Eléanor LIARD (n°2205781)

Academic year 2024/2025

How ChatGPT is transforming academic research papers

CMI EFiQuaS scientific documentation project report, third year

Université Panthéon-Assas, Paris II
Under the direction of José De-Sousa and Pierre Deschamps

Abstract:

Launched in November 2022, ChatGPT is a powerful LLM tool that can assist in helping or even completely writing papers. The comprehension of its effects on academic research papers is crucial to help detect who is using it and how to use it in the most efficient way possible. To analyze its effect on academic research I [1] scraped papers on the SSRN website to gather a database, [2] analyzed its effect on all types of papers over time and then [3] analyzed its effect on papers based on their linguistic proximity to English. For all my computations, I used Python, and all my scripts are available on GitHub. I found that ChatGPT influences the difficulty of readability and requires a higher level in education to understand abstracts. It had a higher impact on countries with high linguistic distances to English.

Table of content

Table of content of figures

1. Introduction

According to Forbes [1], within five days of its launch, ChatGPT had attracted over one million users. It is a Large Language Model (LLM) developed by OpenAI, based on the Generative Pre-trained Transformer (GPT) architecture.
It answers questions, writes content, summarizes information, translates languages, and can assist with problem-solving, based on prompts. For example, it can answer the prompt: "give me concise bullet points for the introduction of my paper on ChatGPT" [2].

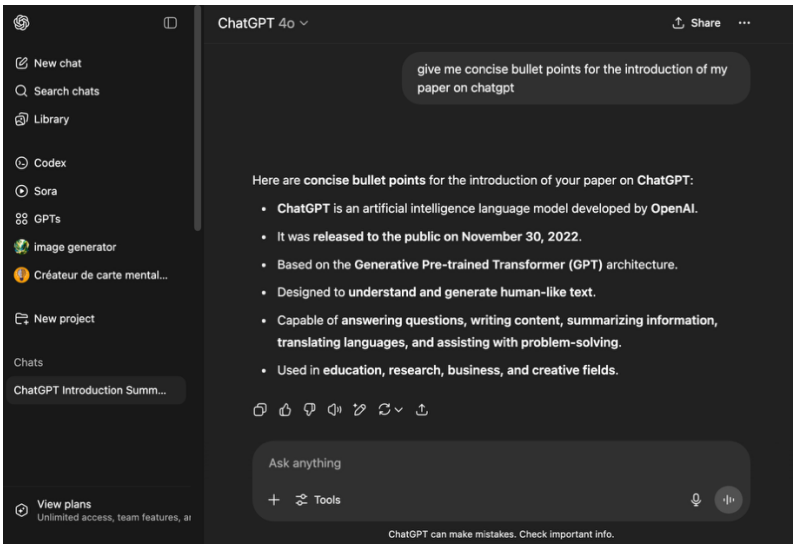

Figure 1: Screenshot of an example of prompt answered by ChatGPT

As it can generate, review, and improve text, it can be used in academic research. Of course, ChatGPT is not the only LLM with an impact on academic research but in this paper will be considered as the main representative of generative AI.

This paper raises a couple of questions concerning ChatGPT's impact on academic research:
- Does ChatGPT impact academic research?
- How does ChatGPT impact academic research?
- Whose academic research does ChatGPT impact?
- Whose academic research does ChatGPT impact the most?

To begin with, I reviewed a related article on readability. Then I performed an automated data retrieval of articles on the SSRN [3] website between 2015 and 2025 and computed measurements of readability. Finally, I analyzed my results: first studying the evolution of readability measurements over time for all papers and then focusing on the evolutions of those metrics depending on paper's linguistic distance with English.

As research papers can be very long and their content difficult to get ahold of, the articles are represented by their abstracts.

As the main part of this project was creating a database and computing measurements, I created a GitHub repository that contains all the code I used and the data I scraped and computed.

2. Literature review

The aim of the literature review was to find out about research into the readability of business articles and the methodology employed.

I read the article *Publishing while female* published in January 2022 by Erin Hengel [4]. It was about investigating the fact that women are underrepresented in top economics journals and whether they show higher writing standards.

In her research, Hengel followed several key steps:

- She defined how to identify an article written by a woman, ensuring clear criteria for sample selection.
- She built a database of article abstracts to serve as her textual data.
- She applied established readability formulas to quantify the clarity and accessibility of academic writing.
- She used Ordinary Least Squares (OLS) regression models to analyze whether readability systematically differed based on the author's gender.

I inspired myself of her work to use the abstracts as a source of data, her way of scraping and using data readability metrics as a proxy for writing clarity.

3. Automated data retrieval

The database was built using the SSRN API, specifically the search function, which allows retrieval of up to 10,000 articles per request.

The first and quite time-consuming task of the project was to create the database from 01/01/2015 to 06/07/2025, allowing the capture of the tendencies well before the launch of ChatGPT.
The database was built using the SSRN's API[1], specifically its research function, which allows retrieval of up to 10,000 articles per request.

Here is the link of the API:
https://api.ssrn.com/content/v1/bindings/205/papers/search?index=0&count=200&sort=0&term=labor

It depends on the parameters:
- `Term`: research of the articles with *term* in their title
- `Count`: number of articles asked
- `Index`: number of the first article in the list

I therefore had to create a list of words to look up and keep the ones from between 2015 and 2025.

3.1. List of keywords

The first step to creating the database was listing the keywords to be researched. The fuller the list, the more data could be retrieved and used for analysis.
To ensure the list of keywords was efficient and could capture as many articles as possible, I selected 10,000 articles and computed how many of those would have been scraped with my list of keywords and the research function of the API.

3.1.1. Creation of the list of keywords

The list of keywords was based on the possible categories of the JEL classification system [5].
In the list, the small words such as "in" or "the" were excluded because they would appear in almost all the most recent 10,000 articles and therefore would disable the intended filtering mechanism of the list. The combined words were also excluded from the list because the list treats them as two different individual words.
The issue with this strategy is that it does not allow to capture all the articles but does capture a great amount.

The first attempts of lists of keywords came from the general categories of the JEL classification system [5]:
- Titles of the categories
- Titles of the categories and economic words from the sub-titles
- Titles and sub-titles of the categories

3.1.2. Test of the list of keywords

To test the list of keywords, I looked at its efficiency at finding the titles in a list of 10,000 articles.

An issue arose when creating the 10,000 articles list: SSRN blocks itself when too many requests are asked. To avoid this blockage, I had to add a `time sleep` in the function that retrieved all the titles of the articles per index.

[1] API (Application Programming Interface): set of rules and tools that allows different software applications to communicate and exchange data with each other.

After having scraped all the articles with the list, I verified its success rate. At first it was very low, but then I modified my code to also create an Excel spreadsheet with the number of times words in the titles not collected occurred. That way, I added more words to my list and got a success result of 99.9%.

The code I used is: `creation_list.py`. This code and its final outputs are also in a file on the GitHub.

The script used for this task was `creation_list.py`, which performs the following steps:

- Capturing the 10,000 most recent articles published on the SSRN website and saving them in a spreadsheet `db_first_10000.xlsx`
- Counting and displaying the number of articles with at least a keyword in their title or not
- Saving the articles found in a spreadsheet `titles_found.xlsx`, and those not found in `titles_not_found.xlsx`
- Counting how many times each keyword is being used and saving it in `Nber_times_keywords.xlsx` and counting how many times each word in the non-found titles appear in `Nber_non_selected_words.xlsx`

The code was launched Saturday the 28 December 2024 and its script and final outputs are in the GitHub repository.

After having finalized the creation of the list of keywords, I had to collect the abstract and information of articles, then affiliate each article to a country and compute measurements to answer the problematic.

