

Writing an MA thesis with the help of Large Language Models? Beyond ChatGPT in Academic English



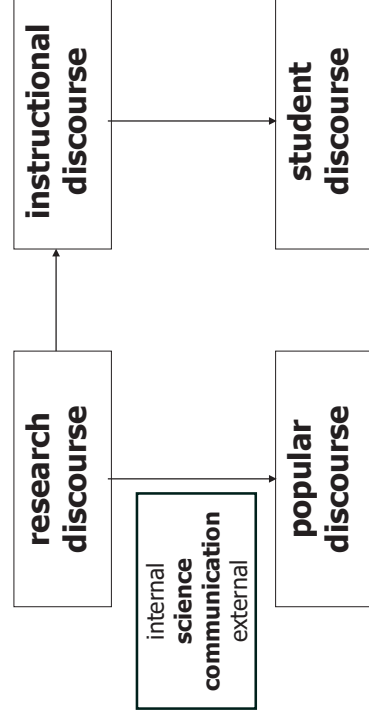
Josef Schmiied
English Language & Linguistics
Chemnitz University of Technology
https://www.tu-chemnitz.de/phl/english/sections/ling/presentations_js.php
josef.schmiied@phil.tu-chemnitz.de
<http://tinyurl.com/Josef-Schmiied>

1. Introduction: Research Experience

BIG Research Projects:

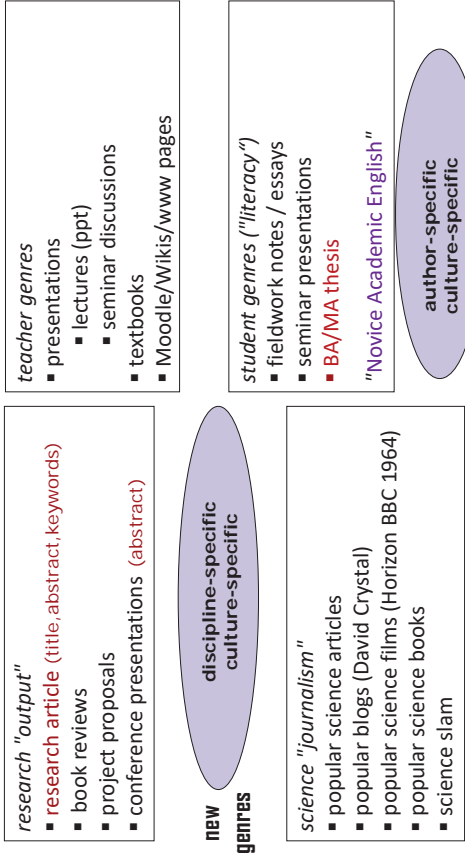
- SFB: Identity in Africa: A5 Sociolinguistics of English variation in Africa, International Corpus of English – East Africa (→Kenya, Tanzania)
- Lampeter Corpus of EModE Tracts (1640-1740) (DFG)
- Internet Grammar of English (DFG; incl. stratified English-German Translation Corpus)
- Academic & Journalistic Writing (DAAD; SE Europe, Africa, China)
- SFB/CRC: Hybrid Societies: Humans interacting with embodied digital technologies: D03 Credibility & Language Varieties: Acceptance of conversational pedagogical agents (Chinese, Italian, Czech “Accents”)
- Marie Curie: BIG DATA, Emilia+ coordinated by Bologna

1. Introduction: Academic Discourses



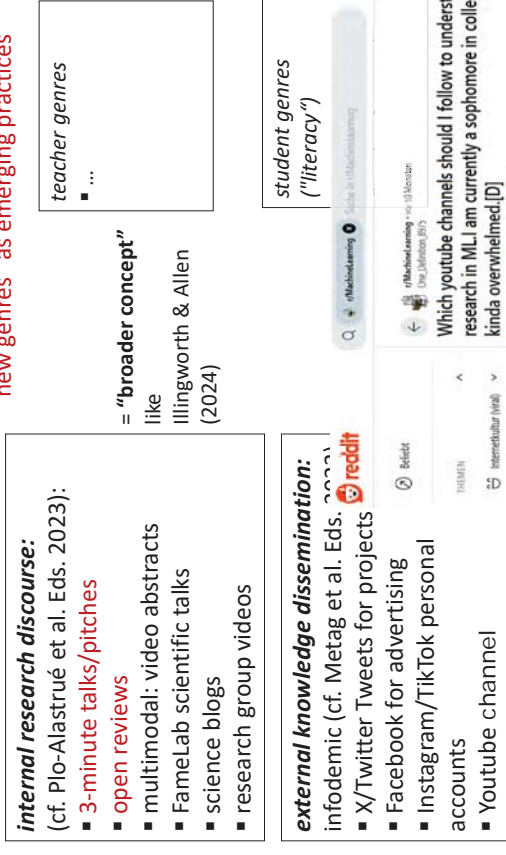
1. Introduction: Academic Discourse Genres (traditional)

