FSU | FLORIDA STATE UNIVERSITY

FSU UNDERGRADUATE RESEARCH OPPORTUNITY PROGRAM
CENTER FOR UNDERGRADUATE RESEARCH & ACADEMIC ENGAGEMENT

The Impact of LLM's on Spoken Language

Bryce Anderson and Dr. Tom Juzek
Department of Modern Languages and Linguistics

Abstract:

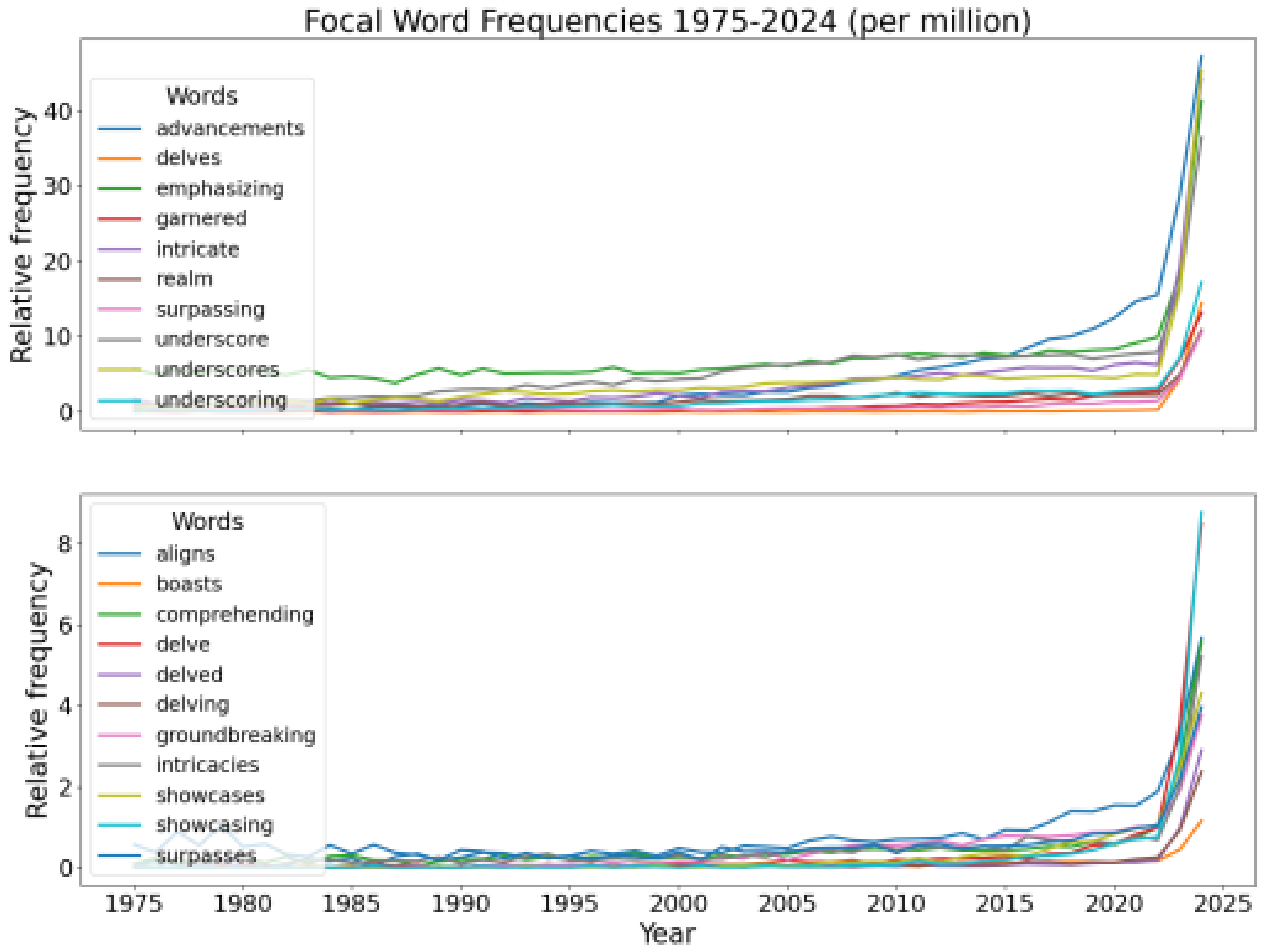
AI favored words have shown a rapid increase in frequency in Scientific English since 2022. Furthermore, the natural progression of language change has recently experienced the phenomenon of accelerated language change, where the introduction of Large Language Models (LLM's) may have caused deviations from expected linguistic results. This has been observed in scripted spoken language, but has yet to be observed in spontaneous spoken language. This gap in research on AI influenced spontaneous spoken language is substantial. This study aims to bridge the gap and observe the impact of Large Language Models on spontaneous speech. We systematically adhere to previously defined methods to characterize potential linguistic changes. Our approach involves transcribing 600 hours of spontaneous scientific podcast interviews recorded before and after the release of ChatGPT. Analysis of these transcriptions involves part of speech tagging to quantify frequency change, which are then tested for significant deviation from expected progression. We compare words from the transcripts that show a significant spike in frequency from 2020 to 2025 with the frequency trends of lexical items being overrepresented by LLM's, in order to identify an overlap. Preliminary results are expected to be incomplete, with the potential to identify the early onset of language change in spontaneous speech. Further data on everyday spontaneous spoken language is necessary to produce more robust results. With technology dominating aspects of everyday life, it is important to understand the impact these technologies could have on human language usage and communication.