3.2. Creating the database

To organize and automatize my different scripts, I organized all of it in a folder called `abstract_analysis`:

```
ABSTRACT_ANALYSIS/
├── annexes/
│   └── creation_list/
│       ├── creation_list.py
│       ├── db_first_10000.xlsx
│       ├── Nber_non_selected_words.xlsx
│       ├── Nber_times_keywords.xlsx
│       ├── titles_found.xlsx
│       └── titles_not_found.xlsx
├── paper/
│   └── ChatGPT_report.pdf
├── data/
│   ├── ling_web.dta
│   ├── v1.67-2025-06-24-ror-data_schema.xlsx
│   └── v1.67-2025-06-24-ror-data.xlsx
├── outputs/
│   ├── graphs/
│   │   ├── all_papers/
│   │   │   ├── monthly_average_all_metrics.png
│   │   │   ├── ...
│   │   │   └── monthly_average_ttr.png
│   │   └── by_cle/
│   │       ├── comparison_monthly_average_automated_reading.png
│   │       ├── ...
│   │       └── comparison_monthly_average_ttr.png
│   ├── aff_1_author.xlsx
```

```
│   ├── affiliations_not_found_word_count.xlsx
│   ├── analysis_all_papers.xlsx
│   ├── analysis_by_cle.xlsx
│   ├── computations.xlsx
│   └── db_info_abstract.xlsx
├── scripts/
│   ├── __pycache__/
│   ├── aff_1_author.py
│   ├── analysis.py
│   ├── computations.py
│   └── info_abstract.py
├── main.py
├── .gitignore
├── README.md
└── requirements.txt
```

To run all computations at once, the `main.py` file must be run with all inputs declared. Here is what I had to fill to get my computations.

main.py > ...

```python
from scripts.info_abstract import info_abstract
from scripts.aff_1_author import aff_1_author
from scripts.computations import computations
from scripts.analysis import analysis


def main():
    output_info_abstract = 'outputs/db_info_abstract.csv'
    output_aff_1_author = 'outputs/aff_1_author.csv'
    output_computations = 'outputs/computations.csv'
    output_analysis = 'outputs/analysis.csv'
    begin_date = '2015-12-01' # 'YYYY--MM-DD'
    end_date = '2025-07-06'
    interest_name = 'ChatGPT'
    interest_year = 2022
    interest_month = 11
    interest_day = 30


    # Appel de la fonction principale
    info_abstract(output_info_abstract, begin_date, end_date)
    aff_1_author(output_info_abstract,output_aff_1_author)
    computations(output_aff_1_author,output_computations)
    analysis(output_computations,output_analysis,interest_name,interest_year,interest_month,interest_day)


if __name__ == "__main__":
    main()
```

Figure 2: `main.py` from GitHub repository

My script `main.py` calls 4 different functions and each one refer to one script.

3.2.1. Information and abstracts

The first script `info_abstract.py` gets all the data from the SSRN website.

I made a list of all the URLs with keywords that would give information about articles, scraped the article metadata in parallel and cleaned the results. Then I fetched the abstract for each article ID using another SSRN API in parallel and exported all the data into a CSV: `db_info_abstract.csv`

In this spreadsheet, each line represents an article and the columns information about the article:

- `Title`
- `Id`
- `Abstract_type`: type/category of the abstract (ex: working paper)
- `Publication_status`: Reference or citation string for the paper
- `Is_paid`: whether access to the paper requires payment (True/False)
- `Reference`: reference or citation string for the paper
- `Page_count`: number of pages
- `url`
- `affiliations`
- `is_approved`: whether the paper has been officially approved on SSRN
- `approved_date`: date when the paper was approved
- `Downloads`: total number of downloads of the paper
- `author_X_id`, `author_X_last_name`, `author_X_first_name`, `author_X_url`
- `paperDate`: date of publication
- `abstract`

As I knew we wanted to compare the different levels of English of the authors of the paper, I needed to affiliate each paper to a country.

3.2.2. Affiliation to a country

Affiliating each paper to a country was difficult and presented multiple obstacles.

The first one was finding separate affiliation: each paper had as many affiliations as authors and all the affiliations were in one cell. Separating each affiliation was too difficult because of their format with hyphens and comas.
After discussion with my teachers, we agreed to keep only articles with one author, as we would be confident about their affiliation.

The second one was linking each affiliation to a country. Finding a library or API with enough knowledge of all the affiliations was impossible and often gave bad results. Hence, I had to create my own library and used the ROR API [6].

My script `aff_1_author.py` keeps articles with only one author, extracts and cleans their affiliation. Then, based on a library of keywords it starts by trying to match them with a country and then in a second step for the unmatched ones applies fuzzy matching against the ROR database. It also computes the number of times each word in the unresolved affiliations appears, to enrich my personal library.

3.2.3. Computations

Computing measurements was easier. I got the proximity of each country to English (of the USA) and indices to measure the way the abstracts were written.

The script used is `computation.py`.

3.2.3.1. Language proximity

I used the data from Melitz, J. et Toubal, F (2014) Native Language, Spoken Language, Translation and Trade. Journal of International Economics, Vol. 92, N°2: 351-363 [7] to rank each country by its degree of proximity to the English language.

I also had to standardize the names of the countries, either because they were named differently in the database than in the data or because the country did not exist in the database, so I had to rename it with a similar country.

The index I selected is '`cle`', which refers to the aggregate index of common language [7]. It summarizes the evidence about the linguistic influences in an index resting strictly on exogenous linguistic factors.

It combines 3 components:

- CNL (common native language): probability that two people randomly drawn from each country share the same native language
- COL (common official language): binary variable equal to 1 if both countries share an official language, 0 otherwise.
- LP (linguistic proximity): continuous measure of how linguistically close two countries' main languages are, based either on language trees (LP1) or phonetic similarity of basic vocabulary (LP2).

$$\mathrm{CLE} = \mathrm{CNL} + (1 - \mathrm{CNL}) \times \frac{\mathrm{COL} + \mathrm{LP}}{\max(\mathrm{COL} + \mathrm{LP})}$$

This index provides a single, continuous measure of the linguistic ease of trade between countries, considering both: direct communication (through shared native or official language), and indirect communication (through linguistic similarity that makes translation easier).

3.2.3.2. Indices

I computed some measurements of readability myself:

- **Average word length** (`avg_length_words`)

$$\text{avg\_length\_words} = \frac{\text{total characters}}{\text{total words}}$$

- **Proportion of long sentences** (`prop_more_15words`): proportion of sentences exceeding 15 words

$$\text{prop\_more\_15words} = \frac{\text{number of sentences with more than 15 words}}{\text{total sentences}}$$

- **Type-Token Ratio** (`ttr`): unique number of words divided by total words, to measure the lexical diversity in a text

$$\text{ttr} = \frac{\text{Nnumber of unique words}}{\text{total words}}$$

I also computed measurement with the `textstat` library [8]:

- **Flesch reading ease** (`Flesch_reading_ease`): ease of reading score [9]

$$\text{Flesch\_reading\_ease} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Score	Difficulty
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Confusing

Figure 3: Table of Flesch reading ease score

- **Flesch-Kincaid grade level** (`fk_grade_level`): U.S. school grade level required to understand the text [10]

$$\text{fk\_grade\_level} = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

School level (US)	Notes
5th grade	Very easy to read. Easily understood by an average 11-year-old student.
6th grade	Easy to read. Conversational English for consumers.
7th grade	Fairly easy to read.
8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
10th to 12th grade	Fairly difficult to read.
College	Difficult to read.
College graduate	Very difficult to read. Best understood by university graduates.
Professional	Extremely difficult to read. Best understood by university graduates.

Figure 4: Table of Flesch-Kincaid grade level

- **Gunning Fog index** (`gunning_Fog`): estimated years of education needed to understand the text [11]

$$\text{gunning\_Fog} = 0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

Fog Index	Reading level by grade
17	College graduate
16	College senior
15	College junior
14	College sophomore
13	College freshman
12	High school senior
11	High school junior
10	High school sophomore
9	High school freshman
8	Eighth grade
7	Seventh grade
6	Sixth grade

Figure 5: Table of Fog index

- **SMOG index** (`smog`): estimated grade level based on complex words [12]

$$\text{smog} = 1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

- **Dale-Chall score** (`dale_chall`): readability score based on the use of familiar vs. difficult words [13]

$$\text{dale\_chall} = 0.1579\left(\frac{\text{difficult words}}{\text{words}} \times 100\right) + 0.0496\left(\frac{\text{words}}{\text{sentences}}\right)$$

Score	Understood by
4.9 or lower	average 4th-grade student or lower
5.0–5.9	average 5th or 6th-grade student
6.0–6.9	average 7th or 8th-grade student
7.0–7.9	average 9th or 10th-grade student
8.0–8.9	average 11th or 12th-grade student
9.0–9.9	average 13th to 15th-grade (college) student

Figure 6: Table of Dale-Chall score

- **Automated Readability index** (`automated_readility`): readability based on characters per word and words per sentence [14]

$$\text{automated\_readility} = 4.71\left(\frac{\text{characters}}{\text{words}}\right) + 0.5\left(\frac{\text{words}}{\text{sentences}}\right) - 21.43$$

Score	Age	Grade Level
1	5-6	Kindergarten
2	6-7	First Grade
3	7-8	Second Grade
4	8-9	Third Grade
5	9-10	Fourth Grade
6	10-11	Fifth Grade
7	11-12	Sixth Grade
8	12-13	Seventh Grade
9	13-14	Eighth Grade
10	14-15	Ninth Grade
11	15-16	Tenth Grade
12	16-17	Eleventh Grade
13	17-18	Twelfth Grade
14	18-22	College student

Figure 7: Table of Automated Readability index

3.2.4. Analysis

To analyze all computations in my `analysis.py` script, I computed the average of each metric on a quarterly basis and plotted them with the library matplotlib and computed descriptives statistics.
It also computes the mean of each metric before and after a date (here November 2022) and does a t-test for each one of them.
The final output is an Excel spreadsheet with the descriptive statistics of each computation and the result of their t-test.

