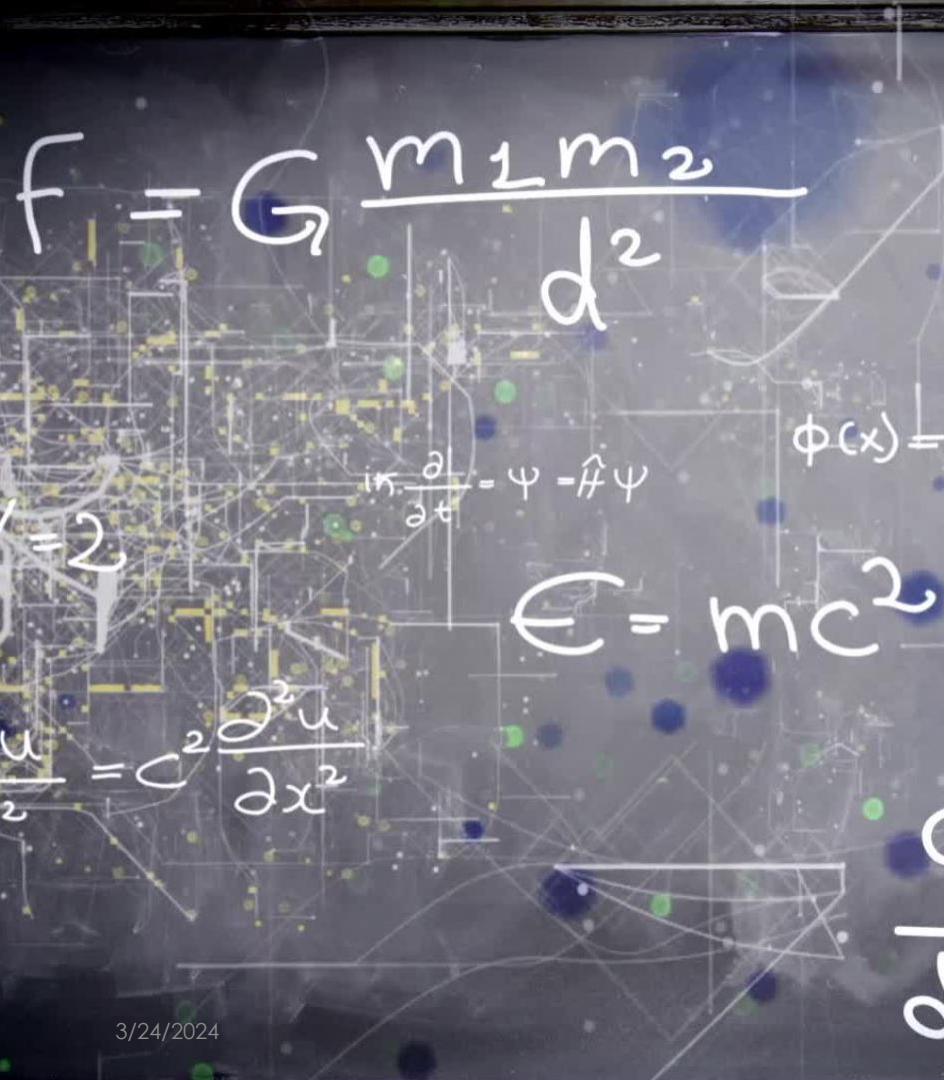
Generative AI and Large Language Models in Industry and Research

József Mezei

Faculty of Social Sciences, Business and Economics, and Law



Åbo Akademi University

My background

- Mathematics (statistics, finance, insurance)
- Actuary in insurance company
- Doctoral degree at Åbo Akademi (decision analytics)
- Post doctoral researcher at ÅA and LUT
- Project based research (optimization, machine learning, network analysis)
- Data scientist at F-Secure (cybersecurity)
- Data scientist at Avaintec Oy (health care analytics)
- Currently: Professor at ÅA in Information Systems

My background: projects

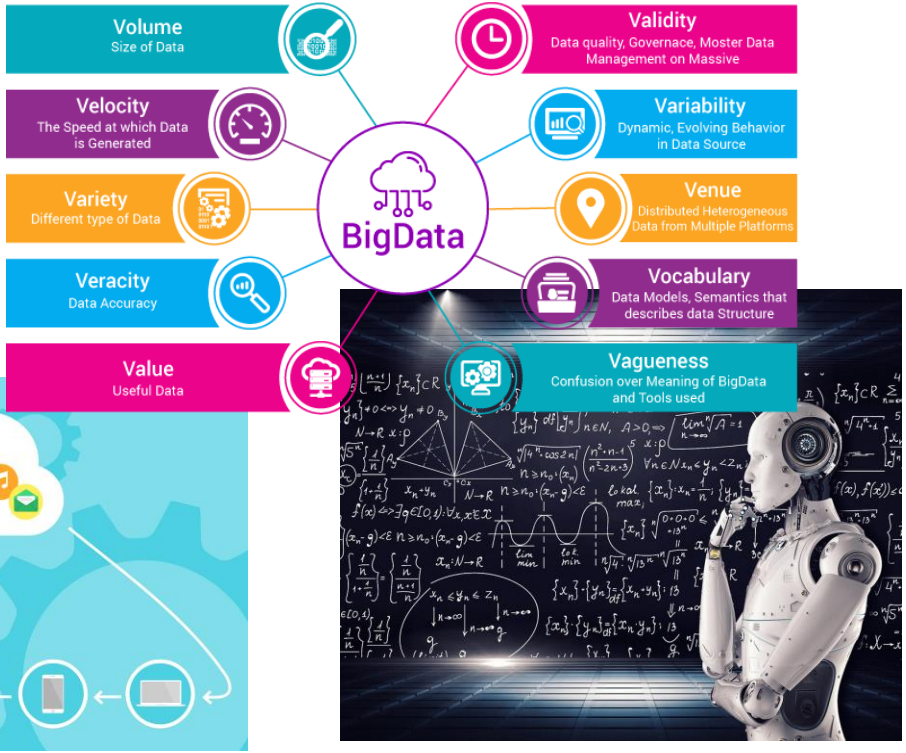
- Car insurance pricing
- Grid computing
- Predictive maintenance
- Knowledge management in industrial processes
- Logistic optimization
- Network analysis in finance (crisis prediction)
- Churn prediction
- Health analytics
- Consumer behaviour

Basic terminology (informal)

- **Business Intelligence**: leveraging past and present data to describe the state of a business today
- **Business analytics**: utilizing data to predict where the business is heading and prescribe actions to optimize future outcomes (Descriptive, Predictive, Prescriptive)
- **Machine learning**: design and development of algorithms that allow computers to evolve behaviours based on empirical data
- **Artificial Intelligence**: the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

Analytics process – Big Data – Cloud Computing – Machine learning

Åbo Akademi University

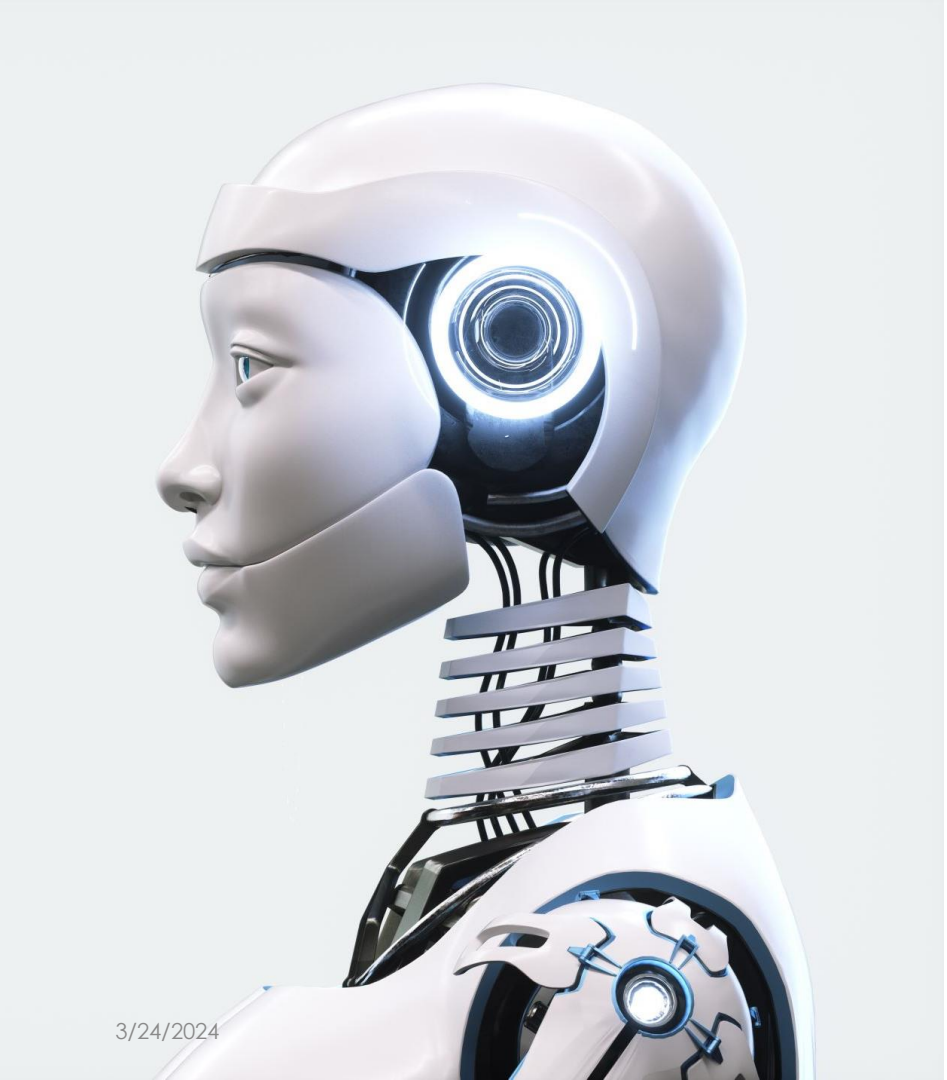


ML-AI applications: Marketing

ML-AI applications: Finance

ML-AI applications: Mind-reading

UTS researcher tests new mind-reading technology. Image: UTS



Åbo Akademi University

What is Artificial Intelligence?