“(staple) genres” as conventionalised practices (cf. Schmieid 2015)

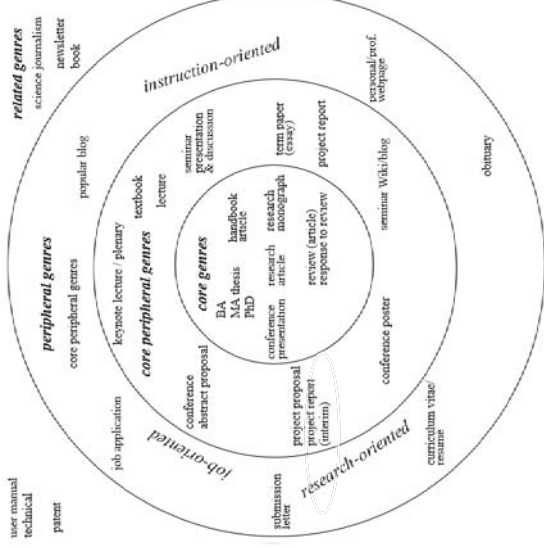


1. Introduction: Academic Discourse Genres (new)

“new genres” as emerging practices



1. Introduction: Academic Writing Genres



1. Introduction: LLM Research Writing Applications

„Artificial Intelligence“

social media

chatbots

Large Language Models (LLMs)

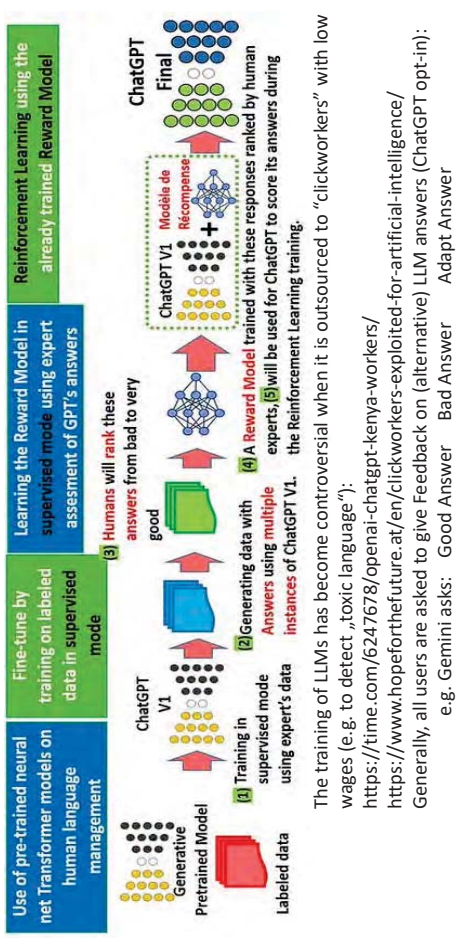
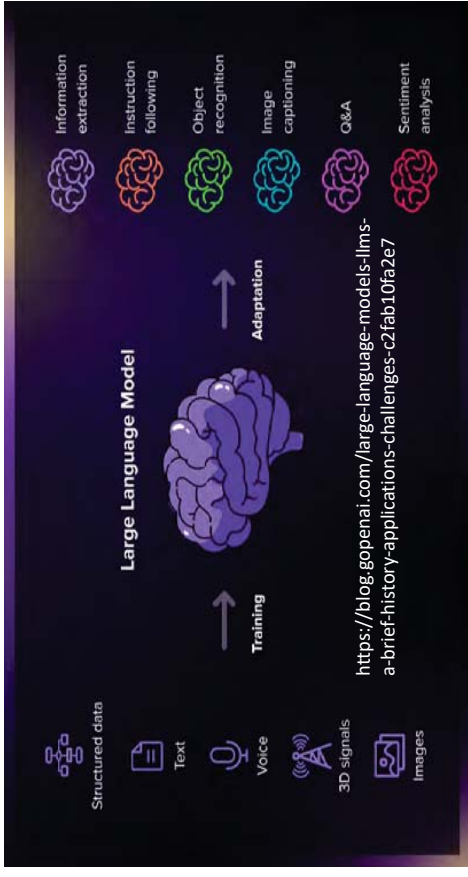
„Office“