The two-sample t-test is used to test whether the means of two independent groups are significantly different from each other [15].

The null and alternative hypotheses are:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

Where $\mu_1,\ \mu_2$ are the population means.

The first step of the test is to compute the t-statistic:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\overline{x_1}, \overline{x_2}$ are the sample means
- s_1^2, s_2^2 are the sample variances
- n_1, n_2 are the sample sizes

The second step of the test is to compute the degree of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

The last step if computing the p-value, which is the probability of observing a value as extreme as the one obtained if the null hypothesis were true, and then comparing the p-value and the significance level:

$$\text{If } p \leq 0.05 \quad \Rightarrow \quad \text{Reject } H_0$$

$$\text{If } p > 0.05 \quad \Rightarrow \quad \text{Do not reject } H_0$$

5. Analysis of all articles

The aim of this first was to see whether ChatGPT had an impact on academic research and if so, how?

The final database contained 80,667 articles, from 01 January 2025 to 06 July 2025.

I plotted the metrics computed and did t-test with the following hypothesis:

The null and alternative hypotheses are:

$$H_0: \mu_1 = \mu_2: the\ means\ of\ the\ metric\ are\ equal\ before\ and\ after\ ChatGPT's\ launch$$
$$H_1: \mu_1 \neq \mu_2: the\ means\ of\ the\ metric\ are\ different\ before\ and\ after\ ChatGPT's\ launch$$

metric	count	mean	std	min	25%	50%	75%	max	t_stat	p_value	mean_before	mean_after	significant
flesh_reading_ease	80667	29,3983	16,0307	-344	19,57	29,55	39,3	121,22	39,8172	0	30,7366	25,577	True
fk_grade_level	80667	15,2582	3,5755	-3,5	13,1	15	17,1	173,3	-19,8631	0	15,1153	15,6661	True
gunning_fog	80667	16,5525	3,685	0,4	14,27	16,21	18,42	180,84	-14,688	0	16,4432	16,8645	True
smog	80667	15,686	4,1638	0	14,6	16,2	17,9	34,7	-19,1919	0	15,5264	16,1416	True
dale_chall	80667	11,0105	1,3978	0,35	10,09	10,81	11,7	35,46	-21,269	0	10,951	11,1804	True
automated_readility	80667	18,2628	4,4928	-6,9	15,5	17,9	20,4	223,8	-18,6601	0	18,0926	18,7489	True
avg_length_words	80667	5,5132	0,4497	2,4225	5,228	5,4926	5,7795	8,4393	-33,3294	0	5,4809	5,6056	True
prop_more_15words	80667	0,7531	0,2153	0	0,6111	0,7778	1	1	2,3804	0,0173	0,7542	0,7501	True
ttr	80667	0,5839	0,0922	0,1768	0,5252	0,5801	0,637	1	0,5805	0,5615	0,584	0,5835	False
page_count	80667	30,2245	36,4055	-1	11	24	41	2802	-8,6359	0	29,5186	32,2398	True
cle	80667	0,5228	0,3893	0	0,1463	0,3647	1	1	7,7039	0	0,5291	0,5049	True

Figure 8: Table of statistics of measurements for all articles

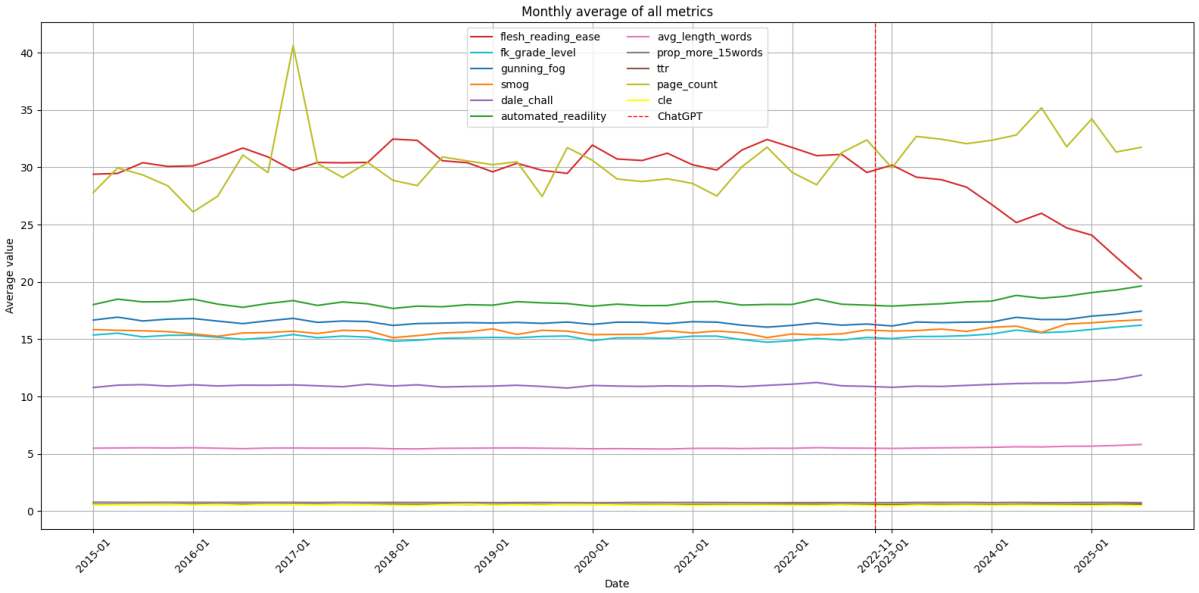

Figure 9: Graphical representation of measurements of readability over time

From a first glance, every curve except automated readability, TTR and CLE seem to vary.

5.1. Readability measurements

5.1.1. Flesch reading ease

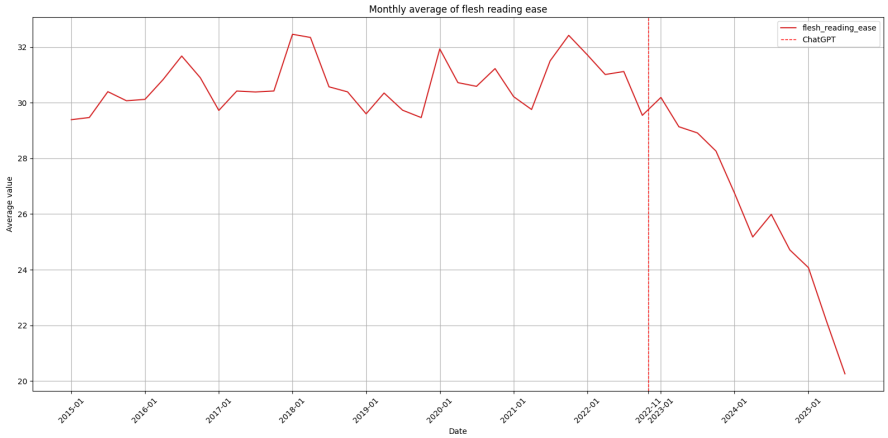

Figure 10: Graphical representation of Flesch reading ease over time

The value of Flesch reading ease, averaged over the period from 2015 to 2025, was 29,39, which corresponds to the level of reading very confusing.

Before ChatGPT, the average amounted to 30,73 and after ChatGPT to 25,57. The difficulty goes from difficult to very confusing. The t-stat is 39,81 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average reading ease before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in automated readability index, indicating abstracts of papers got more difficult to read.

5.1.2. Flesch-Kincaid Grade Level

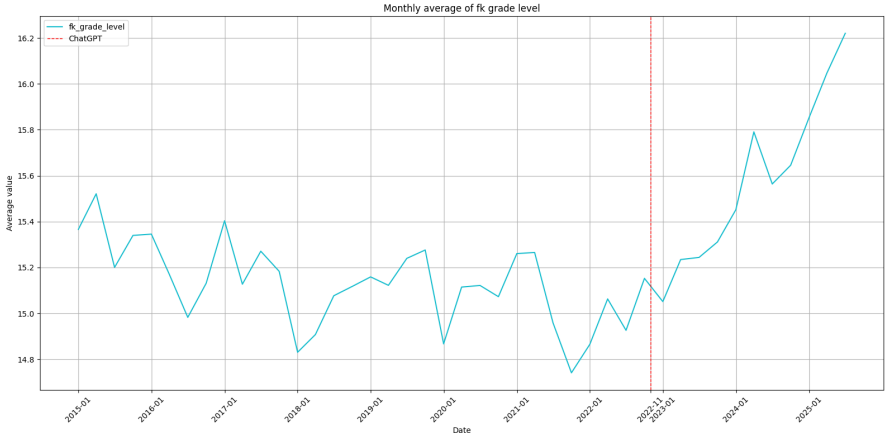

Figure 11: Graphical representation of Flesch-Kincaid grade level over time

The value of the Flesch-Kincaid grade level, averaged over the period from 2015 to 2025, was 15,25, which corresponds to the level of reading of a college student.