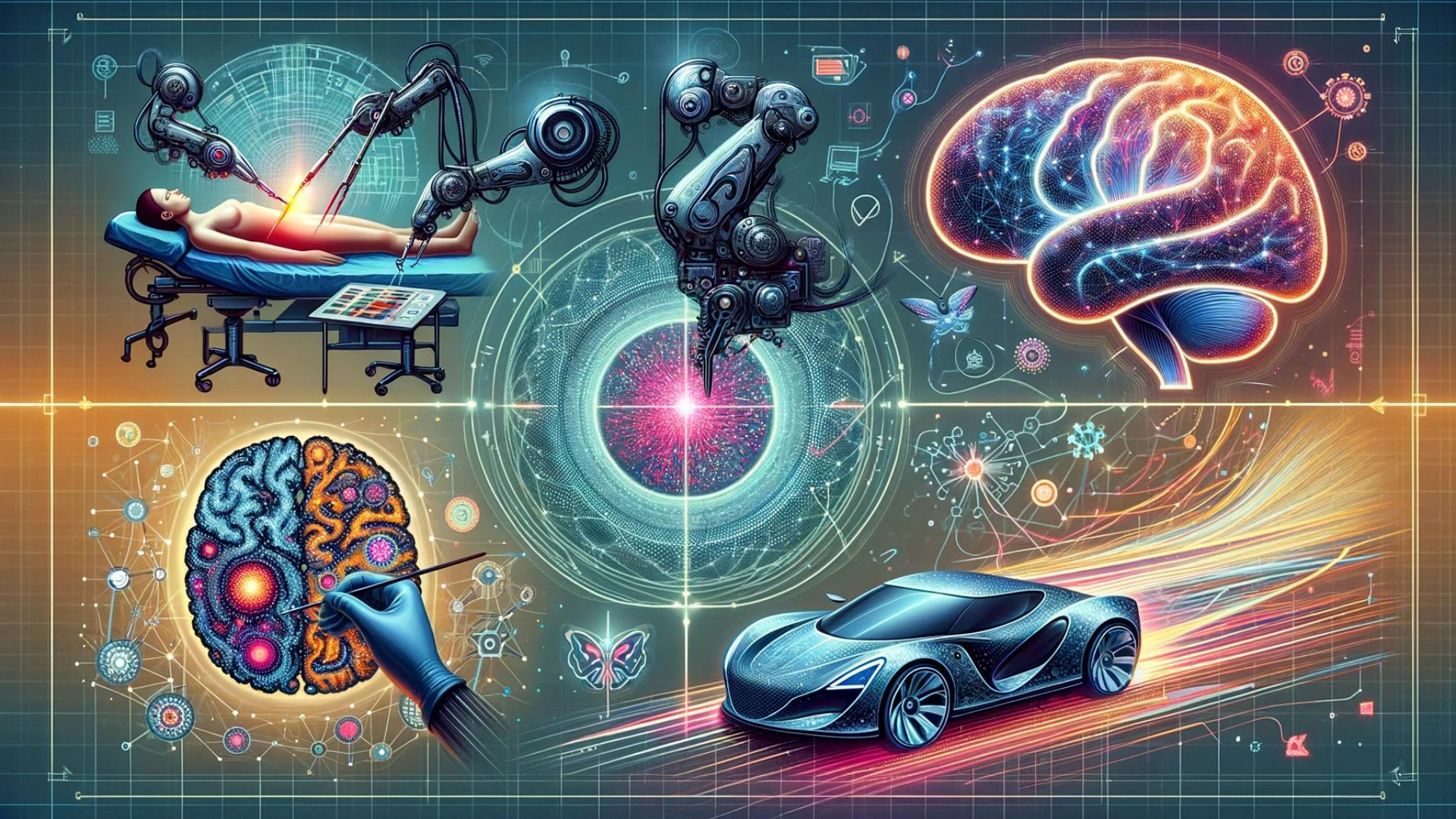
- The simulation of human intelligence in machines that are programmed to think like humans and mimic their actions
  - Learning
  - Reasoning
  - Self-correction
  - Problem-solving
  - Perception
  - Language understanding
- Weak/Narrow AI vs. Strong/General AI


Never Send A
Human To
Do A
Machine's
Job.

Åbo Akademi University

Generative AI

- Generative AI refers to a subset of AI technologies capable of creating new content, from text to images, by learning from existing data
- Built upon neural networks, particularly deep learning, using layers of interconnected nodes that simulate human brain function to process and generate complex data representations



Discriminative vs. Generative models

Discriminative models

Answer closed-ended questions

Learn from training data

Guess correct answer/category

"Is there a cat or dog on the picture?"

Generative models

Guess data for a prediction

Learning from training data

Generate new content

"Draw me a picture of a dog!"

Large Language Models

- Large language models are machine learning models that can process and generate natural language text
- Trained on massive amounts of data, such as books, articles, and websites, to learn patterns and relationships in language
- Some of the most well-known large language models include GPT-4 and BERT
- These models have millions or even billions of parameters, which allows them to generate highly complex and nuanced text (GPT-3: 175 billion parameters, GPT-4: 100 trillion(?) parameters)

LLM applications

- Chatbots and virtual assistants
- Content creation
- Language translation
- Text summarization
- Creative AI

Benefits of LLMs

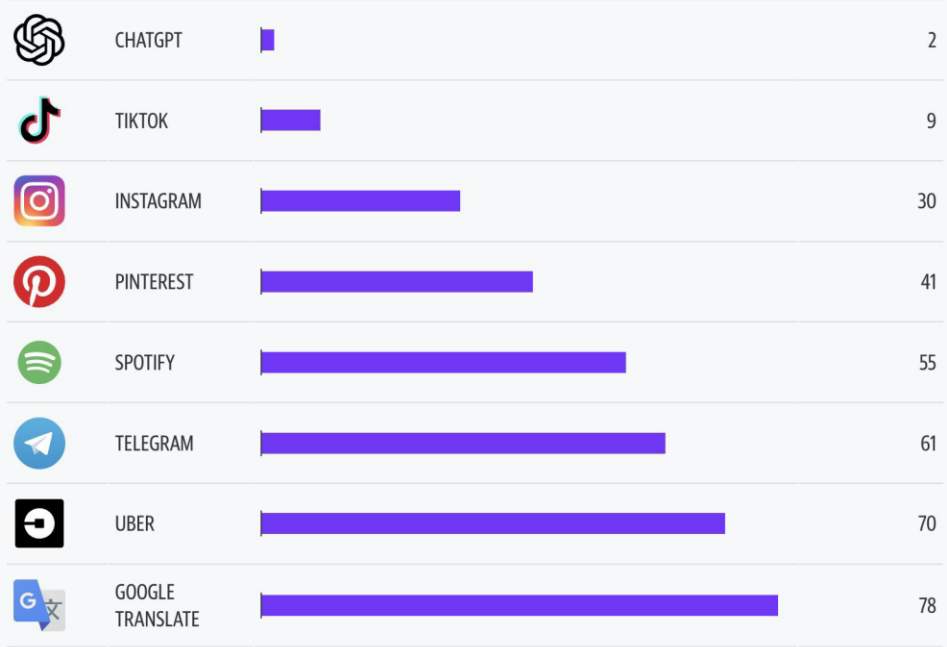
- Time savings: large amounts of high-quality text quickly, saving time and effort
- Cost savings: reduce the cost of content creation, language translation, etc.
- Improved accuracy: highly accurate results, improving the quality of content
- Personalization: Chatbots and virtual assistants provide personalized experiences to users

Challenges of LLMs

- Bias: may perpetuate biases present in the training data, which can lead to unfair or discriminatory outcomes
- Misinformation: Large language models can be used to generate false or misleading information
- Ethical concerns: use raises ethical questions around privacy, security, and accountability
- Computing resources: training and running large language models requires significant computing resources

HOW LONG IT TOOK TOP APPS TO HIT 100M MONTHLY USERS

ChatGPT is estimated to have hit 100M users in January, 2 months after it's launch. Here's how long it took other top apps to reach that:



SOURCE: UBS

yahoo! finance

Åbo Akademi University

Key factors driving development

- Computing power
- Dataset availability
- Competitive interests
- Model design

Components of modeling

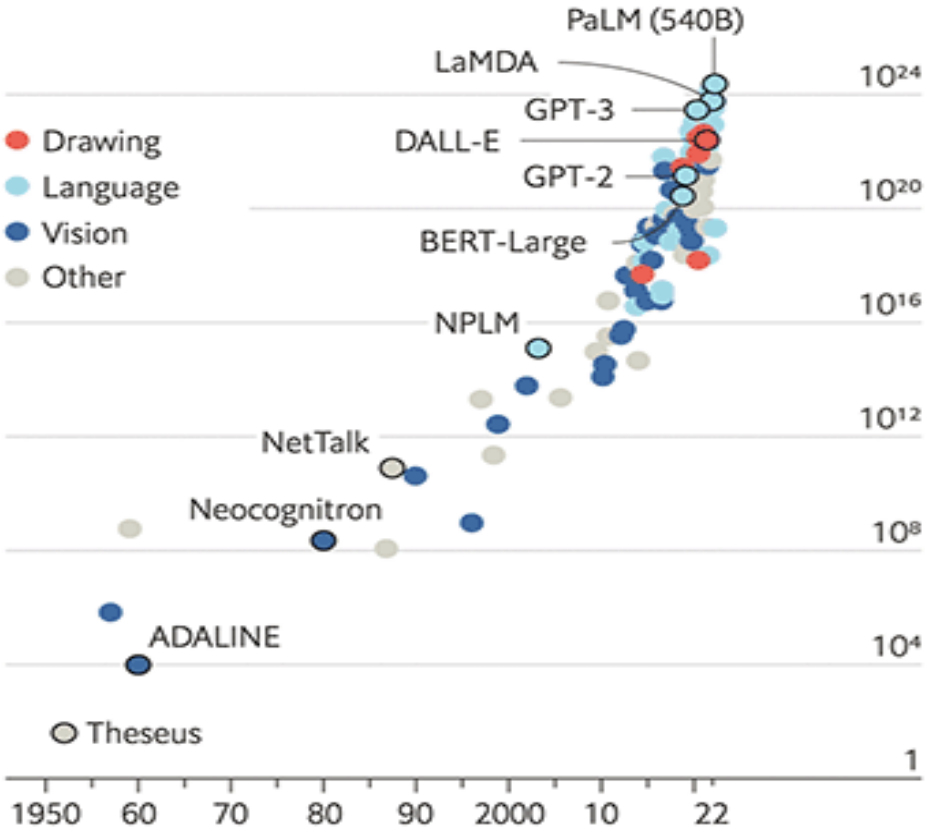
Generative Adversarial Networks

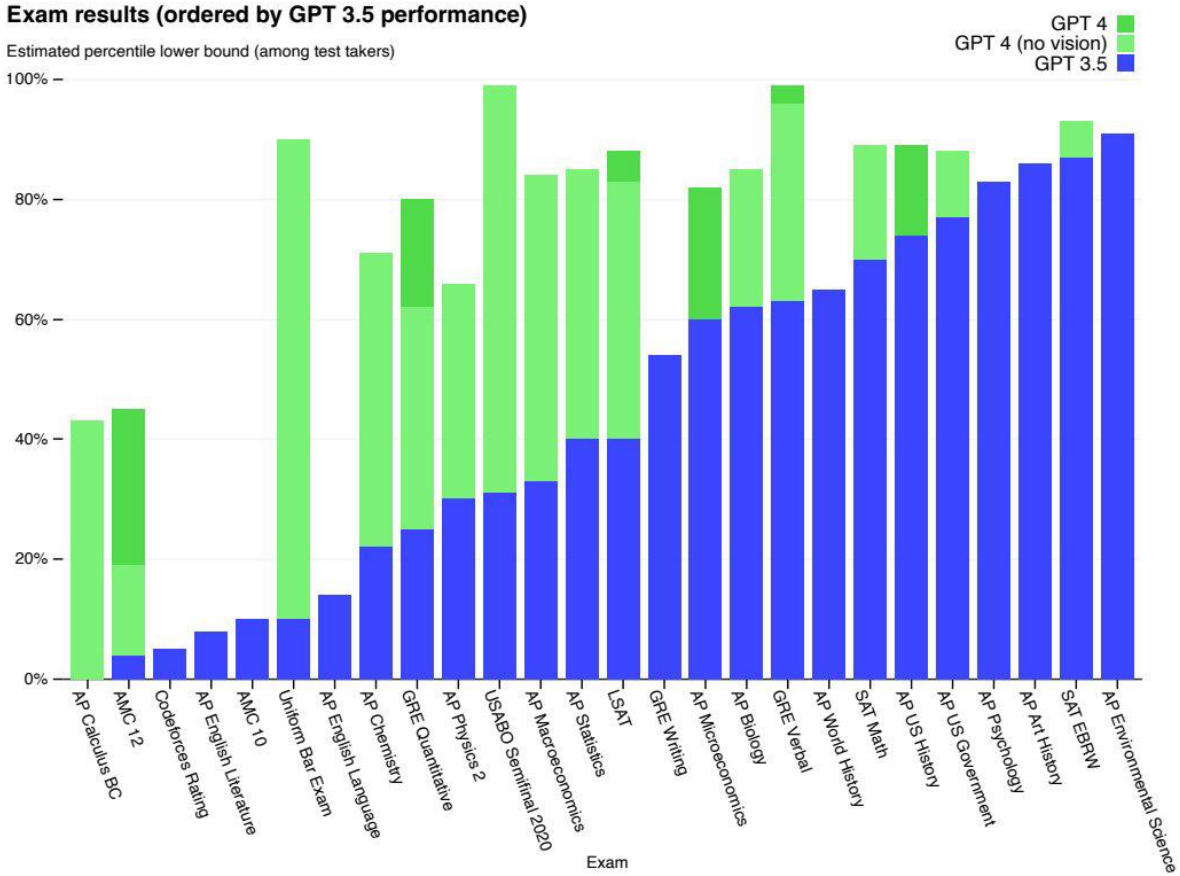
Transformers

Reinforcement learning with human feedback

AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale

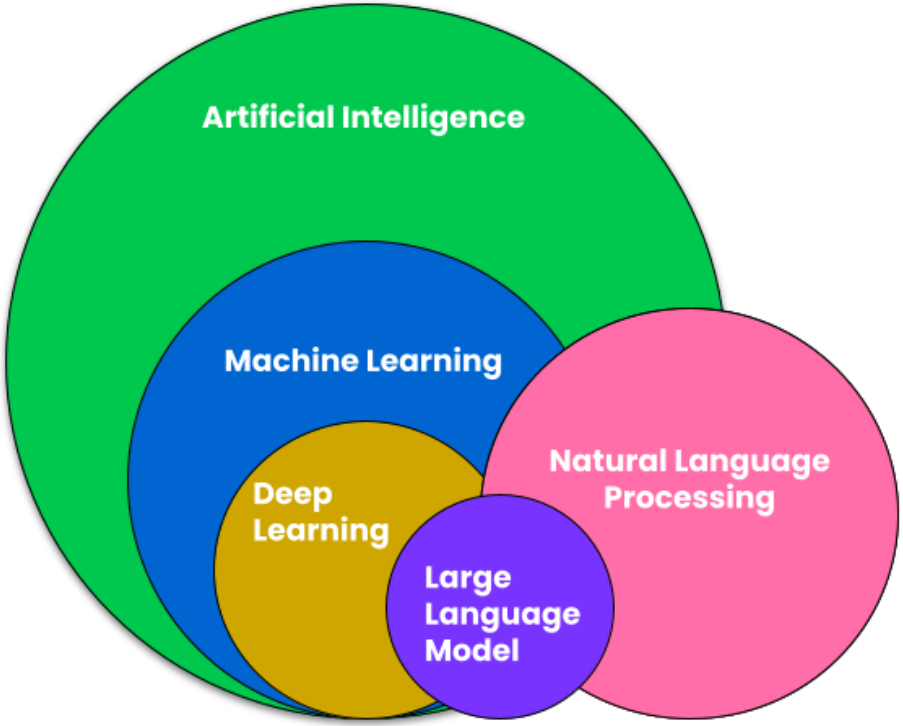




Åbo Akademi University

Large language models

Åbo Akademi University



Applications

- Sentiment analysis
- Identifying themes
- Translating text or speech
- Generating code
- Next-word prediction

Finance

Investment outlook

Annual reports

News articles

Social media posts

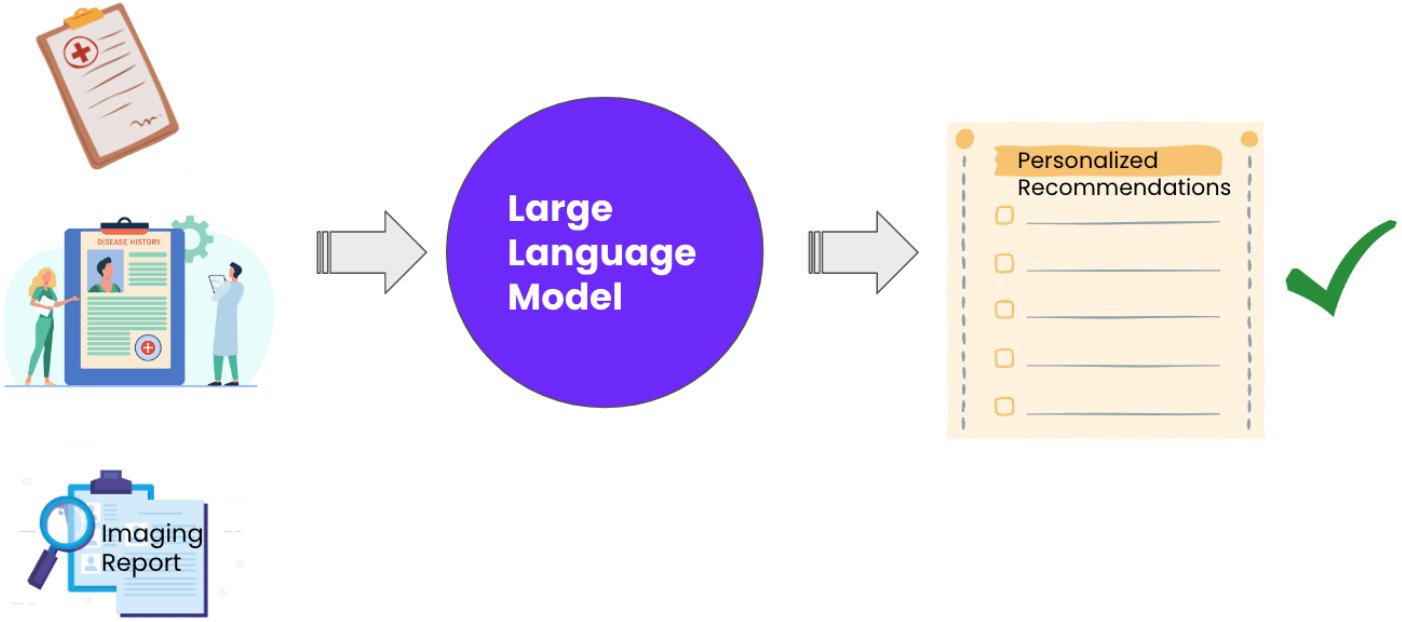
Large Language Model

Market analysis

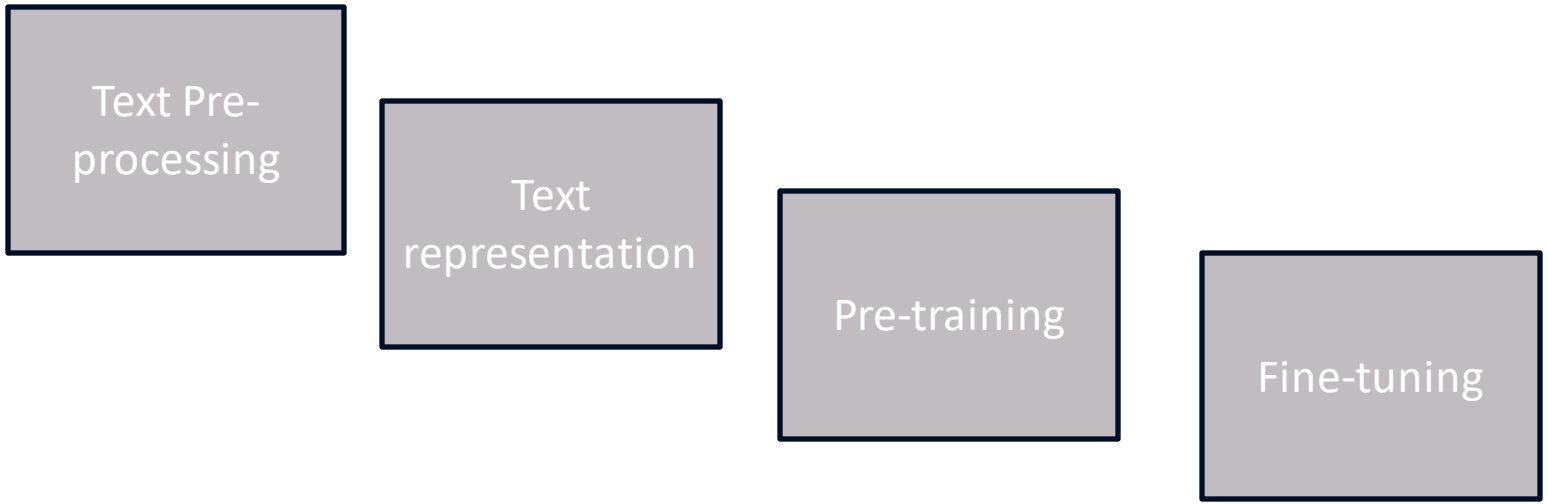
Portfolio management

Investment opportunities

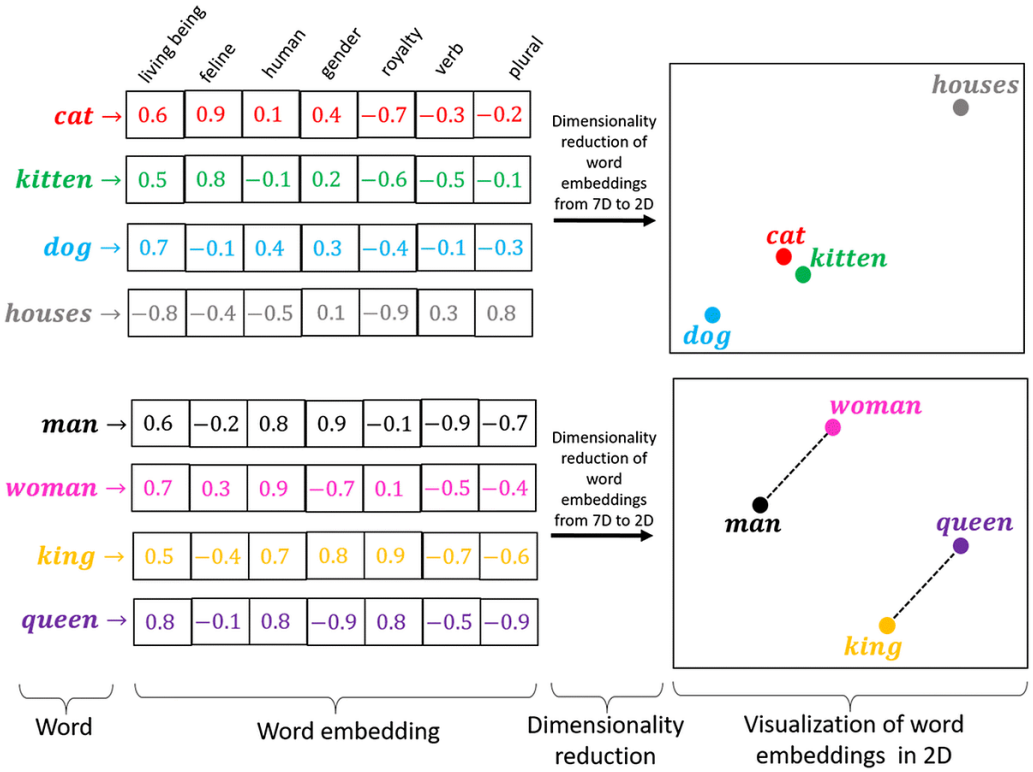
Health care


DISEASE HISTORY
Imaging Report
Large Language Model
Personalized Recommendations

Building blocks of LLMs



Text representation



Model training

Memory, processing power, infrastructure

Expensive (355 years of processing time on a single GPU)

LLM:

- 100,000's Central Processing Units (CPUs)
- 10,000's Graphic Processing Units (GPUs)

A personal computer: 4-8 CPU and 1-2 GPUs

Model training

- Need of high-quality data
- To learn the complexities and subtleties of language
- A few hundred gigabytes (GBs) of text data
- More than a million books

Pre-training vs. Fine-tuning

Pre-training

Compute: thousands of CPUs and GPUs

Training time: weeks or months

Data: hundreds of GBs

Fine-tuning

Compute: 1-2 CPUs and GPUs

Training time: hours or days

Data: couple of GBs

Transformers

Åbo Akademi University

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

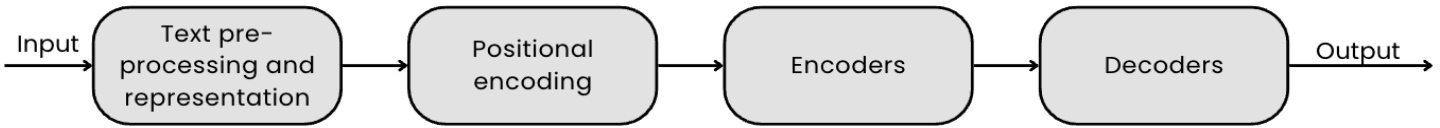
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