Teacher

Researcher for Ac. Writing (+critical thinking)
NOT Data Processing & Analysis, Outreach, etc.
(cf. Ray 2023, Farrokhnia et al. 2024)

Lecture Planning Course-Curriculum-Translation
Skills Practice „Rules“, etc.
Specialised Concepts

Academic Style
Publishing Conventions
abstract title key words metadiscourse

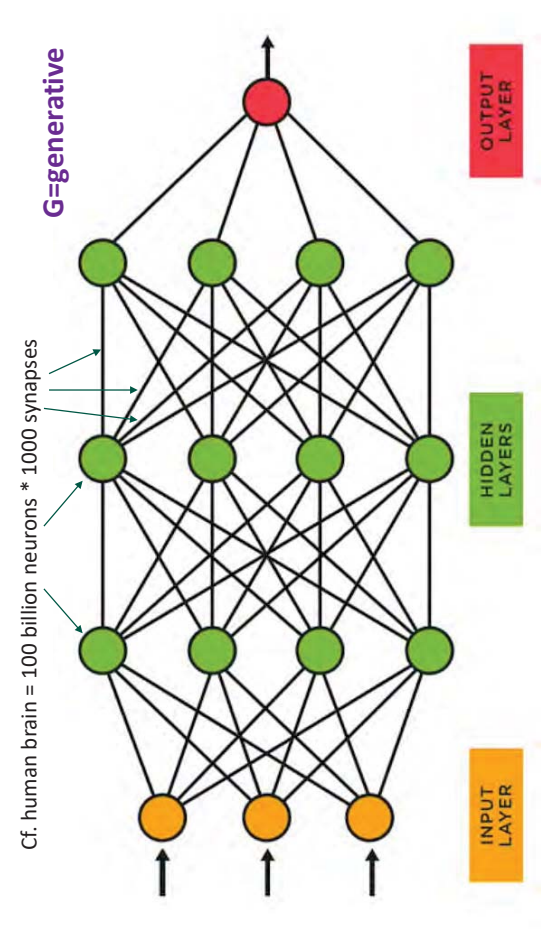


P=pre-trained

What Are Large Language Models?

An LLM is a machine learning model that processes and generates human language from training on massive volumes of data. It uses a deep learning algorithm to learn the patterns and nuances of language and provide human-like responses to a wide variety of queries and prompts. LLMs can also learn new skills such as creative writing or sentiment analysis.

The "large" in its name refers to the size of the datasets on which it's trained. The goal of an LLM at its core is to infer what comes next in a sequence of texts. ChatGPT is a conversational agent built on top of an LLM that can generate text in a way that's extremely similar to humans. When it comes to expanding the search experience, Coveo is testing generative AI capabilities that surface answers directly within search results.



2. Large Language Models: T=Transformer Visualisation

Chapter: Output

LLM

How to predict

tokens

pos embed

tok embed

transformer i

multi-head, causal self-attention

layer norm

feed forward

layer norm

linear softmax

Table of Contents

Introduction

Preliminaries

Components

Embedding

Layer Norm

Self Attention

Projection

MLP

Transformer

Softmax

Output

T=transformer

<https://bbcroft.net/llm>

2. Large Language Models: Linguistic Principles „MI“

English-Corpora.org

corpora PDF guides videos related resources users my account upgrade help

In our corpora, Mutual Information is calculated as follows:
 $MI = \log \left(\frac{AB}{A \cdot B} \right) / \log(2)$

Suppose we are calculating the MI for the collocate *color* near *purple* in BNC.
 $A =$ frequency of node word (e.g. *purple*): 1246
 $B =$ frequency of collocate (e.g. *color*): 112
 $AB =$ frequency of collocate near the node word (e.g. *color* near *purple*): 22
 $sizeCorpus =$ size of corpus (# words; in this case the BNC: 96,263,399
 $span =$ span of words (e.g. 3 to left and 3 to right of node word): 6
 $\log(2)$ is literally the natural log of the number 2: 0.6931

$MI = 11.30 = \log \left(\frac{22 \cdot 96,263,399}{(1246 \cdot 112 \cdot 6)} \right) / 0.6931$