The t-stat is -19,86 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average reading ease before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in Flesch-Kincaid grade level, indicating readers of abstracts of papers need a higher level of education.

5.1.3. Gunning Fog

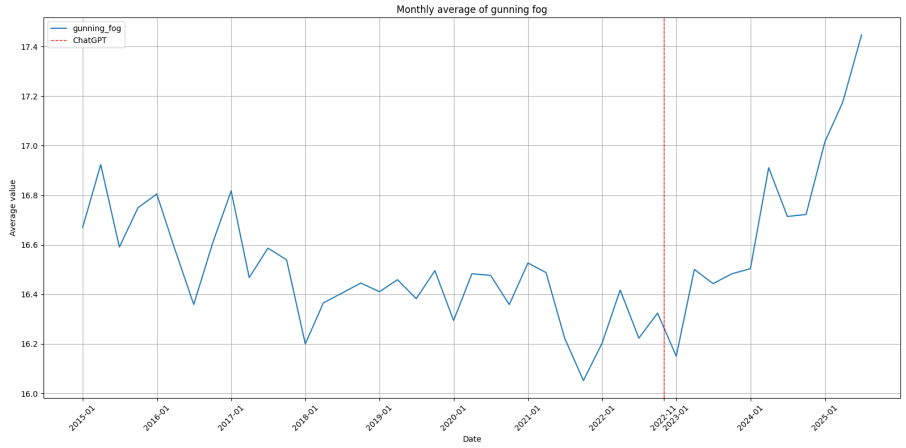

Figure 12: Graphical representation of Gunning Fog score over time

The value of the Gunning Fog index, averaged over the period from 2015 to 2025, was 16,55, which corresponds to the level of reading of a college senior.

Before ChatGPT, the average amounted to 16,44 and after to 16,86. The t-stat is -14,68 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average necessary level of reading before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the Gunning Fog readability index, indicating that reading the abstracts of papers required a higher educational level.

5.1.4. SMOG

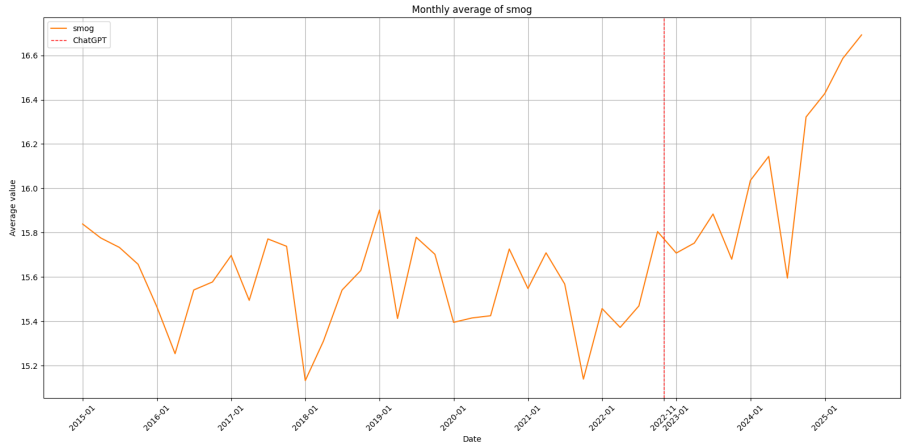

Figure 13: Graphical representation of SMOG index over time

The value of the SMOG index, averaged over the period from 2015 to 2025, was 15,68, which corresponds to the level of reading of a college senior.

Before ChatGPT, the average amounted to 15,52 and after to 16,14. The t-stat is -19,19 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average necessary level of reading before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the SMOG readability index, indicating that reading the abstracts of papers required a higher educational level.

5.1.5. Dale Chall

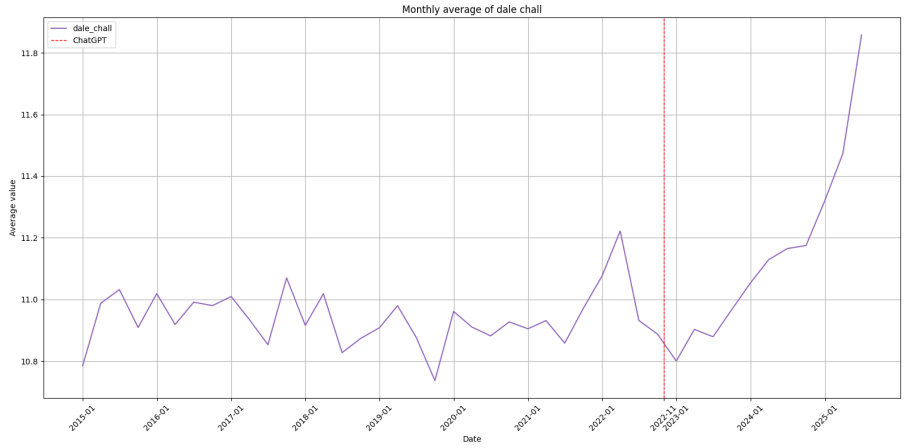

Figure 14: Graphical representation of Dale Chall score over time

The value of Dale-Chall, averaged over the period from 2015 to 2025, was 11,01, which corresponds to a level of reading above college.

Before ChatGPT, the average amounted to 10,95 and after to11,18. The t-stat is -21,26 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the level of reading necessary before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the Dale Chall index, indicating that reading the abstracts of papers required a higher educational level after the introduction of ChatGPT.

5.1.6. Automated readability

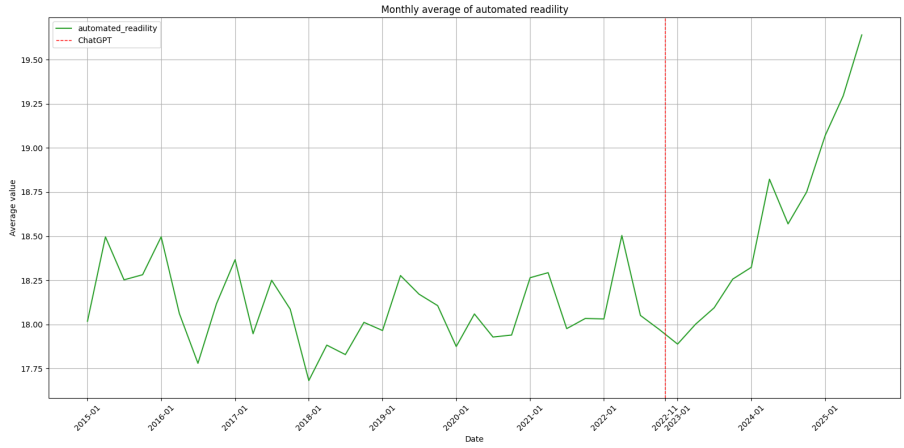

Figure 15: Graphical representation of automated readability over time

The value of automated readability, averaged over the period from 2015 to 2025, was 18,26, which corresponds to the level of reading of a college senior.

Before ChatGPT, the average amounted to 18,09 and after to 18,74. The t-stat is -18,66 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average automated readability before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in automated readability index, indicating that reading the abstracts of papers required a higher educational level.

5.2. Structural measurements

5.2.1. Average length words

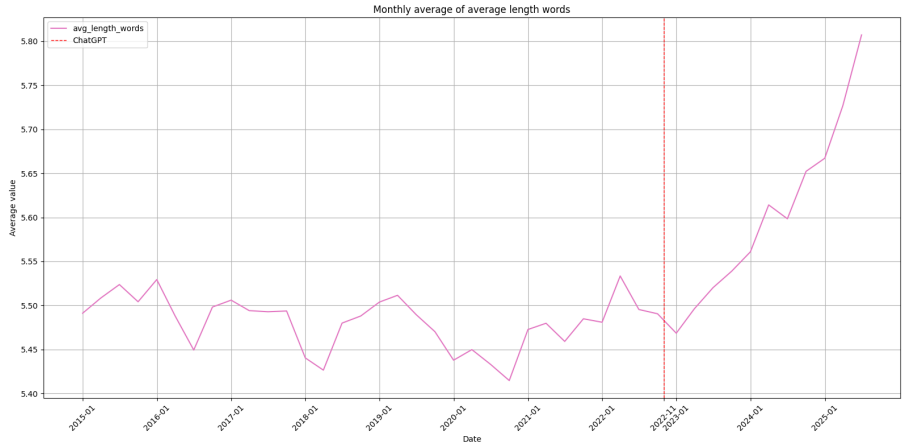

Figure 16: Graphical representation of average length of words over time

The value of average length words, averaged over the period from 2015 to 2025, was 5,51 characters per word.