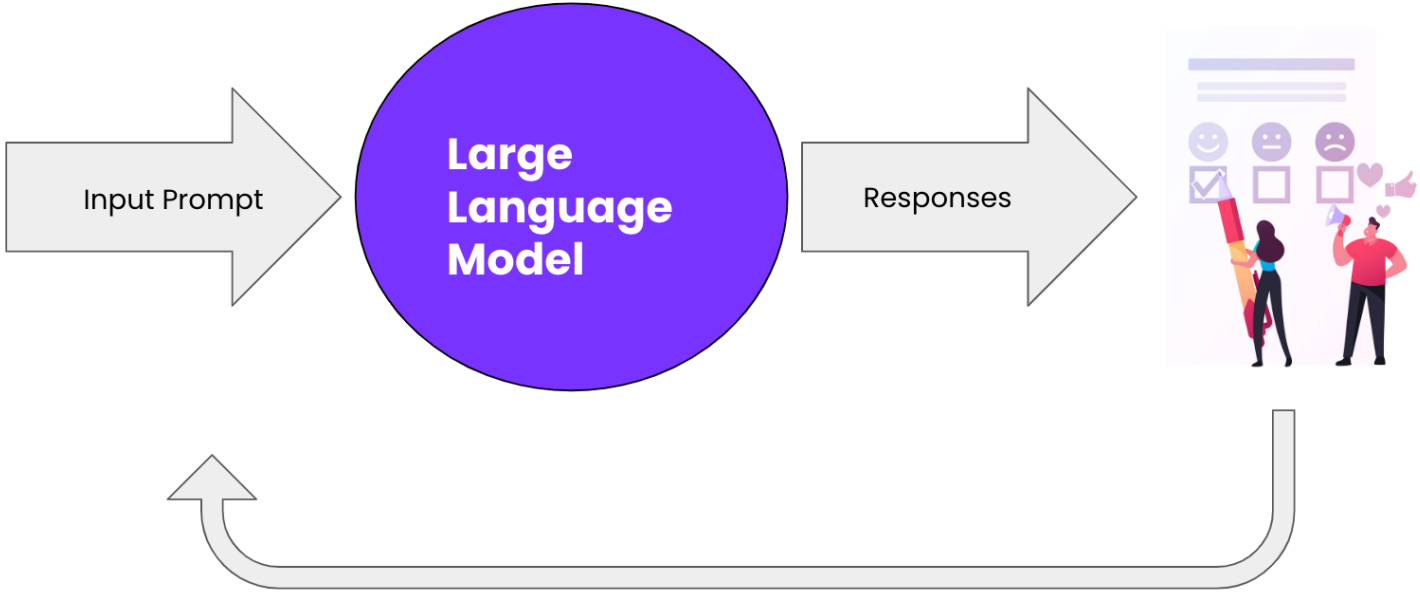
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

6 Dec 2017



Reinforcement learning through human feedback



Articles about the impact of Generative AI and LLMs

Åbo Akademi University

- Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. "Gpts are gpts: An early look at the labor market impact potential of large language models." arXiv preprint arXiv:2303.10130 (2023).
- Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. Generative AI at work. No. w31161. National Bureau of Economic Research, 2023.
- Ritala, Paavo, Mika Ruokonen, and Laavanya Ramaul. "Transforming boundaries: how does ChatGPT change knowledge work?." Journal of Business Strategy ahead-of-print (2023)
- Wahid, Risqo, Mero, Joel, and Ritala, Paavo. "Editorial: Written by ChatGPT, llustrated by Midjourney: generative AI for content marketing"

Articles about the impact of Generative AI and LLMs

Åbo Akademi University

- Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Technology & Operations Mgt. Unit Working Paper 24-013 (2023).

- Girotra, Karan, Lennart Meincke, Christian Terwiesch, and Karl T. Ulrich. "Ideas are dimes a dozen: Large language models for idea generation in innovation." *Available at SSRN 4526071* (2023).

White papers/industry reports about Generative AI and LLMs

- McKinsey & Company: Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom
- BCG AI Radar: From Potential to Profit with Generative AI
- STM Association: Generative AI in Scholarly Communications

Current trends

Åbo Akademi University

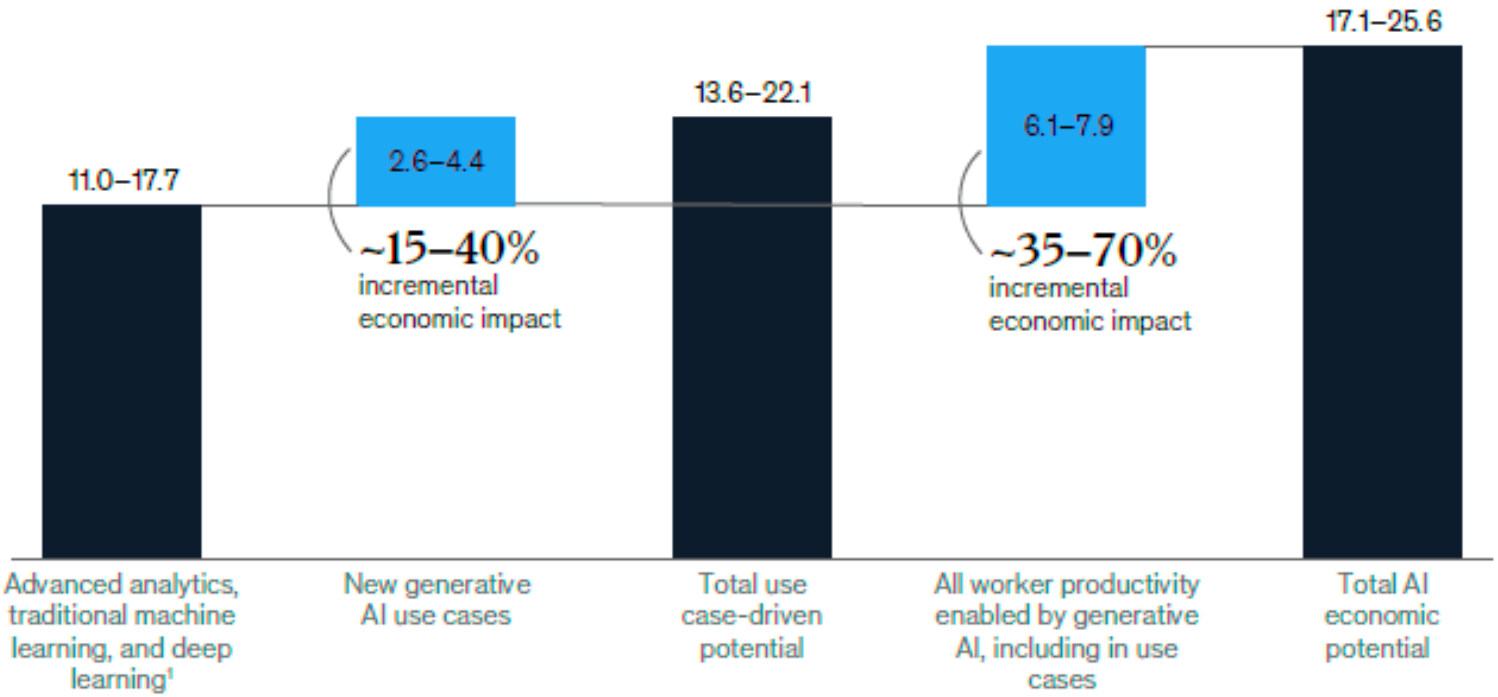
- **89%** of executives rank AI and GenAI as a top-three tech priority for 2024
- **54**% of leaders expect AI to deliver cost savings in 2024 (productivity gains in operations, customer service etc.)
- Only **6**% of companies have managed to train more than 25% of their people on GenAI tools so far
- **45**% of leaders say that they don't yet have guidance or restrictions on AI and GenAI use at work
- **66**% of the executives believe that it will take at least two years for AI and GenAI to move beyond the hype, focused on pursuing limited experimentation and small-scale pilots
- On average, approx. **50**% of employees will need to be reskilled in the next 3 years
- The total percentage of working hours that could theoretically be automated by integrating technologies that exist today is **50% to 70%**

Implementing Generative AI

Åbo Akademi University

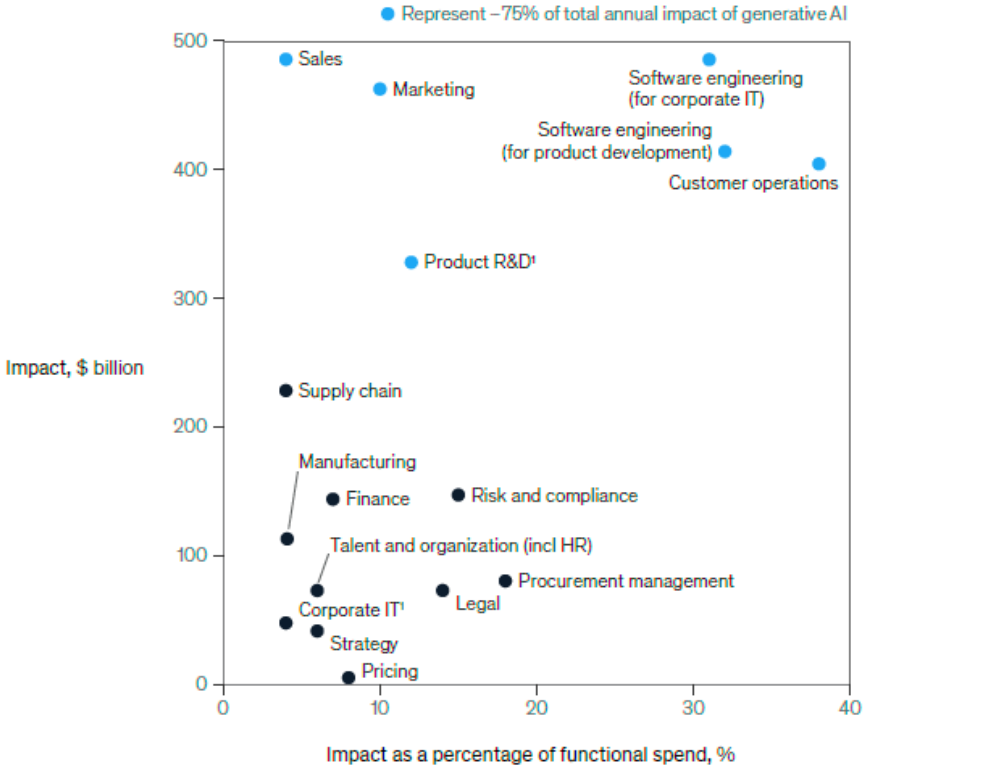
- **Taker**
  - Off-the-shelf coding assistant for software developers
  - General-purpose customer service chatbot with prompt engineering only and text chat only
- **Shaper**
  - Customer service chatbot fine-tuned with sector-specific knowledge and chat history
- **Maker**
  - Foundation model trained for assisting in patient diagnosis