BNCweb

HELP	GO	WORDS	THRO	ALL	%	WP	
1	1	ADVERB	7	87.62	3.06	1.00	
2	2	ADJECTIVE	11	36.44	12.25	1.00	
3	3	VERB	14	35.71	12.17	1.00	
4	4	PRONOUN	5	34	26.47	11.73	1.00
5	5	CONJUNCTION	22	112	19.64	11.30	1.00
6	6	NUMERAL	13	372	3.49	8.31	1.00
7	7	PREPOSITION	6	184	3.26	8.71	1.00
8	8	ARTICLE	8	339	2.36	9.25	1.00
9	9	INTERJECTION	1	1	0.00	0.00	1.00
10	10	PROPER NOUN	43	2103	1.29	7.49	1.00
11	11	PERSON	13	1005	1.29	7.36	1.00
12	12	ORGANIZATION	13	1598	1.19	7.26	1.00
13	13	LOCATION	9	857	1.05	7.08	1.00
14	14	DATE	7	706	0.98	7.00	1.00
15	15	PLANT	7	706	0.98	7.00	1.00

2. Large Language Models: T=Transformer Visualisation

Chapter: Output

LLM

How to predict

tokens

pos embed

tok embed

transformer i

multi-head, causal self-attention

layer norm

feed forward

layer norm

linear softmax

Table of Contents

Introduction

Preliminaries

Components

Embedding

Layer Norm

Self Attention

Projection

MLP

Transformer

Softmax

Output

T=transformer

<https://bbcroft.net/llm>

2. Large Language Models: SWOT Analysis ChatGPT in education

Helpful to achieve goals

Harmful to achieve goals

Internal factors

External factors

Strengths

Weaknesses

Opportunities

Threats

Generating plausible responses

Self-improving capability

Providing personalised responses

Providing real-time responses

Lack of deep understanding

Difficulty in evaluating the quality of responses

The risk of biases and discrimination

Lack of higher-order thinking skills

Increasing accessibility of information

Facilitating personalised learning

Facilitating complex learning

Decreasing teaching workload

Lack of understanding of the context

Threatening academic integrity

Perpetuating discrimination in education

Democratisation of plagiarism in education

Declining in high-order cognitive skills

2. Large Language Models: Linguistic Principles „collocates“

Corpus of Contemporary American English

SEARCH WORD CONTEXT ACCOUNT

alcohol (NOUN)

drinking, drug, drink, liquor, substance, cocaine, abuse, smoking, addiction, marijuana, heroin, smoking, consumption, drink, beer, drunk, treatment, excessive use, harmful

COLLOCATES (more)

NOUN drug, use, abuse, consumption, tobacco, problem, blood, level

VERB drink, consume, rub, avoid, abuse, smell, smoke, associate

ADJ fetal, excessive, illegal, heavy, moderate, illicit, smoking, related

ADV eg, ie, prenatally, voluntarily, negatively, excessively, moderately, capita

RELATED WORDS

alcoholic, alcoholism, alcohol-related, alcohol-free, alcohol-based, blood-alcohol

ChatGPT

Sure, here are some examples of things you can tell me to do:

1. Generate a poem about nature.
2. Summarize the plot of "To Kill a Mockingbird" by Harper Lee.
3. Provide a list of the top 10 tourist attractions in Paris.
4. Explain the concept of quantum entanglement.
5. Create a simple recipe for chocolate chip cookies.
6. Generate a random 5-digit number.
7. Translate the phrase "Hello, how are you?" into Spanish.
8. Summarize the key points of a TED Talk on artificial intelligence.
9. Describe the process of photosynthesis in plants.
10. Provide a brief biography of Albert Einstein.

Which task would you like to start with?



Mizumoto/Eguchi (2023):

The world was taken by surprise by the release of ChatGPT (<https://chat.openai.com/>) at the end of the year 2022. ChatGPT is a chatbot that answers human questions with an AI that seems to have a perfect understanding of the language. ...

The emergence of an AI tool capable of comprehending and producing a language like ChatGPT marks the beginning of a new era in which humans and AI coexist in L2 language learning and teaching, as well as in its research.

However, the use of ChatGPT is currently being evaluated, and some academic journals have cautioned against its irresponsible application (e.g., Nature Editorial, 2023). Despite criticisms from scholars like Chomsky, who denounce ChatGPT as a form of "high-tech plagiarism" (EduKitchen, 2023), it is evident that AI language models are here to stay, and it is our responsibility to educate students on their ethical use because "the future is now" (Pavlik, 2023, p. 9). At the same time, learners, teachers, and researchers can benefit from acquiring effective methods for utilizing GPT in their work, particularly non-native English speakers, who can be linguistically empowered. Language models such as GPT can be best understood as an ever-present, logical assistant, as AI is not currently a substitute for human expertise.