Introduction:

Research Question: Does AI Usage Influence Spontaneous Spoken Language?

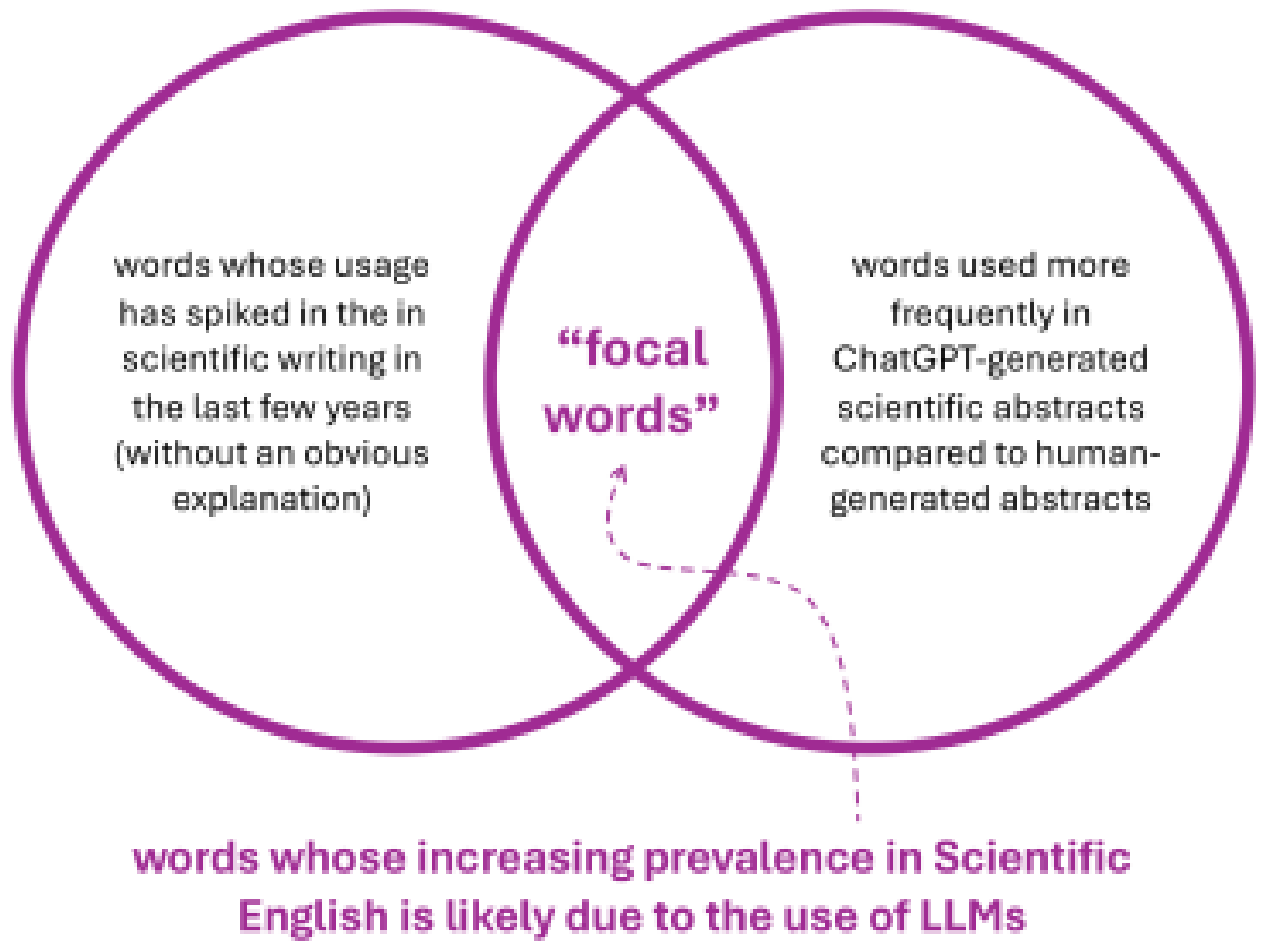
- Language change is complex and propelled by a combination of societal, cultural, and cognitive factors (Li et al, 2023; Zhang, 2014; Amato et al, 2018).
- When the printing press was introduced it expanded access to written literature, developed standardization for written language at the time, and altered the exchange of communication globally (Eisenstein, 1980).
- Similarly, the introduction of the telephone, and later on, cell phones revolutionized how humans speak and interact with each other (Sayers, 2014; Al-Sharqi et al, 2020).
- Some commonly recognized examples include abbreviated internet slang like "LOL", emoticons, hashtags, and textese (i.e. "gr8" instead of "great") (Djalolovna, 2024; Crystal, 2008).
- Societal behaviours reflected these technological changes as well, and neologisms were introduced such as "selfie" or "meme" (Crystal, 2008).
- Prior research has not only recognized the widespread use of LLM generation in academic writing, but have identified a list of focal words being overused by LLM's. This same list of focal words mirrors large frequency spikes of some of the same words in Scientific English abstracts (pre/post 2022) (Juzek, Ward, 2024).
- Existing literature identifies a connection between LLM's and scripted spoken language (Yakura et al, 2024; Geng et al, 2024).
- This is where the current research plans to provide a greater understanding of the impact AI has on spontaneous spoken language. We develop the question, will transcriptions of Scientific Podcasts before 2022, and after 2022, contain any words that show a significant increase in frequency when compared, and overlap with the previously determined list of overused AI words