Before ChatGPT, the average amounted to 5,48 and after to 5,60. The t-stat is -33,32 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the average length of words before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the average length of words, indicating that words in abstracts are getting longer.

5.2.2. Proportion of more than 15 words

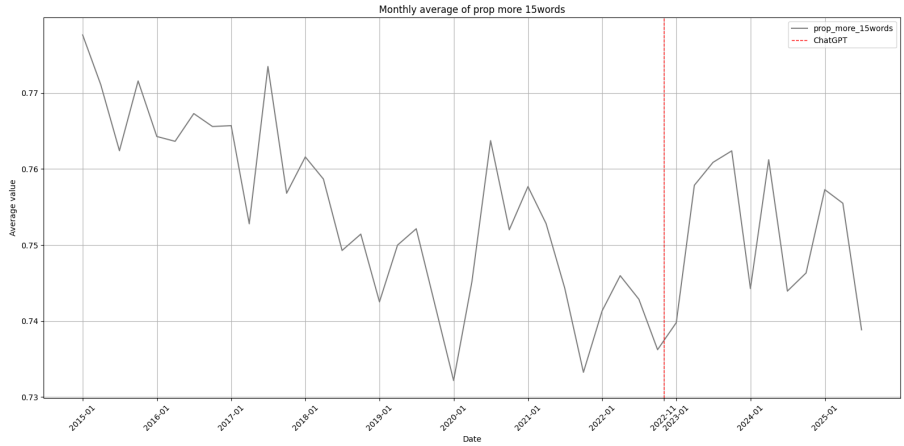

Figure 17: Graphical representation of proportion of more than 15 words per sentence over time

The value of the average of proportion of more than 15 words, averaged over the period from 2015 to 2025, was 75%.

Before ChatGPT, the average amounted to 75,42% and after to 75,01%. The t-stat is 2,38 and the p-value of 0,01. The difference is therefore statistically significant, and we can reject the null hypothesis that the proportion of sentences with more than 15 words before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the proportion of sentences with more than 15 words.

5.3. Other measurements

5.3.1. TTR (Type-Token Ratio)

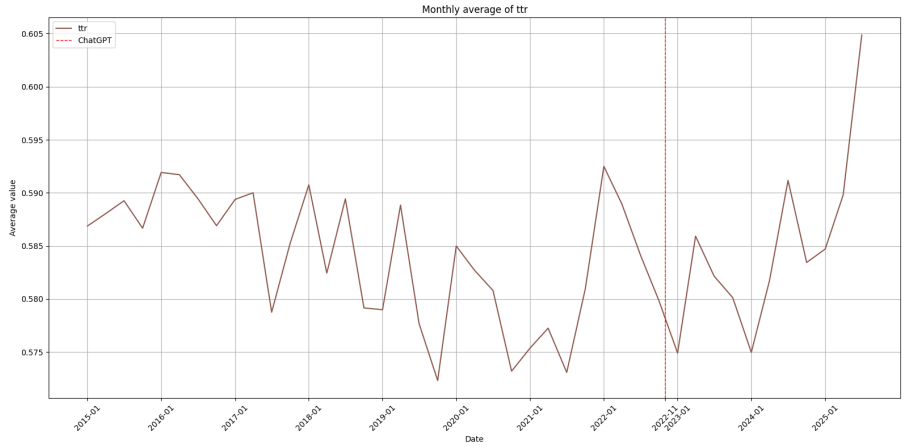

Figure 18: Graphical representation of TTR over time

The value of the TTR, averaged over the period from 2015 to 2025, was 0,58.

Before ChatGPT, the average amounted to 0,584 and after to 0,583. The t-stat is -21,26 and the p-value of 0,584. The difference is therefore **not statistically significant**, and we cannot reject the null hypothesis that the lexical diversity before and after ChatGPT is the same.

The launch of ChatGPT did not lead to a variation in the type-Token Ratio, indicating that the lexical diversity in abstract has not changed.

5.3.2. Page count

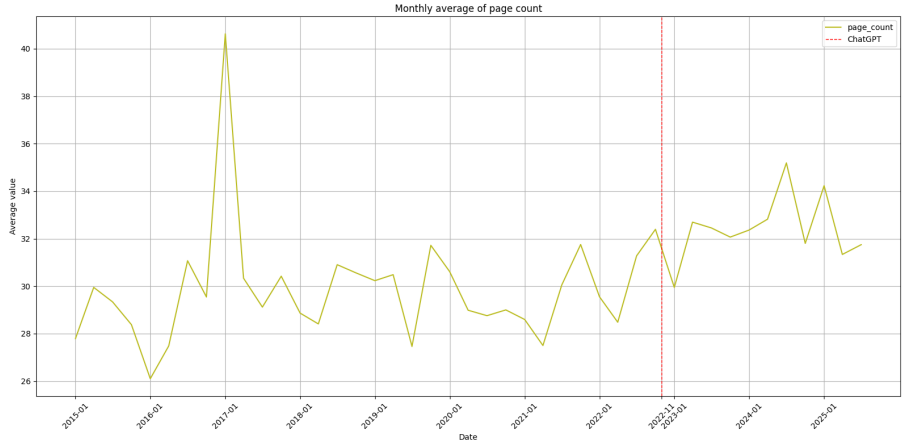

Figure 19: Graphical representation of number of pages over time

The value of the average of page count, averaged over the period from 2015 to 2025, was 30,22 pages per article.

Before ChatGPT, the average amounted to 29,51 and after to 32,23. The t-stat is -8,63 and the p-value of 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that the number of pages before and after ChatGPT is the same.

The launch of ChatGPT led to an increase in the number of pages per paper.

5.3.3. Common level index

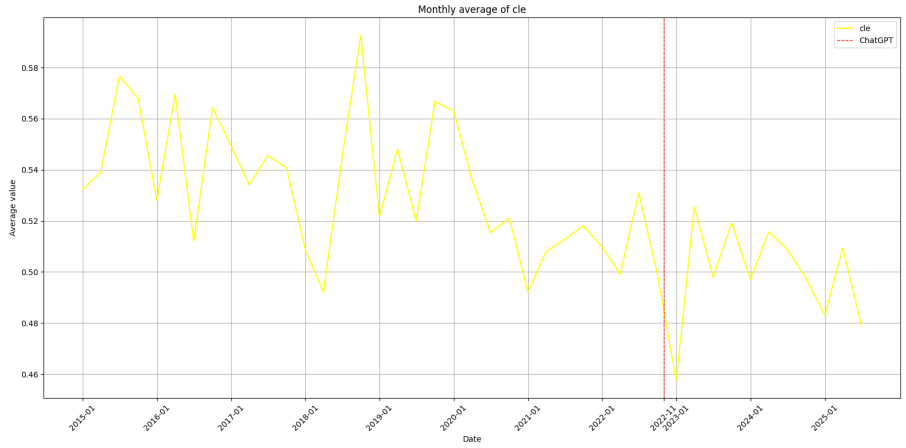

Figure 20: Graphical representation of the mean of CLE over time

The value of CLE, averaged over the period from 2015 to 2025, was 0,5228.

Before ChatGPT, the average amounted to 0,5291 and after to 0,5049. The t-stat is 7,7 and the p-value 0,00. The difference is therefore statistically significant, and we can reject the null hypothesis that average of CLE before and after ChatGPT is the same.

The launch of ChatGPT led to a decrease in the average of common level index, therefore in an increase of papers written in languages less close to English.

Computing these measurements by common level index are therefore interesting to pursue the analysis of the impact of ChatGPT on academic research.

7. Analysis by linguistic distance

To deepen the subject, I tested whether the launch of ChatGPT helped English or closely related language speakers or not.
The linguistic distance is measured by the common level index (CLE). As it is a continuous variable, for further analysis I used a discrete variable based on it that I computed in `computations.py`.