Large Language Models (LLMs) are powerful tools, but their use in research articles raises some ethical concerns. Here are some key areas to consider:

- Authorship and Credit:** Who gets credit for research findings when an LLM helps generate content? Should the LLM be a co-author? How much human oversight is needed? These are open questions that can blur the lines of traditional authorship. Oxford study on LLM responsibility [invalid URL removed]
- Transparency and Reproducibility:** Research relies on being able to repeat and verify findings. If an LLM is involved in data analysis or literature review, how can researchers ensure transparency about the LLM's role and limitations? Can others replicate the research if they don't have access to the same LLM?
- Bias and Accuracy:** LLMs are trained on massive amounts of data, which can reflect societal biases. These biases can be unknowingly incorporated into research articles, leading to skewed or inaccurate findings. Researchers need to be aware of these potential biases and take steps to mitigate them.
- Overtrust and Misinformation:** It's easy to be impressed by an LLM's ability to generate text. However, LLMs can still produce factually incorrect or misleading content. Researchers need to critically evaluate the LLM's outputs and avoid uncritically accepting them as truth.

These are just some of the ethical considerations surrounding LLM use in research. As this technology continues to develop, researchers and publishers will need to grapple with these issues and develop best practices for ethical and responsible LLM use.

As an AI assistant, I can help you with various aspects of writing a research article:

1. **Literature Review**
 2. **Structuring the Article:** I can provide guidance on the typical structure of a research article, including the introduction, literature review, methodology, results, discussion, and conclusion sections. I can help you organize your ideas and ensure a logical flow of information.
 3. **Writing and Editing:** I can help you with the actual writing process, from drafting paragraphs and sections to improving clarity, coherence, and language usage. I can also assist with editing and proofreading to ensure your article is well-written and free of errors.
 4. **Formatting and Referencing**
 5. **Data Analysis and Visualization**
 6. **Manuscript Submission**
 7. **Collaboration and Feedback**
- It's important to note that while I can provide valuable assistance, I cannot replace your subject-matter expertise or critical thinking. My role is to support you throughout the writing process, but the research, analysis, and conclusions should ultimately come from you and your co-authors.

3. Brainstorming: Ask ChatGPT 4 ethical issues

Using Large Language Models (LLMs) for research articles raises several ethical considerations that researchers should carefully address. Here are some key ethical issues:

1. Bias and Fairness
2. Plagiarism
3. Authorship and Attribution
4. Privacy and Confidentiality
5. Manipulation and Misinformation
6. Impact on Academic Publishing
7. Environmental Impact
8. Equitable Access

Addressing these ethical issues requires a combination of **awareness, transparency, accountability, and proactive measures** to mitigate potential risks and promote responsible use of LLMs in research articles.

3. Brainstorming with Bard (=Gemini): writing for journals

write on digital dissemination practices for linguists for publication in Peer-Reviewed Journals

choice of options

Vorschlag 1

Sure, here are some digital dissemination practices for linguists for publication in peer-reviewed journals:

Vorschlag 2

Sure, here are some digital dissemination practices for linguists for publication in peer-reviewed journals:

Vorschlag 3

Sure, here are some digital dissemination practices for linguists for publication in peer-reviewed journals:

informal style

Sure, here are some digital dissemination practices for linguists for publication in peer-reviewed journals:

- **Publish in open access journals:** Open access journals make their articles freely available online, which can help to increase the visibility and impact of research. There are a number of open access journals that specialize in linguistics, such as the *Journal of Linguistics* and the *International Journal of Language and Communication Disorders*.
- **Create and share research data:** Linguists can share their data through online repositories, such as the Linguistic Data Consortium and the Open Linguistics Archive. This can make it more accessible to other researchers and the public, and can facilitate further research on the data.
- **Use social media:** Social media platforms like Twitter, Facebook, and LinkedIn can be used to share research findings with a wider audience. Linguists can create accounts for themselves and their research groups, and can use these accounts to post links to their articles, data, and other resources. They can also use social media to interact with other linguists and the public, and to answer questions about their research.
- **Create and maintain a website:** Linguists can create websites to showcase their research and make it easier for people to find their work. The

4. Research Publishing: ChatGPT academic style

Consensus GPT excels in a variety of research applications, including:

- **Real Study Insights:** Ask Consensus to interpret findings from genuine studies on topics like the cognitive benefits of creatine.
- **Enhanced Academic Writing:** Utilize Consensus to draft literature reviews or academic pieces with valid citations.
- **Targeted Paper Searches:** Effortlessly locate specific research papers without relying on precise keyword matches.
- **Advanced Filtering:** Apply filters to refine searches based on publication type, journal prestige, and study design.
- **Research Paper Analysis:** Have Consensus examine and summarize key elements of a research document.
- **Concept Extraction:** Request a compilation of strategies or findings across multiple studies.
- **Science-based Content Creation:** Draft authoritative blogs on health and wellness topics grounded in scientific evidence.