AI's potential impact on the global economy, $ trillion



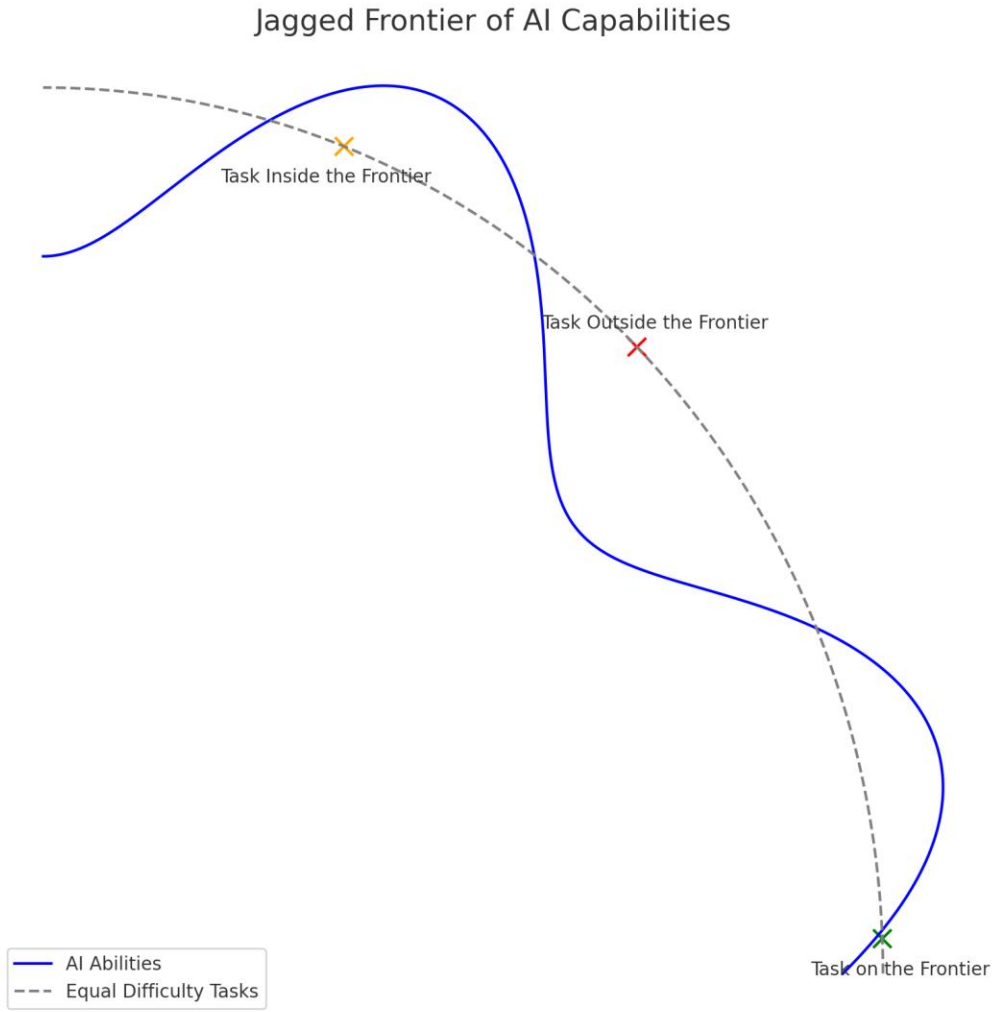
[1]Updated use case estimates from "Notes from the AI frontier: Applications and value of deep learning," McKinsey Global Institute, April 17, 2018.

Using generative AI in just a few functions could drive most of the technology's impact across potential corporate use cases.



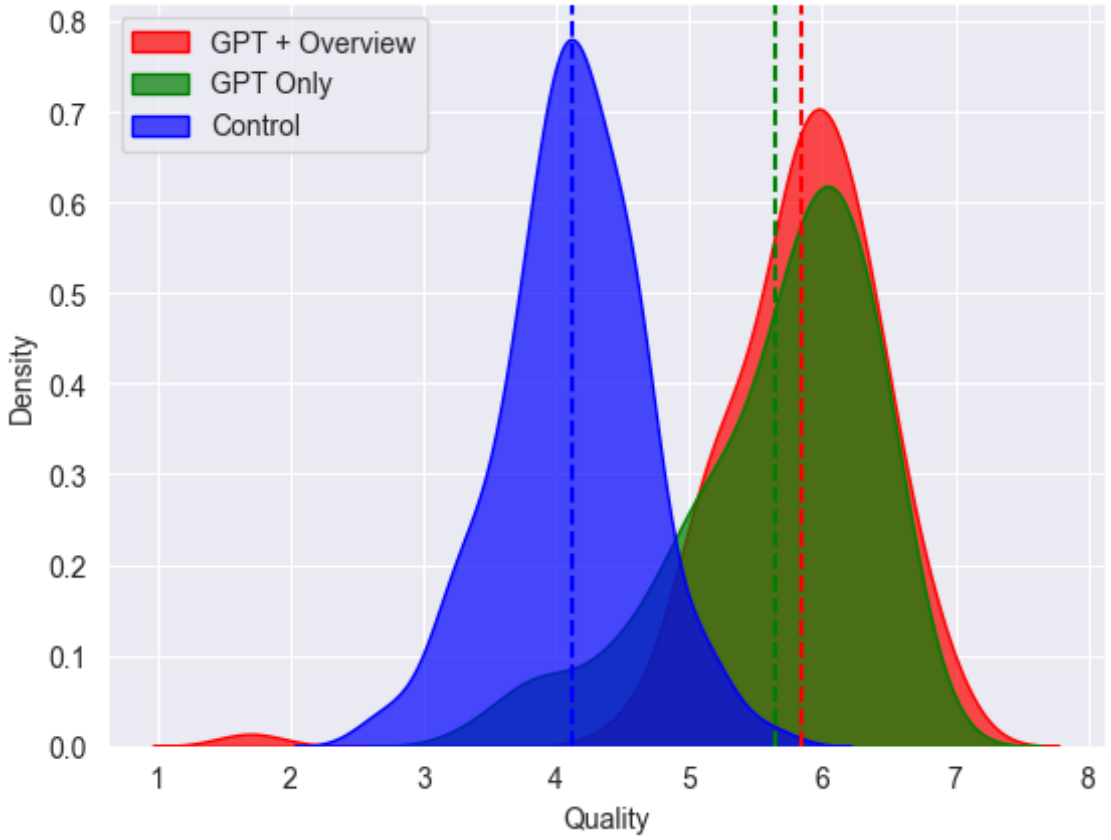
Note: Impact is averaged.
[1]Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

Jagged AI frontier (Dell'Acqua et al., 2023)



Åbo Akademi University

Performance comparison (Dell'Acqua et al., 2023)



Åbo Akademi University

Centaur practices

Åbo Akademi University

User BA1: You are writing a memo to a CEO of a company to inform him of where his company should focus as they attempt to drive revenue growth for one of their three brands, kleding man. You want to provide him with the following facts for why he should focus on driving revenue growth for kleding man:

1. Several of his team members recognize that their brand awareness in this category has slipped and they are no longer serving the older men demographic

2. Their price point is too high for younger men, their current target demographic which is why they have not been able to even retain constant sales

3. Their share of revenue for Kleding Men has dropped 20% or 11 percentage points, or a 5.44% CAGR.

4. Their revenue for Kelding Man has dropped by 22% or a 6% CAGR since 2013, or roughly ~$19M in absolute terms

5. Compared to market, the total % of the clothing market that is made up of menswear is 52%, they are at 44% of their bands indicating there is room to grow and bounce back

Generative AI: Subject: Driving Revenue Growth for Kleding Man

Dear [CEO's Name],

I hope this message finds you well. Our team has recently conducted an in-depth analysis of your company's performance, and we recommend that you focus on driving revenue growth for one of your key brands, Kleding Man. Our assessment has identified several areas where addressing opportunities can lead to significant growth and improvement.

1. Brand Awareness: Team members have noted that brand awareness for Kleding Man has slipped in recent times, particularly within the older male demographic. Regaining a foothold in this market segment is crucial for future success.

2. Price Points: Our analysis indicates that Kleding Man's pricing is too high for younger men, the current target demographic. The inability to maintain constant sales is a direct result of the pricing being prohibitive for this demographic. Addressing this issue could lead to better customer retention and growth.

3. Declining Share of Revenue: Kleding Man's share of revenue has dropped 20%, approximately 11 percentage points or a 5.44% CAGR. Reversing this trend is vital for the brand's overall health and prosperity....

Cyborg practices

User BA2: Is that all? Can you make sure to revisit notes and make sure you added all important takeaways?

Generative AI: Kleding creates collections by defining and describing the target customer, including their age, activities, style, and hobbies.

The company shifted target demographics for Kleding Man to younger, trendier consumers. This strategy was a mistake as it led to low brand awareness among 18-30-year-old men and high prices.

Kleding Woman's switch in target customer has been successful, with good image and notoriety among 18-35-year-old women.

Kleding Kids is the smallest segment, and marketing efforts have been de-emphasized, leading to a decline in that segment.

Kleding Man's collection size was expanded from ~800 items to ~1600 items in 4 years, but it will now be reduced to around 900-1000 items by 2020.

The company plans to refocus on its original target customer of 35-40-year-old educated men and adjust store sizes accordingly.

Kleding Woman conducted an exercise to size the ideal store based on collection size, and a similar exercise could be beneficial for Kleding Man.

User BA2: There was a point made in the interview about how men's sales have decreased due to lack of kids sale. Add that to the above bullets as well.

Åbo Akademi University

Current capabilities

Åbo Akademi University

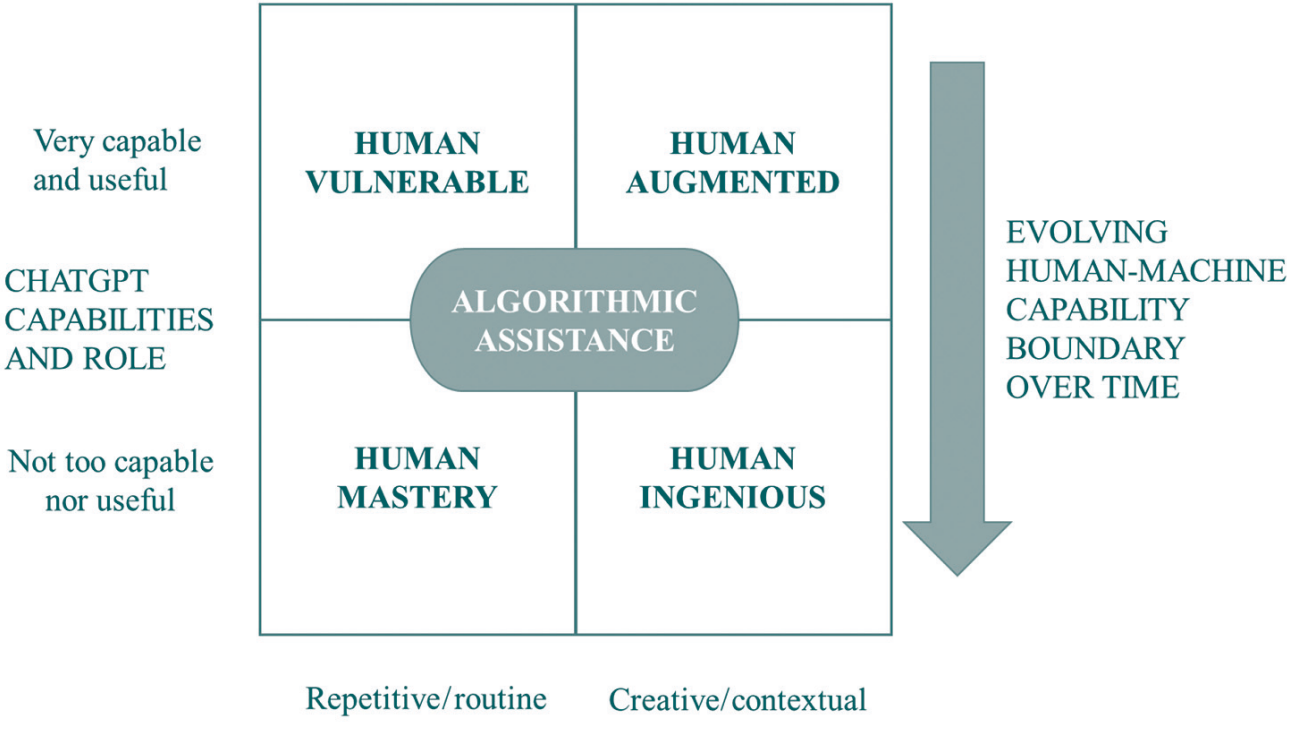
- Answer questions and generate output from prompts provided in natural language
- Serve as a search engine for inspiration, creativity and overviews on a wide range of topics
- Content production tool to generate new texts (early drafts)
- Content reorganization tool by iteratively working with users to refine textual materials for style and content

Current capabilities

Write, review and correct software code
Make learning something new much easier and faster
Handle routine tasks like responding to emails
Take over repetitive tasks
Generate text that is free of syntax or language errors
Replace Google Search
Access-real time information

Current Limitations

- Cannot answer complicated questions or offer unique insights into a topic; however, if prompted with new data, this limitation can be partially overcome
- Cannot provide consistently correct or unbiased answers or consistently answer questions that are asked multiple times
- Cannot meet strict privacy, compliance or ethical requirements



Source: Authors' own work

Åbo Akademi University

Boundaries of algorithmic assistance (Ritala et al., 2023)

Generating (business) ideas (Karan et al., 2023)