The initial `cle` variable had 92 values, the secondary `cle_bis` had 4 values:

Value in `cle`	Value in `cle_bis`	Count
[0.00; 0.25]	1	32,881
]0.25; 0.50]	2	10,045
]0.50; 0.75]	3	4,417
]0.75; 1.00]	4	33,324

The highest CLE is made of English-speaking countries (USA, UK, Ireland and New Zealand). It is the level of comparison for other countries that are considered as not primary English-speakers.

My goal for this part was to test two propositions:
- Whose academic research does ChatGPT impact?
- Whose academic research does ChatGPT impact the most?

To analyze these questions, I computed for each metric and each range of CLE:
- A t-test for each CLE range for each metric to test whether their average is the same before and after the launch of ChatGPT
- The percentage change between the date of the launch and the last day of the database to quantify the impact

The analysis comes from the output `analysis_by_cle.xlsx`.

metric	cle_group	mean_before	mean_after	t_stat_before_after	p_val_before_after	significant	%_change
flesh_reading_ease	0-0.25	32,4386	24,9405	35,3311	0	True	-23,11
flesh_reading_ease	0.25-0.50	29,021	27,0789	5,6101	0	True	-6,69
flesh_reading_ease	0.50-0.75	28,7788	24,4413	8,0818	0	True	-15,07
flesh_reading_ease	0.75-1.00	29,9033	25,9706	20,2479	0	True	-13,15
fk_grade_level	0-0.25	14,9203	15,808	-19,2386	0	True	5,95
fk_grade_level	0.25-0.50	15,3483	15,3779	-0,3989	0,69	False	0,19
fk_grade_level	0.50-0.75	15,5332	15,8931	-3,1767	0,0015	True	2,32
fk_grade_level	0.75-1.00	15,1734	15,5679	-9,6512	0	True	2,6
gunning_fog	0-0.25	16,2897	16,9982	-14,9647	0	True	4,35
gunning_fog	0.25-0.50	16,7095	16,5852	1,5978	0,1102	False	-0,74
gunning_fog	0.50-0.75	16,862	17,0116	-1,2641	0,2064	False	0,89
gunning_fog	0.75-1.00	16,4515	16,7826	-7,776	0	True	2,01
smog	0-0.25	15,3361	16,2499	-18,2264	0	True	5,96
smog	0.25-0.50	15,7712	15,9766	-2,3072	0,0211	True	1,3
smog	0.50-0.75	15,8585	16,1781	-2,2366	0,0254	True	2,02
smog	0.75-1.00	15,5881	16,0674	-9,6088	0	True	3,07
dale_chall	0-0.25	11,127	11,229	-5,6387	0	True	0,92
dale_chall	0.25-0.50	11,0281	11,117	-2,849	0,0044	True	0,81
dale_chall	0.50-0.75	10,7636	11,136	-8,8055	0	True	3,46
dale_chall	0.75-1.00	10,7851	11,1514	-23,6937	0	True	3,4
automated_readility	0-0.25	18,1861	18,8891	-11,9142	0	True	3,87
automated_readility	0.25-0.50	18,0761	18,3103	-2,4958	0,0126	True	1,3
automated_readility	0.50-0.75	18,2875	19,0691	-5,4469	0	True	4,27
automated_readility	0.75-1.00	17,982	18,6851	-13,6727	0	True	3,91
avg_length_words	0-0.25	5,4707	5,6192	-24,1103	0	True	2,71
avg_length_words	0.25-0.50	5,4793	5,5641	-8,696	0	True	1,55
avg_length_words	0.50-0.75	5,4892	5,6055	-7,3605	0	True	2,12
avg_length_words	0.75-1.00	5,49	5,6029	-20,1337	0	True	2,06
prop_more_15words	0-0.25	0,7538	0,7614	-2,8898	0,0039	True	1
prop_more_15words	0.25-0.50	0,757	0,746	2,2206	0,0264	True	-1,45
prop_more_15words	0.50-0.75	0,7842	0,761	3,0887	0,002	True	-2,96
prop_more_15words	0.75-1.00	0,7497	0,7377	4,3556	0	True	-1,6
ttr	0-0.25	0,5783	0,5747	3,2082	0,0013	True	-0,63
ttr	0.25-0.50	0,5899	0,5892	0,3745	0,708	False	-0,13
ttr	0.50-0.75	0,586	0,5855	0,1568	0,8754	False	-0,09
ttr	0.75-1.00	0,5872	0,5913	-3,5239	0,0004	True	0,69
page_count	0-0.25	24,6502	27,7567	-5,797	0	True	12,6
page_count	0.25-0.50	32,454	35,2586	-3,8627	0,0001	True	8,64
page_count	0.50-0.75	30,9516	38,3619	-3,5416	0,0004	True	23,94
page_count	0.75-1.00	33,069	35,4818	-5,926	0	True	7,3

Figure 21: Table results of CLE

7.1. Readability measurements

7.1.1. Flesch reading ease

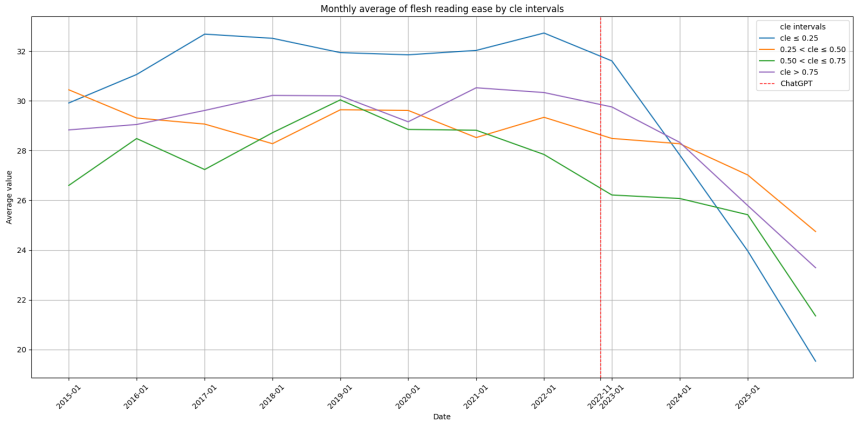

Figure 22: Graphical representation of Flesch reading ease score by CLE over time

Overall impact on all ranges:
There are statistically significant changes in Flesch reading ease before and after the launch of ChatGPT, as the p-values of all t-test are null. **All countries write abstracts that are harder to understand**.

Quantifying the impact:
On average, before the launch of ChatGPT, countries the furthest away from English-speaking had the easiest text difficulty and after, they were almost the one with the most text difficulty.
The countries that had **the biggest change in text difficulty are the countries the furthest away from English** (−23%).

7.1.2. Flesch-Kincaid grade level

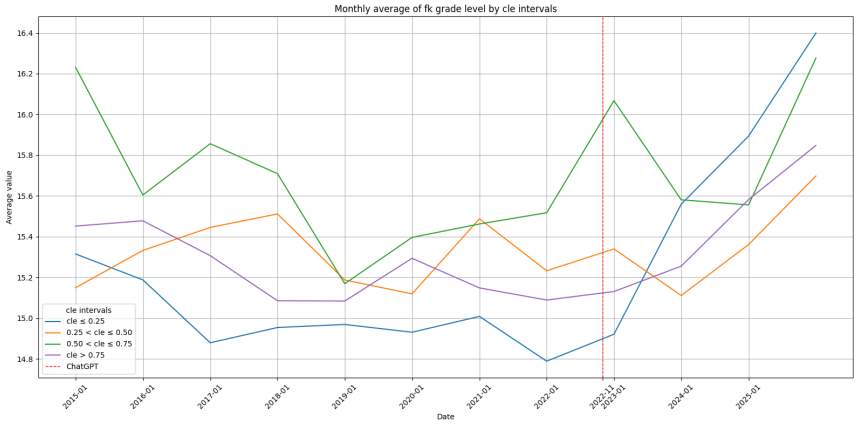

Figure 23: Graphical representation of Flesch-Kinkai grade level by CLE over time

Overall impact on all ranges:
There is no change in grade level requirement before and after the launch of ChatGPT for 0.25-0.5 CLE countries, as the p-values of all t-test are above 0.05. But for the others, there is statistically significant **increase in grade level requirement** to read abstracts.

Quantifying the impact:
On average, before the launch of ChatGPT, countries the furthest away from English-speaking had the least grade level requirement and after, they were the second least ones, before the ones far away from English (CLE of 0.25-0.5).
Based on the Flesch-Kinkai analysis, the countries that had **the biggest change in grade level are the countries the furthest away from English** (5.95%).