<https://consensus.app/home/blog/introducing-researchgpt-by-consensus/>
 We are excited to announce the official launch of Consensus, formerly known as ResearchGPT, as the leading GPT product available in OpenAI's GPT marketplace.
 We're not just a GPT. Did you know you can try the full Consensus search engine experience for free at <https://consensus.app/home/?>

4. Research Publishing: GPT 4 Research Article Structure

Research articles in digital humanities typically follow a structure that combines traditional humanities scholarship with innovative digital methodologies and tools. While the specific structure may vary depending on the nature of the research and the preferences of individual researchers or journals, here's a prototypical structure commonly used in digital humanities research articles:

1. Title Page
2. Abstract
3. Keywords
4. Introduction
5. Literature Review
6. Methodology
7. Case Study or Analysis
8. Discussion
9. Conclusion
10. References
11. Appendices (if applicable)

<https://token.ujk.edu.pl/volumes/volume-16/>
TOKEN 16 (2023)
Special issue: Patterns of Language Variation and Change in Academic Writing
Guest editors: Josef Schmieid (Chernitz University of Technology), Marina Bondi (University of Modena and Reggio Emilia), Olga Dontcheva Navratlova (Masaryk University) and Carmen Pérez-Llantada (University of Zaragoza)

This structure provides a framework for organizing research articles in digital humanities, integrating traditional humanities scholarship with innovative digital methodologies and technologies. Researchers should **adapt and refine this structure to fit the specific requirements** and conventions of their discipline, as well as the aims and scope of their research projects.

4. Research Publishing: the Art of Prompting

- Title**
Suggest a title for this research journal abstract
- Abstract**
Suggest a 300 word abstract for this title: Using large Language Models to write Research articles
- Keywords**
Suggest 5 keywords for this research journal abstract
- Introduction**
Write an (historical?) introduction for this research journal abstract
- Literature Review**
Write a literature review for this research journal abstract
- Methodology**
Which methodologies can I use for this article?
- Case Study or Analysis**
Find 3 case studies for this issue
- Discussion**
- Conclusion**
Which conclusions may I arrive at in this article?
- References**
Which 10 references are crucial for this article?

4. Research Publishing: GPT4o Conclusions (without analysis?)

Based on the findings from the case studies and the literature review, several conclusions can be drawn regarding the utilization of Large Language Models (LLMs) in research article composition:

- Enhanced Writing Efficiency and Productivity:**
 - LLMs offer researchers a powerful tool for streamlining the writing process, leading to increased efficiency and productivity.
 - Researchers using LLMs can generate coherent and contextually relevant text more quickly, enabling them to produce research articles in a more timely manner.
- Facilitated Interdisciplinary Collaboration**
- Improved Writing Creativity and Exploration**
- Ethical Considerations and Challenges**
 - The use of LLMs in research article composition raises important ethical considerations related to bias mitigation, plagiarism prevention, and authorship attribution.
 - Ethical guidelines and safeguards are essential to ensure responsible use of LLMs and mitigate potential risks associated with bias, plagiarism, and data privacy.
- Opportunities for Future Research and Development**

In conclusion, the integration of LLMs into research article composition holds immense potential for enhancing writing efficiency, promoting interdisciplinary collaboration, and fostering creativity in the scholarly community. However, it is crucial for researchers to navigate the ethical considerations associated with LLM-assisted writing responsibly and transparently. By critically evaluating the opportunities and challenges of LLMs in academic writing, researchers can harness their transformative potential to advance knowledge and scholarship in the digital age.

DeepL Write

English (britisch) | Anmelden | 756 / 2000

Optimierte Texte mit DeepL Write

The rapid development of AI opens up new possibilities for learning and using foreign languages. Although tools like ChatGPT, Bard, Perplexity, or DeepL are currently free, their applications need to be learned to use them effectively for personal or professional needs.

In this course, we will explain the basics of the most common AI tools (often called "stochastic parrots"). We'll apply these tools to various types of English texts, such as travelogues, book reviews, summaries, etc., and discuss the results generated by AI. We don't write - we improve. We don't translate - we compare versions. This course is suitable for anyone who wants to create foreign-language texts, especially at an advanced level, either for personal or professional purposes.