	Human Generated Ideas	ChatGPT-4	ChatGPT-4 trained with examples
N Ideas	200	100	100
Average Length of Description	63 words	69 words	71 words
Average Quality	0.404	0.468	0.493
Standard Deviation of Quality	0.112	0.108	0.120
Best Idea	0.64	0.70	0.75
Average Quality of Top Decile	0.62	0.64	0.66
Average Novelty of Top Decile	0.45	0.35	0.33
Fraction of the top decile of pooled ideas from this source	5/40	15/40	20/40
P-value (Is the average quality different?)		vs. humans <0.001	vs. humans <0.001 vs. baseline LLM 0.11

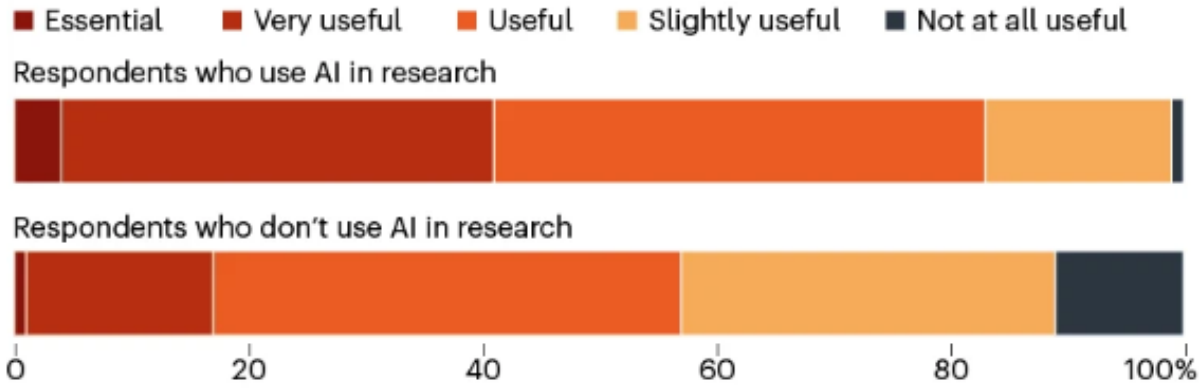
How LLMs are used in research?

Åbo Akademi University

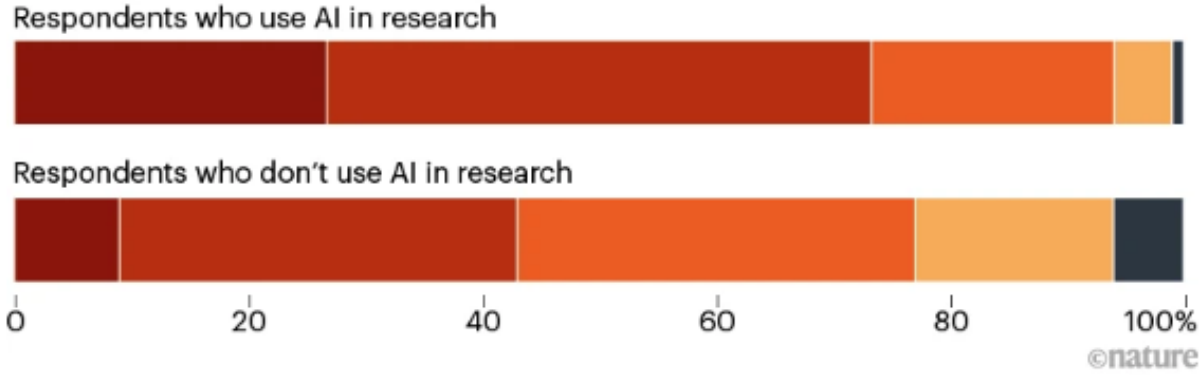
- Watkins, Ryan. "Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows." *AI and Ethics* (2023): 1-6.
- Guler, Nazmiye, Samuel Kirshner, and Richard Vidgen. "Artificial Intelligence Research in Business and Management: A Literature Review Leveraging Machine Learning and Large Language Models." *Available at SSRN 4540834* (2023).
- Van Noorden, Richard and Perkel, Jeffrey M. "AI and Science: what 1600 researchers think". https://www.nature.com/articles/d41586-023-02980-0

AI ANTICIPATIONS

Q: How useful do you think AI tools are for researchers in your field?



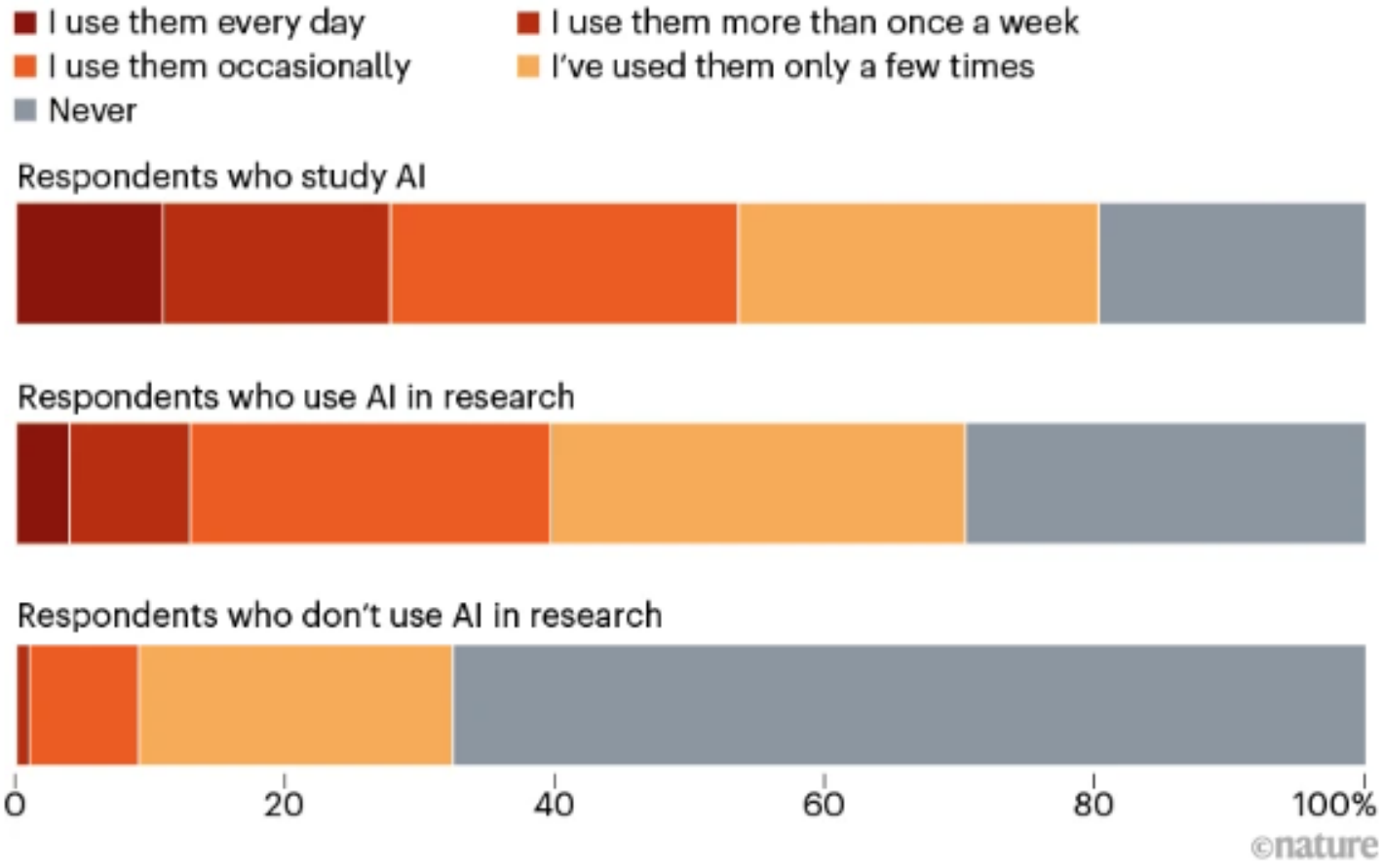
Q: How useful do you think AI tools will become for researchers in your field in the next decade?



Åbo Akademi University

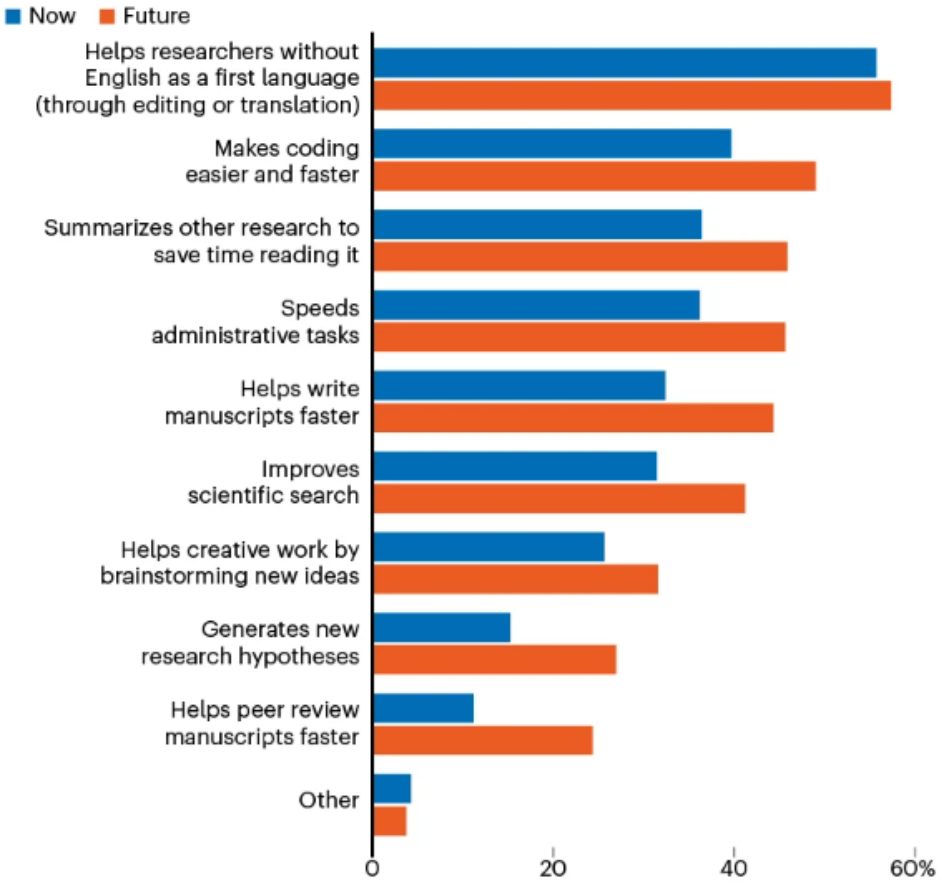
USING GENERATIVE AI

Q: How often do you use generative AI tools (such as ChatGPT) at work?



Åbo Akademi University

Q: What do you think are currently the biggest benefits of generative AI for research? In the future, where do you think generative AI will have the biggest beneficial impacts for research?



*1,659 respondents. For more on *Nature*'s survey, see go.nature.com/45232vd

Åbo Akademi University

How it can help in research?

- Idea generation and topic selection
  - Generate list of potential research topics/questions based on latest trendsand gaps in existing literature
- Literature review
  - Summarize articles, find relevant articles, pinpoint key findings
- Research design and methodology
  - Suggest research desgin by analysing similar studies and their outcomes
- Data collection
  - Designing surveys, experiments, help scraping public data from websites

How it can help in research?