7.1.3. Gunning Fog

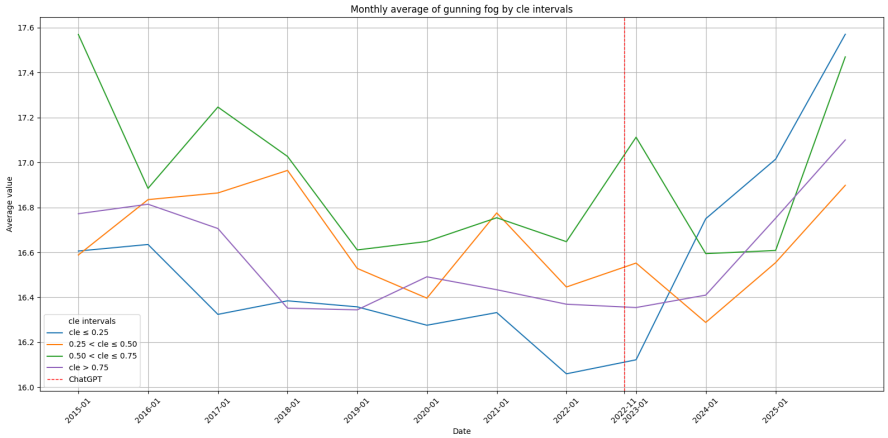

Figure 24: Graphical representation of Gunning Fog by CLE over time

Overall impact on all ranges:
There is no change in grade level requirement before and after the launch of ChatGPT for 0.25-0.75 CLE countries, as the p-values of all t-test are above 0.05. But for the **least close and lest furthest from English,** there is statistically significant **increase in grade level** requirement to read abstracts**.**

Quantifying the impact:
On average, before the launch of ChatGPT, countries the furthest away from English-speaking had the least grade level requirement and after, they were the second more demanding ones, after the native English-speaker's countries.
Based on the Gunning Fog analysis, the countries that had **the biggest change in grade level requirement are the countries the furthest away from English** (4.35%).

7.1.4. SMOG

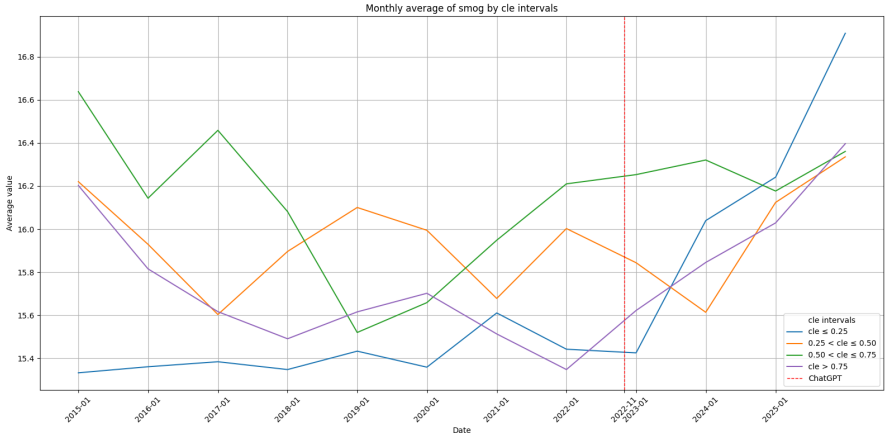

Figure 25: Graphical representation of SMOG by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are null**,** there are statistically significant **increase in grade level** requirement to read abstracts.

Quantifying the impact:
On average, before the launch of ChatGPT, countries the **furthest** away from English-speaking had the least grade level requirement (15.3) and after, they were the **most demanding** ones (16.24).
Based on the SMOG analysis, the countries that had **the biggest change in grade level requirement** are the countries the **furthest away from English** (5.96%).

7.1.5. Dale Chall

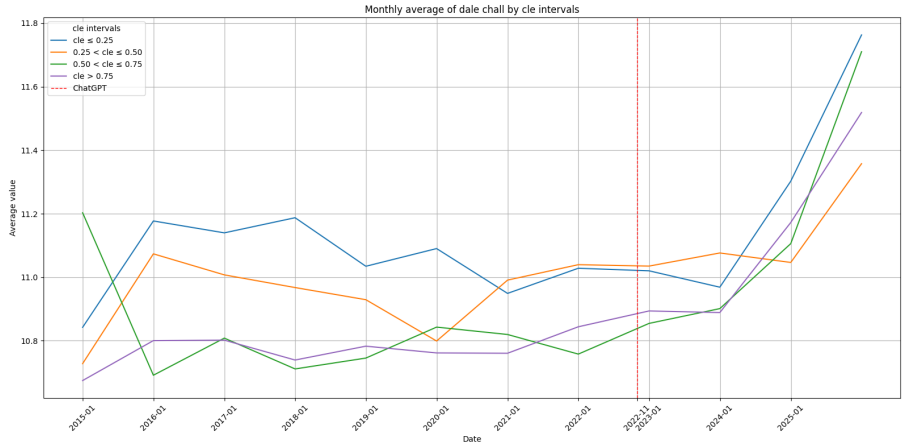

Figure 26: Graphical representation of Dale Chall inde by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are under 0.05, there are statistically significant **increase in grade level requirement** to read abstracts.

Quantifying the impact:
On average, before and after the launch of ChatGPT, countries the furthest away from English-speaking had the biggest grade level requirement (11.12).
Based on the Dale Chall analysis, the countries that had the **biggest change grade level requirement** are the countries **close to English (CLE 0.5-0.75)** (3.46%).

7.1.6. Automated readability

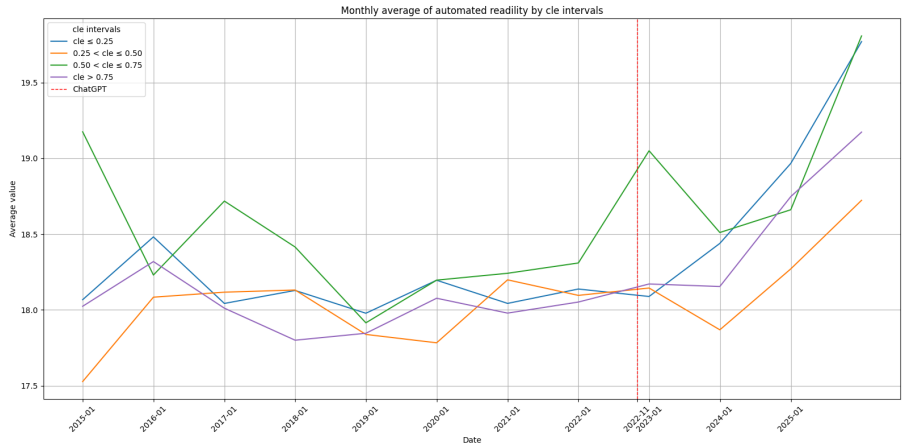

Figure 27: Graphical representation of automated readability by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are under 0.05, there are statistically significant **increase in grade level requirement to read abstracts.**

Quantifying the impact:
On average, before and after the launch of ChatGPT, countries the furthest away from English-speaking had the second highest grade level requirement (18.18 and 18.88).
Based on the automated readability analysis, the countries that had **the biggest change in grade level requirement** are the countries **close to English (CLE 0.5-0.75)** (4.27%).

7.2. Structural measurements

7.2.1. Average length word

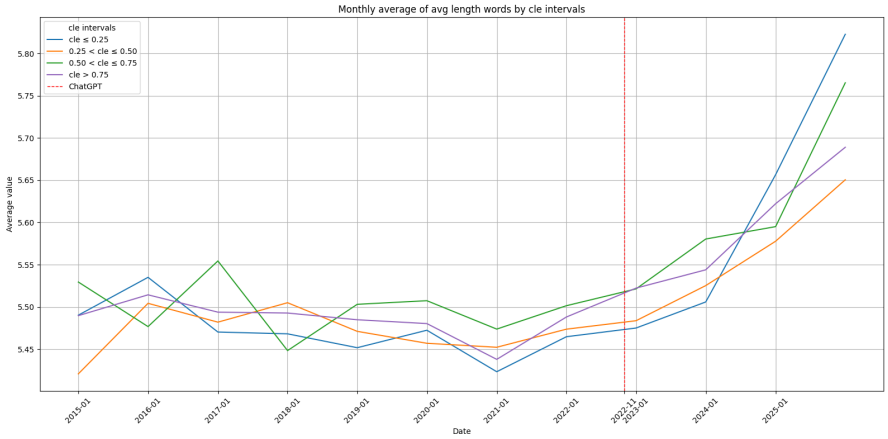

Figure 28: Graphical representation of average length words by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are under 0.05, **there are statistically significant increase in average word length.**

Quantifying the impact:
On average, before and after the launch of ChatGPT, countries the furthest away from English-speaking had the lowest average length of words (5.47 and 5.6).
The countries that had **the biggest increase in average length of words are the countries furthest away from English** (2.71%).