756 / 2000

4. Research Publishing: Improve AI Proposals

DeepL Write

English (britisch) | Anmelden | 756 / 2000

Optimierte Texte mit DeepL Write

The rapid development of AI opens up new possibilities for learning and using foreign languages. Although tools like ChatGPT, Bard, Perplexity, or DeepL are currently free, their applications need to be learned to use them effectively for personal or professional needs.

In this course, we will explain the basics of the most common AI tools (often called "stochastic parrots"). We'll apply these tools to various types of English texts, such as travelogues, book reviews, summaries, etc., and discuss the results generated by AI. We don't write - we improve. We don't translate - we compare versions. This course is suitable for anyone who wants to create foreign-language texts, especially at an advanced level, either for personal or professional purposes.

756 / 2000

4. Research Publishing: Evaluate AI Proposals

DeepL Write

English (britisch) | Anmelden | 756 / 2000

Optimierte Texte mit DeepL Write

The rapid development of AI opens up new possibilities for learning and using foreign languages. Although tools like ChatGPT, Bard, Perplexity, or DeepL are currently free, their applications need to be learned to use them effectively for personal or professional needs.

In this course, we will explain the basics of the most common AI tools (often called "stochastic parrots"). We'll apply these tools to various types of English texts, such as travelogues, book reviews, summaries, etc., and discuss the results generated by AI. We don't write - we improve. We don't translate - we compare versions. This course is suitable for anyone who wants to create foreign-language texts, especially at an advanced level, either for personal or professional purposes.

756 / 2000

Style!! academic

Prototypical (culture-, discipline-specific) expectations vs. Author Identity

- adapt
- Set specific academic style: technical terms + metadiscourse
 - Customise voice (not Standard US)

Several large language model (LLM) applications allow users to upload their own writings to teach the model their personal style. These applications offer fine-tuning or customization options, which can be especially useful for personalizing outputs. Here are a few prominent ones:

1. **OpenAI's GPT-3 and GPT-4 (via API and Playground):**
 - OpenAI allows users to fine-tune models using their own data. By providing a set of examples of your writing, you can customize the output to better match your personal style.
 - **Playground:** Offers a user-friendly interface for testing and tweaking model outputs with your custom prompts and data.
 - **API:** Provides a more extensive setup for fine-tuning and deploying customized models in various applications.

how can I express my academic identity in my academic English through metadiscourse?

good concepts
typical examples
no references

Expressing your academic identity in your academic writing, especially through metadiscourse, is crucial for establishing credibility, engaging readers, and conveying your authority on the subject. Metadiscourse refers to the language used to guide readers through your text and convey your attitude and stance toward the material. Here are some strategies for using metadiscourse effectively to express your academic identity in academic English:

1. **Hedging and Boosting:** Use hedging to indicate uncertainty or caution, and boosting to emphasize your confidence in your claims. For example:
 - Hedging: "It appears that the results may suggest a correlation..."
 - Boosting: "The robust findings unequivocally support the hypothesis..."
2. **Engagement Markers:** Engage with your readers by addressing them directly or signaling your awareness of their presence. This can foster a sense of connection and collaboration. For instance:
 - "In this paper, we will demonstrate..."
 - "As you can see from the data..."
3. **Epistemic Modality:** Indicate your degree of certainty or belief in the information you're presenting. Use adverbs and modal verbs like "clearly," "possibly," "likely," or "certainly" to express your stance:
 - "This finding clearly demonstrates..."
 - "It is likely that this phenomenon occurs because..."

Research Context: Conversational Pedagogical Agents (CPA) in Hybrid Societies (cf. Ivanova, Albrecht, Schmiel fc.).

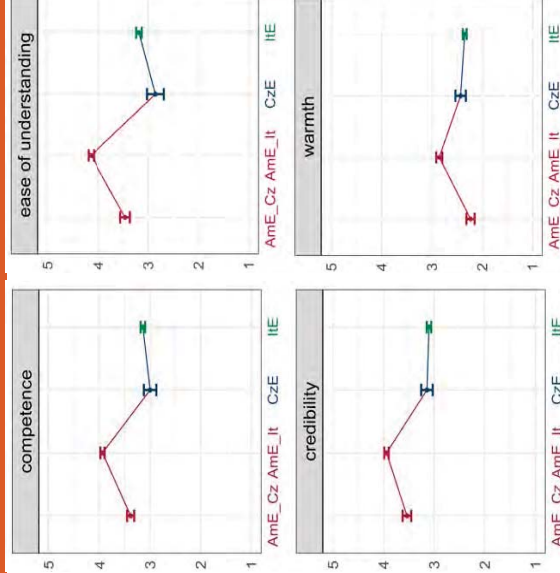
Mission
In hybrid societies, humans and embodied digital technologies should interact as seamlessly as humans among each other.

