitos pelas próprias desses documentos e, a partir desse resumo, produzir o espelhamento. Mas aí não teríamos mais um espelhamento de registro e sim o espelhamento do RESUMO de um registro. O que acha disso?

b) Mais uma questão sobre o balanço do Corpus, enquanto a ONU tem muitos textos, o IPCC tem poucos (40 mais ou menos). São extensos, mas são poucos. Como poderíamos balancear esse desequilíbrio para a análise fatorial?

c) A sugestão seria pegarmos, então, material de divulgação de imprensa oficial da ONU tanto da COP quanto do IPCC por serem textos menores e com maior chance de serem reproduzidos por LLMs sem alucinação. A questão é que o IPCC não tem textos de divulgação suficiente para 402 textos. Outra solução seria, como na tese do Carlos, tentarmos fracionar os relatórios.

4 – Sobre os prompts

a) Vamos pedir para as LLMs espelharem os textos humanos (podemos chamar esse prompt de persona ou não) que vamos fornecer. Acha que seria necessária mais uma categoria de prompt?

b) Como seria, no caso de mais um prompt (default)? Apresentaríamos algum texto também, ou apenas damos o input do ator social, registro e tema, sem nenhum texto para que ele espelhe? Temos os estudos dos discursos presidenciais americanos, sobre os professores, sobre relatos da Covid, mas todos eles têm turnos, o que é diferente do meu. Será que no seu primeiro estudo comparando AI e textos humanos (2024) temos um prompt que posso nos ajudar como guia, ainda que tenha feito?

c) Se optarmos por duas categorias de prompts (espelhamento de textos e default), isso deve afetar os cálculos, dobrar o corpus, mais ou menos. pergunte se eles. Acha necessário para mais uma categoria de análise?