Word Frequency Graph



(Juzek, Ward, 2024. *Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models*)

Focal Word Chart


words whose usage has spiked in the in scientific writing in the last few years (without an obvious explanation)
"focal words"
words used more frequently in ChatGPT-generated scientific abstracts compared to human-generated abstracts
words whose increasing prevalence in Scientific English is likely due to the use of LLMs

(Juzek, Ward, 2024. *Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models*)

Sources:

SCAN ME

Methods

Our research follows a previously defined (Juzek, Ward, 2024), and reproducible, process that involves:

- Identifying Scientific Podcasts that contain unscripted interviews with guest researchers, business owners, academics, and industry professionals.
- Download 600 hours worth of podcast material (300 from pre-22, 300 from post-22) and convert into MP3 format using Cobalt Tools.
- Using Automatic Speech Recognition (ASR) via OpenAI's speech-to-text Whisper model and Python to transcribe 600 hours worth of podcast MP3's.
- Once transcribed, remove metadata and format to be placed into a script that calculates the percentage increase for a token from pre-2022 to post-2022 (Juzek, Ward, 2024).
- Run a script to check the significance of the frequency increase via Python and a Chi-Square Significance formula (Juzek, Ward, 2024).
- If a word is significant and has an unexplainable reason for its increase, it can be identified as a focal word (Juzek, Ward, 2024).
- Graph word frequencies over a time period of pre-22 / post-22.
- Compare the newly collected list of focal words and compare to the previously defined list of focal words.
- Any overlap highlights potential AI influence on spontaneous spoken language.

Analysis and Results

- Research is still ongoing as of early March, but results are expected to be incomplete or potential indicators of language change. Both results would be interesting in determining the approach for future research into technology accelerating language change.

There are two main hypotheses that we can expect to see out of the results:

- Null hypothesis where zero overlap occurs between AI overused words and words with a significant increase in usage. This would then be attributed to factors out of scope of our research, such as being a feature of formal spoken language.
- Alternative hypothesis where there is an overlap with an unexplainable cause for increase, which could be attributed to LLM influence causing early onset language change. This result would show us that AI has potentially slipped into the human language system.