- RQ1** Which specific non-native linguistic cues of CPAs influence the learning performance of non-native human learners?
- RQ2** Which specific non-native linguistic cues influence attributed credibility and acceptance of CPAs by non-native human learners?
- RQ3** How much does a linguistically credible CPA influence the learning performance in non-native educational contexts?

Non-native English lecture



Italian and Czech students give lower rates to their own varieties of English than to American English, for competence and warmth



6. Conclusion: The human role remains essential

ChatGPT thinks human judgment remains essential (Zappavigna 2023)

1. Defining the task
2. Data pre-processing
3. Evaluation and refinement
4. Interpretation and analysis

“Generative AI and the end of corpus-assisted data-driven learning? Not so fast!” (Crosthwaite & Baisa 2023)

- Knowing the data
- Authenticity
- Replicability
- Multimodality
- Safety
- Hallucinations
- Active vs. passive learning

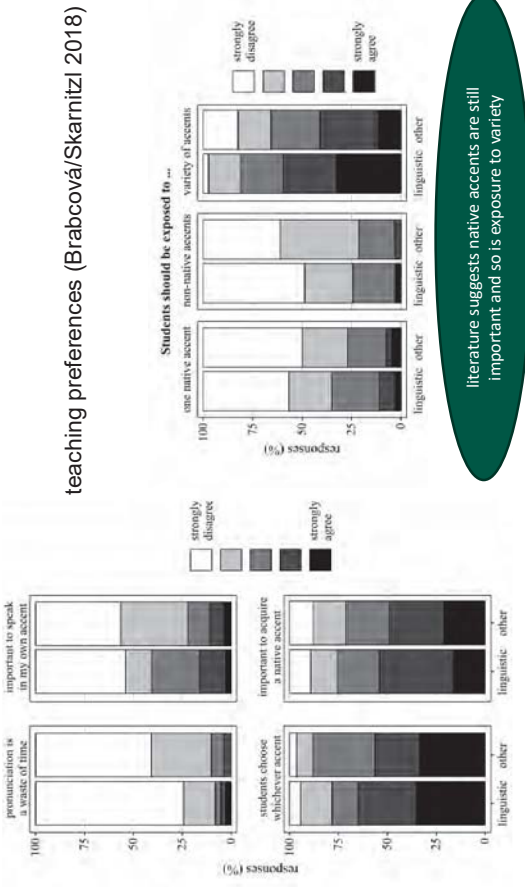
“traditional” reliable resources are available online

6. Conclusion: AI detection tools

The screenshot shows the GPTZero interface with a 100% AI likeness score. It lists top sentences driving AI probability with their scores:

- 1. It helps to soften assertions and make them more diplomatic. (0.41)
- 2. In business texts, hedging is commonly employed to maintain a sense of professionalism and objectivity, and to acknowledge uncertainty or variability in certain situations. (0.30)
- 3. Hedging is a linguistic phenomenon where cautious or non-committal expressions are used to avoid making absolute statements or to mitigate potential risks or uncertainties associated with statements. (0.30)

5. Outlook: Reading of Personalised Agents and Credibility



6. Conclusion: Ethical and legal issues

- In the end the AI user decides – and is responsible!
- Copyright issues and unclear!
- <https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it>
- In academic writing and publishing all aids and resources used must be specified clearly!

KD nuggets Blog Topics Datasets Events Resources

After the launch of ChatGPT, the Pandora box opened. We are now observing a technological shift in the ways we do work. People are creating websites, apps, and even writing novels using ChatGPT.

With all the hype and introduction of AI generative tools, we have seen a rise in bad actors. If you are following the latest news, you must have heard that ChatGPT has passed the Wharton MBA exam. There is a whole list of exams passed by ChatGPT, from medical to a law degree - List: [Here Are the Exams ChatGPT Has Passed so Far](#).

Apart from exams, students are using it to submit assignments, writers are submitting generative content, and researchers are producing high-quality papers just by typing prompts.

To counter the abuse of generative content, I am introducing you to 5 free AI content detector tools. You can use them to check the validity of the content and improve your SEO ranking.

<https://www.kdnuggets.com/2023/02/15-free-tools-detecting-chatgpt-gpt3-gpt2.html>