- Data analysis
  - Perform preliminarydata analysis, suggest statistical techniques, interpret the results
- Writing the draft
  - Drafting sections of the paper, check for grammar and coherence
- Revision and editing
  - Suggest improvements to the draft, paraphrasing, help with adherence to journal guidelines
- Submission and peer review
  - Prepare a list of suitable journals, guide thorugh the submission process

<u>Study 1:</u> Making use of BERT LLM to understand customer journey on e-commerce platforms

Åbo Akademi University

Åbo Akademi University

Motivation

- **Popularity** of the online shopping
- **E-commerce sectors** and **COVID-19**
- The importance of customer interaction and customer satisfaction in online shopping
- Customer online reviews:
  - Provide valuable information
  - Help to understand the needs of customers

Natural Language Processing

Åbo Akademi University

- **Raw text** makes it **infeasible** and **impractical** for manual inspection
- Natural Language Processing (NLP) tools and machine learning techniques
  - Sentiment analysis
  - Topic modeling
  - Text summarization
  - Automated translation

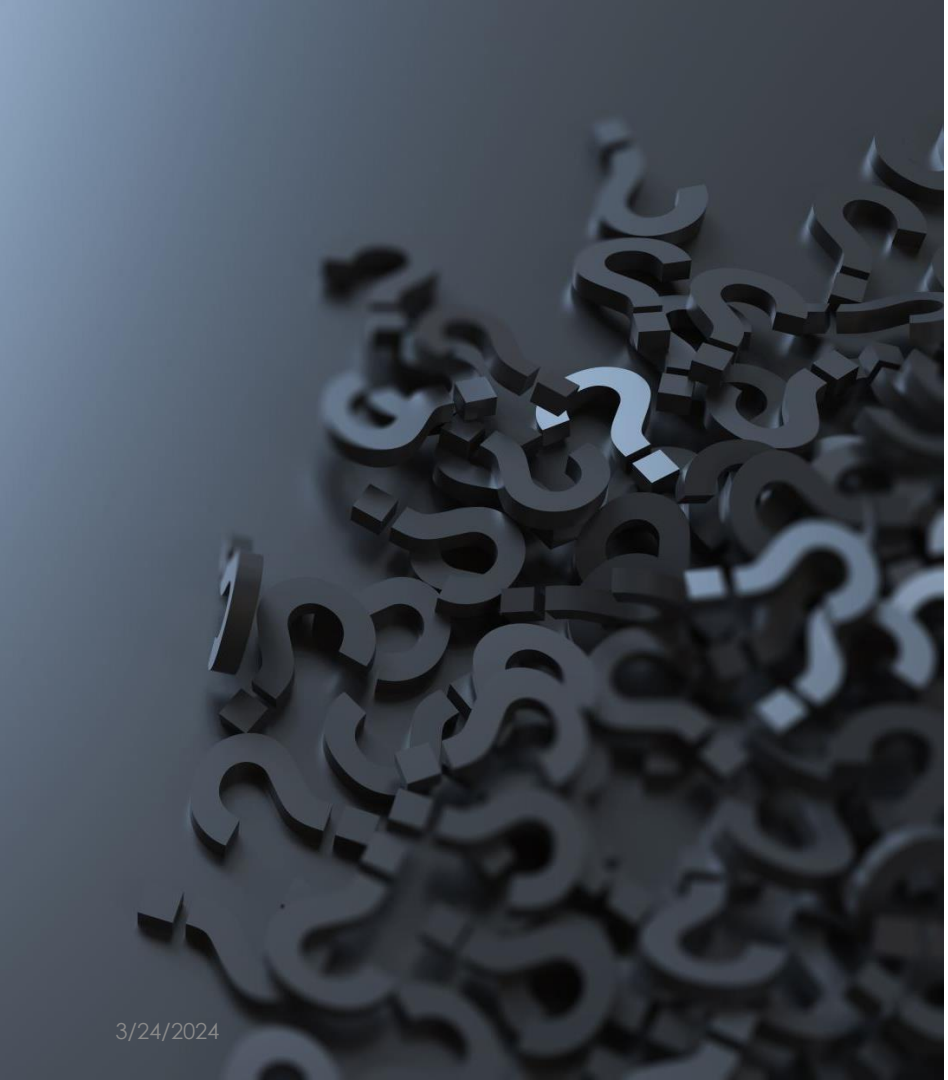
Goal

Åbo Akademi University

- Aspect extraction techniques and machine learning models over manually annotated customer reviews to enhance the understanding of customer opinions in e-commerce:
  - Identifying the **best-performing** model
  - To help e-commerce platforms
- **3500 user reviews** from **Trustpilot**
- The messages were randomly selected, and aspects were extracted and manually annotated using the polarities **'positive'** and **'negative'**

Sentiment Analysis

- Opinion Mining or Sentiment Analysis (SA)
- Most fundamental application of sentiment analysis:
  - Gather people's opinions e.g., customer reviews
- Target-based sentiment analysis (TBSA)/Aspect-based Sentiment Analysis (ABSA)
  - **Aspect/target term extraction (ATE)**
  - **Opinion term extraction (OTE)**
  - **Aspect/target term sentiment classification (ATC)**



Aspect extraction

Åbo Akademi University

- Objective of ATE:
  - **Identify and extract terms that represent aspects of a given sentence**
- ATE involves two sub-tasks:
  - **Extracting all aspect terms**
  - **Grouping aspect terms with similar meanings into categories**
  - Example: "*Excellent customer service and delivery*"

Aspect Extraction

Åbo Akademi University

- Deep learning models have become one of the most effective approaches for natural language processing tasks:
  - **Supervised training process**
  - **Large amounts of training data.**
  - **Problem: acquiring a large amount of supervised data can be a difficult and time-consuming process**
- Solution: transfer learning(e.g. BERT)
  - **allowing a model to be pre-trained on a large amount of unsupervised data**
  - **fine-tuned for a specific task under supervised conditions.**

Aspect Extraction

Åbo Akademi University

- **BERT** is a language model that can evaluate the context of a word from both the left and right sides simultaneously
- BERT uses the masked language modeling (MLM) technique
- The model then attempts to predict the masked word based on the context from both sides of the masked word
- Next-sentence prediction (NSP) task: which involves predicting whether a given sentence follows another sentence or not to capture the relationship between sentences

Aspect Extraction

Åbo Akademi University

- RoBERTa is one of the extensions of BERT
- Sufficient performances compared to BERT in various NLP tasks
- RoBERTa use transformer models as the main architecture
- RoBERTa is trained differently compared to BERT:
  - *Longer, with larger batches and data.*
  - *The next sentence prediction objective is removed*
  - *Using entire sentences as input*

Aspect Extraction

Åbo Akademi University

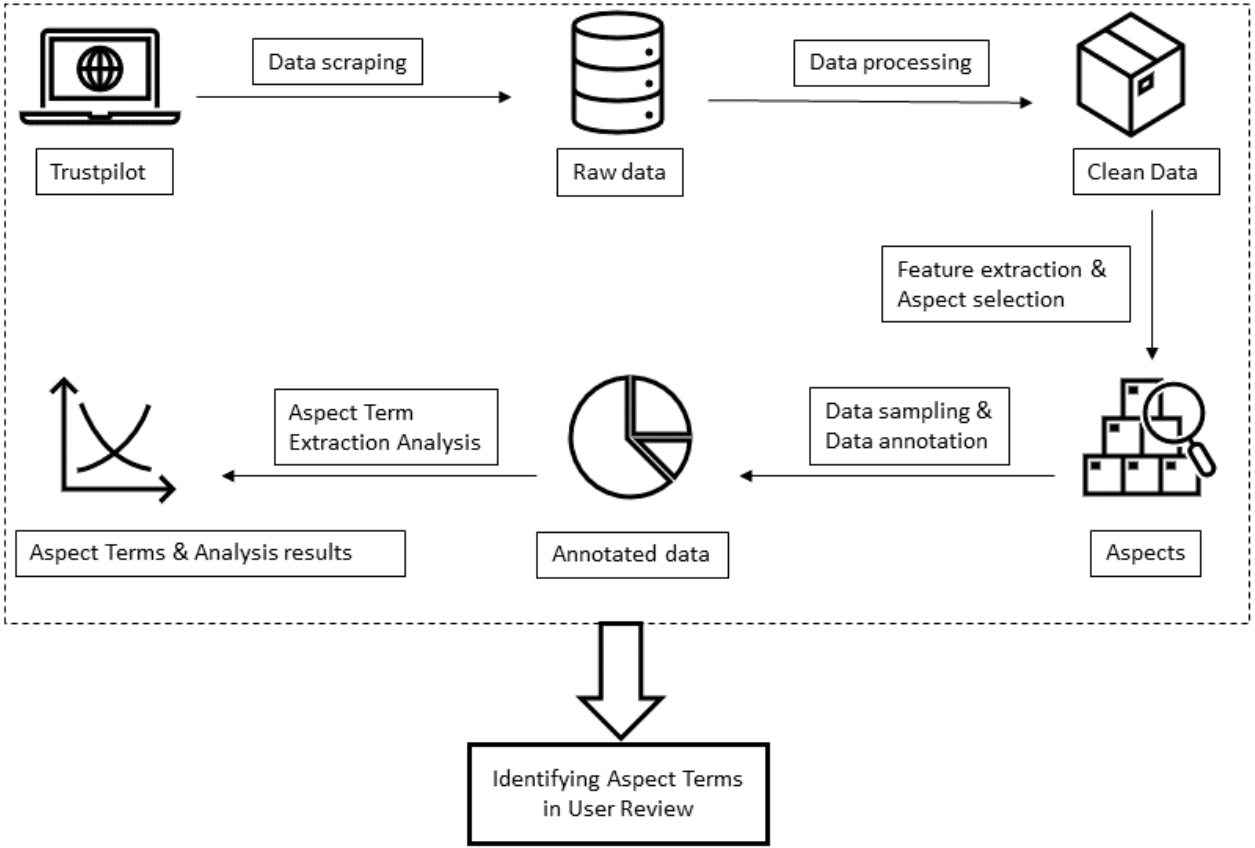
- **BERT_Review** is one of the extensions of BERT
- Sufficient performances compared to BERT in various NLP tasks
- BERT_Review use transformer models as the main architecture
- BERT_Review is designed to address Review Reading Comprehension (RRC) :
  - **Post-training method using BERT to improve the fine-tuning process for RRC**
  - **Post-training approach is used to other review-based tasks, such as aspect extraction and aspect sentiment classification**

Aspect Extraction

Åbo Akademi University

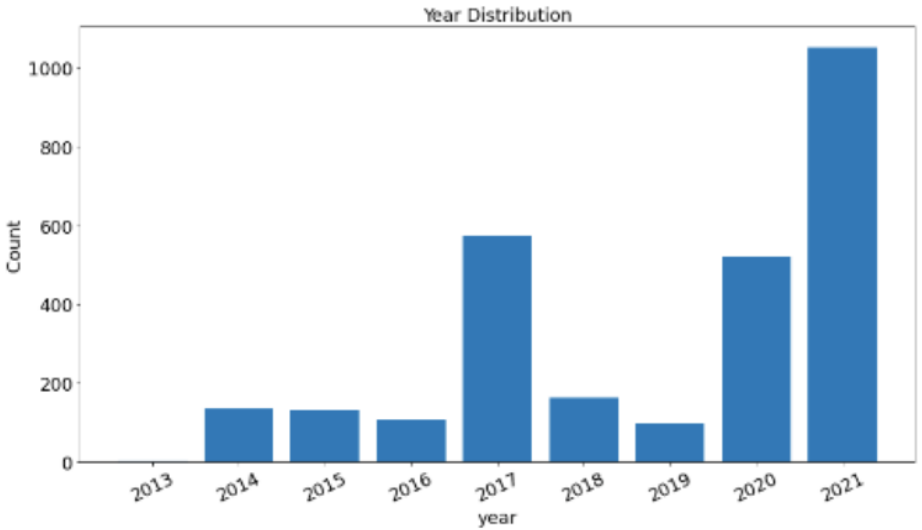
- **BERT and RoBERTa** are among the most frequently employed models for aspect extraction in user reviews (Chauhan et al., 2020; Tian et al., 2020; Yanuar et al., 2020; Lopes et al., 2021)
- The F1 score is achievable in the range of *0.738-0.85* depending on the domain and the language of the reviews

Methodology



Åbo Akademi University

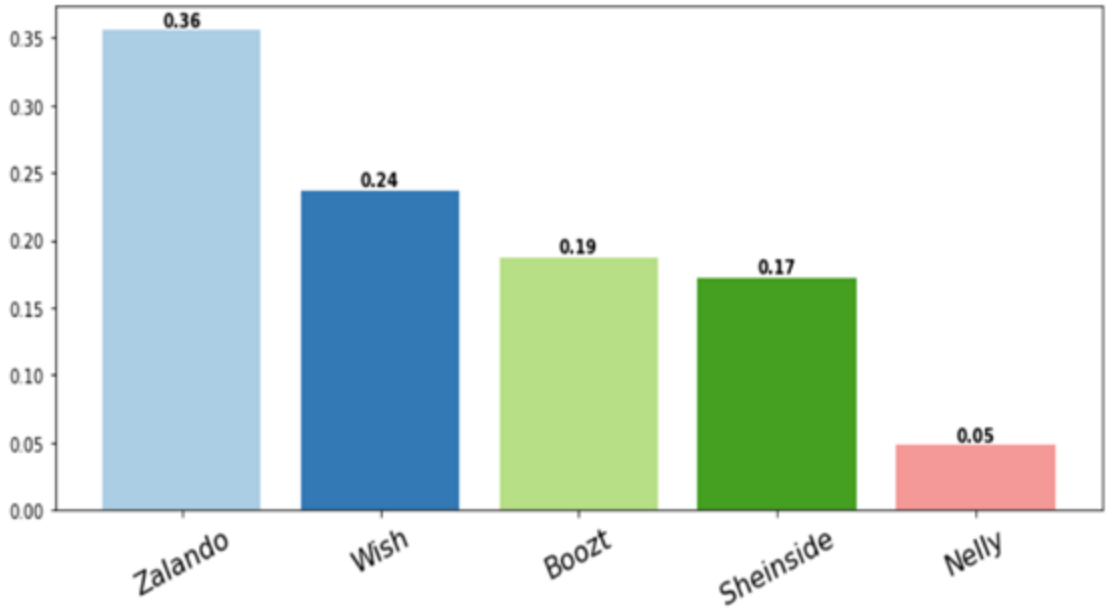
Data collection



- Reviews written in **English** from **Trustpilot**
- Shops that have **only online stores**

Data collection

Åbo Akademi University



Data annotation

Åbo Akademi University

.