7.2.2. Proportion of more than 15 words

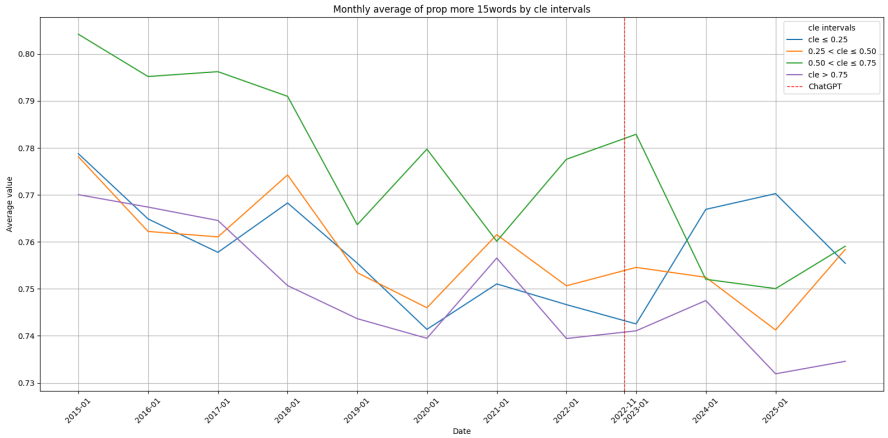

Figure 29: Graphical representation of proportion of sentences of more than 15 words by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are under 0.05, there are statistically significant **change in proportion of sentences of more than 15 words**.

Quantifying the impact:
On average, before the launch of ChatGPT, countries the furthest away from English-speaking had the smallest proportion of sentences with more than 15 words (0.753) and then after they had the biggest proportion (0.7614).
The countries that had **the biggest positive percentage change** are the countries **furthest away from English** (1%). The other countries have a smaller proportion of sentences with 15 words or more.

7.3. Other measurements

7.3.1. TTR (type-token ratio)

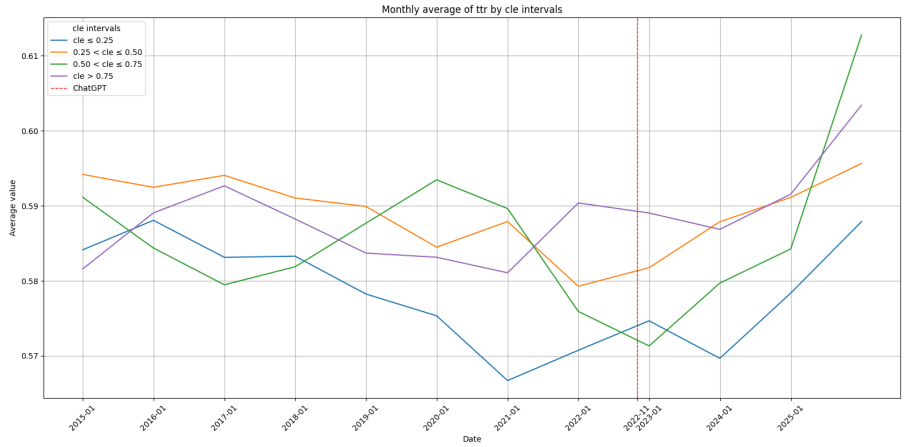

Figure 30: Graphical representation of TTR by CLE over time

Overall impact on all ranges:
Country away from English with a CLE of 0.25 to 0.75 had **no change in their lexical diversity**.
For the others, there are statistically significant changes.

Quantifying the impact:
On average, before and after the launch of ChatGPT, countries the furthest away from English-speaking had the smallest lexical diversity (0.57).
All types of countries had very **small percentage changes** for this metric.

7.3.2. Page count

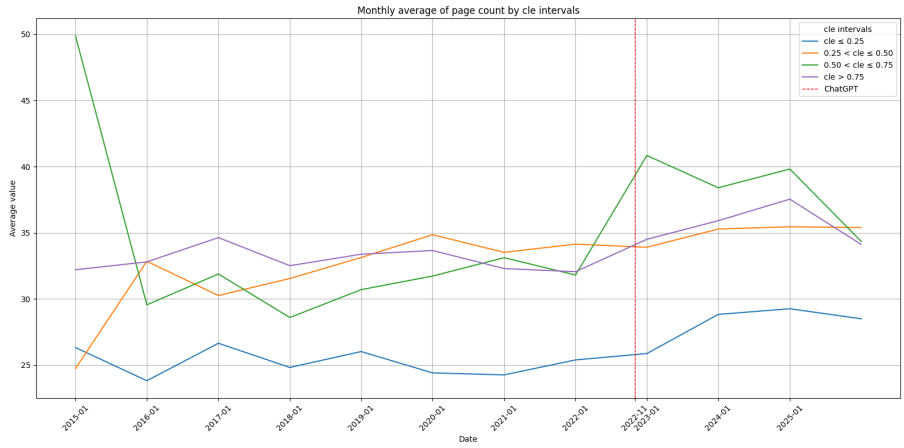

Figure 31: Graphical representation of average page count by CLE over time

Overall impact on all ranges:
For all types of countries because the p-value of the t-test are under 0.05, **there are statistically significant change in the average number of pages.**

Quantifying the impact:
On average, before and after the launch of ChatGPT, countries the furthest away from English-speaking have the smallest average amount of pages (24 and 27).
The type of country with highest percent increase of average number of pages are the ones **pretty close to English** (CLE between 0.50 and 0.75) (23%).

7.4. Comparing the average % change

cle_group	avg_abs_%_change
0-0.25	6,11
0.25-0.50	2,28
0.50-0.75	5,71
0.75-1.00	3,98

Figure 32: Table of average percent change

The highest average percent of change in all the metrics is the group of countries the furthest away from English. Therefore, **ChatGPT impacts more non-native English-speakers countries**.

8. Conclusion

In conclusion, after having scraped hundreds of thousands of articles and computed readability scores I can answer the questions raised.

Does ChatGPT impact academic research?

Yes, from the first analysis of all papers, it can be inferred that **the launch of ChatGPT had statistically significant impacts on academic research**.

How does ChatGPT impact academic research?

On the one hand, abstracts of papers became **more difficult to read,** as their necessary grade level to understand them increased, based on the increase in complexity of words and characters per words and their number of pages, although the lexical diversity did not vary significantly. On the other hand, it decreased the average common level index, indicating **more papers come from countries away from English**.

Whose academic research does ChatGPT impact?

From the second analysis of this paper, it can be drawn that abstracts **became harder to read across all countrie**s. Text difficulty, grade level, word length, number of pages increased, except for lexical diversity.

Whose academic research does ChatGPT impact the most?

Although every type of CLE wrote harder abstracts, the one **furthest away from English** had the most significant change on average in all metrics.

An interesting finding was the stability of the lexical diversity, whether it was in all the papers or per type of country. It might be related to the fact that ChatGPT does not generate nor invent, as it only regurgitates them from the internet. It would be interesting to deeper investigate this aspect.

9. Bibliography

[1] Marr, Bernard. 2023. "A Short History of ChatGPT: How We Got to Where We Are Today." Forbes, May 19, 2023. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

[2] OpenAI. ChatGPT. https://chatgpt.com/

[3] SSRN. https://www.ssrn.com/index.cfm/en/

[4] Hengel, Erin. 2022. "Publishing While Female." January 18, 2022. https://www.erinhengel.com/research/publishing_female.pdf

[5] American Economic Association. "JEL Classification Codes Guide." https://www.aeaweb.org/econlit/jelCodes.php?view=jel

[6] ROR API. Zenodo. https://zenodo.org/records/15731450

[7] Melitz, Jacques, and Farid Toubal. 2014. "Native Language, Spoken Language, Translation and Trade." Journal of International Economics 92 (2): 351–363. https://www.cepii.fr/cepii/fr/bdd_modele/bdd_modele_item.asp?id=19

[8] Textstat. https://pypi.org/project/textstat/

[9] Wikipedia. "Flesch Reading Ease." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests#Flesch_reading_ease

[10] Wikipedia. "Flesch–Kincaid Grade Level." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests#Flesch%E2%80%93Kincaid_grade_level

[11] Wikipedia. "Gunning Fog Index." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Gunning_fog_index

[12] Wikipedia. "SMOG Index." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/SMOG

[13] Wikipedia. "Dale–Chall Readability Formula." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Dale%E2%80%93Chall_readability_formula

[14] Wikipedia. "Automated Readability Index." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Automated_readability_index

[15] Wikipedia. "Student's t-test." Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Student%27s_t-test