	Aspect	Definition
1	Shipping	Quality of the delivery e.g. cost and timeline
2	Trust	Customers' general opinion about the store
3	Item quality	Products' quality
4	Customer service	Quality of customer's direct interaction with store's representative
5	Pricing	Price offerings, availability of discounts and campaigns
6	Product features	Quality of product's image and size guide provided on the website
7	Refund process	Refund speed and quality of handling refund issues
8	Return process	Speed, convenience and cost of the return
9	App experience	User experience in interacting with store's website
10	Delivered product status	Condition of delivered products e.g broken, smelling
11	Information	Availability and quality of the information e.g. misleading ads
12	Packaging	Attractiveness and quality of the packaging
13	Payment	Quality of financial transaction
14	Product availability	Variety of offering products or brands

Data annotation

Åbo Akademi University

- ***“Love this site fantastic saving for quality stuff.“***
- The text is first transformed into a set of words
- Then the annotators assign corresponding labels to each word.
- The label sequence for this example is as follows:
- ['O', 'O', 'Trust_B', 'O', 'Pricing_B', 'O', 'Item_quality_B', 'Item_quality_I', 'O']

Machine learning models

Åbo Akademi University

- Three transformer-based machine-learning models:
  - **BERT, RoBERTa, and BERT_Review**
- We performed five-fold cross-validation to get a more reliable estimation of the models' performance:
  - Dataset was split into training and validation sets by using a 5-fold cross-validation
  - with ten steps of iteration per fold
  - To calculate the aggregate summary of model performance:
    - calculate the mean value across all folds for each performance metric

Results

Åbo Akademi University

Method	Loss-validation set	Accuracy-validation set	F1-validation set	Execution time (in seconds)
BERT_Review	**0.104**	**0.972**	**0.841**	**3,077**
BERT	0.113	0.969	0.829	3,075
RoBERTa	0.109	0.969	0.828	2,905

Results

Åbo Akademi University

- Misclassifications generated by the BERT_Review model:
- Firstly, identifying neutral aspects in the messages
  - ***"I ordered around 300 dollars worth of clothes for my kids and I need to return 100 dollars worth of clothes."***
  - This review was misclassified by the model as a *return process*.
  - No aspects in this message
  - The sentiment of the *return process* is neutral:
- Secondly, less frequent aspects such as payment.
- The model did not have sufficient samples to accurately identify these aspects.

Discussion

Åbo Akademi University

- Results revealed that the BERT_Review model outperformed the other two models under consideration
- Among all categories, the *Trust_B* aspect attained the highest F1 score of 0.92
- *Payment_I* had the lowest score of 0.4
- Inclusion of infrequent aspects may negatively impact the model's performance
- One potential solution: Automatically generate labeled data for infrequent aspects

Discussion

Åbo Akademi University

- Previous research in this area used datasets: SemEval 2014, SemEval 2015, and SemEval 2016 (Wang et al, 2021; Dai et al., 2019) for their experimental studies
- Our study focused on e-commerce businesses that do not have physical shops when collecting data
- We aimed to identify the most prevalent sources of satisfaction and dissatisfaction
- The comprehensiveness and broadness of this data help to analyze smaller online retailers and comparable enterprises

Discussion

Åbo Akademi University

- Automated annotations are less time-consuming and costly than manual annotations, but they are generally less accurate.
- We decided to perform manual annotation to gain:
  - A comprehensive understanding of the review content
  - Find its relevance to the target companies
  - Our manually annotated dataset includes extracted aspect terms with negative and positive sentiments:
    - Companies can ensure that they concentrate on the correct elements by obtaining precise aspect detection results.
      - The combination of sentiment classification of the extracted aspects can enable automated and accurate identification of the sources of dissatisfaction.

Future

- All three models performed well, with potential for further improvement
- Adding more samples to the dataset, particularly for infrequent aspects such as payment and packaging.
- Using manually annotated datasets to generate automated labeled data
  - E.g., by fine-tuning Large Language Models like GPT-3

Åbo Akademi University

Study 2: Exploring the Performance of Large Language Models for Data Analysis Tasks through the CRISP-DM Framework

Åbo Akademi University

Motivation and Objective

Åbo Akademi University

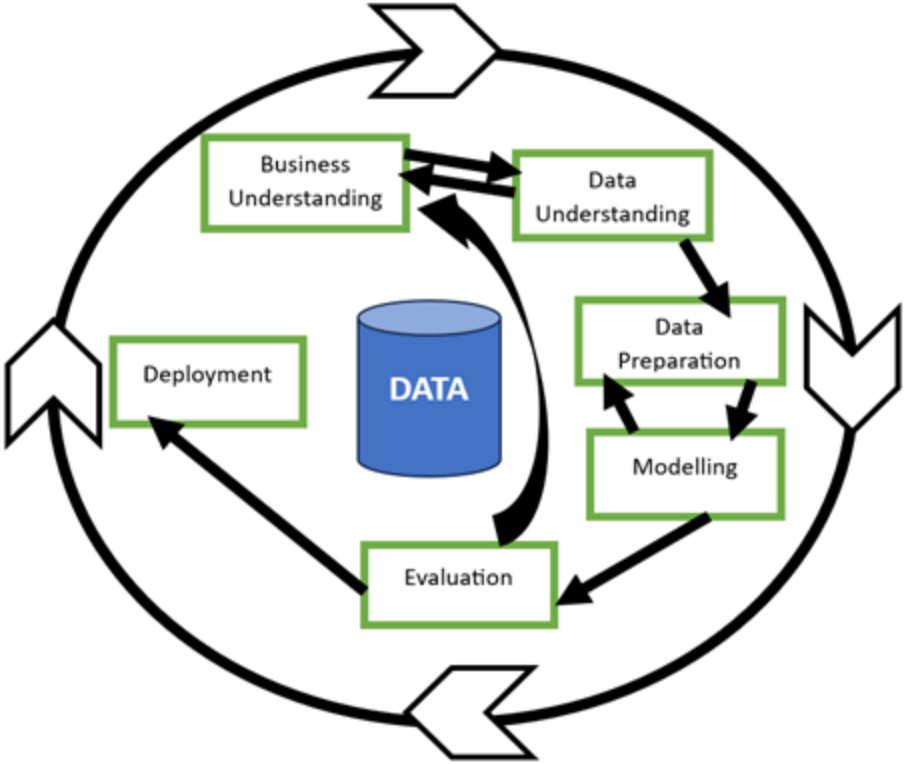
- AI impacts employment, and it creates new jobs and professions
- Studies research skill in data-related professions. However,...
- LLMs increase the efficiency of workers, and have text-to-code capabilities
- Objective: to delineate the shifts in skills demanded following the introduction of text-to-code language models and the consequent automation of data-related tasks.

To what extent and how efficiently do LLMs, such as GPT, perform data-related tasks, and how does this impact the expected skill set of data analysts?

CRISP-DM Framework

Åbo Akademi University

Most of the tasks performed on a regular basis by a data analyst can be classified into the stages of the CRISP framework



Large Language Models (LLM) for coding

Åbo Akademi University

- Transformer architecture, efficiency in computational processing and identification of long-range relationships between words, other successful pretrained models
- The models, including GPT Codex and the latest GPT-4, have been studied in the context of coding
- LLMs' capabilities from the perspective of data analysis: mathematical knowledge, creativity and logical thinking, within the mid-term horizon, testing and enhanced insight vs coding skills

Methodology

Åbo Akademi University

- Tasks that closely align with the curriculum of a Master's program specializing in the impact of digitalization on organizations
- Results of the research are compared, studied and generalized from the CRISP-DM framework perspective
- ChatGPT's performance center on the accuracy of the responses, correctness of the solutions produced, but also whether ChatGPT adheres to the requisite problem-solving steps
- ChatGPT 3.5 and ChatGPT 4 (advanced data analysis)

Methodology

Åbo Akademi University

- *Course 1*: a general introduction to programming concepts using Python. The assignment tasks encompass the essential principles of programming
- *Course 2*: programming tools for data analysis in Python. The assignment tasks are related to data visualization, data manipulation, data preprocessing, as well as the basics of building and evaluating machine learning models
- *Course 3*: advanced machine learning concepts, NLP and applications with Python. The assignment tasks within this course challenge students with more complex and specialized problem-solving scenarios
- *Course 4*: data management and SQL. Students learn about entity-relationship (ER) diagrams, and querying databases with SQL. The assignment tasks in this course emphasize practical querying skills

Results - Data Understanding & Preparation

Åbo Akademi University

- *Data description*: the tool effectively utilized Python libraries like Pandas, and Numpy. Specifically, the GPT models have no problem in understanding what summary statistics to generate for different variable types.
- *EDA tasks*: the models were adept at creating visual representations of data (e.g., in Matplotlib, Seaborn).
- *Data quality verification*: the GPT models can assess the completeness and correctness of the data.

Results - Data Understanding & Preparation

Åbo Akademi University

- Aptly handled selecting and cleaning data, integrating data, and converting the data to suitable format for modeling.
- Perform operations like data cleaning, transformation, using NLTK for tasks (e.g. tokenization).
- Code interpreter was correct in all tasks, while the GPT Text-to-code model was only partially correct in some instances of descriptive analysis. For instance, number of visualizations, or incorrect use of the correlation measure.

Results - Modeling

Åbo Akademi University

- Strong capabilities in complex tasks like association rule mining and classification (e.g. Sklearn).
- ML tasks, regression analysis, statistical modeling
- Also advanced modeling and performance evalution techniques used
- Some of the steps and guidance were missed, clear instructions were required

Results - Modeling

Åbo Akademi University

- *When was human intervention required?*

- Unavailable modules in ChatGPT (ADA) model. (nltk stopwords, tokenization vs split function).
- Defining stopwords, or NaN is missing or can be replaced with 0. i.e. domain knowledge, critical thinking and decision-making may be required in some cases.
- Although the models used Grid Search, human engagement may contribute in defining parameters and metrics that are most suitable for the task.

General observations

Åbo Akademi University

- Areas of partial correctness - GPT Text-to-code model in advanced modeling tasks, certain aspects of descriptive analysis, SQL (e.g. involving working with only a single table vs requiring joining the tables).

- Some assignments involve detailed description of the methods and stages as a guide for performing the tasks. These also underscores the need for human oversight, especially in complex or critical analytical scenarios.

- Most of the GPT failures (e.g., incorrect answers, errors, incompleteness) have been addressed by additional questions or clarifications. However, ChatGPT (ADA) enables running codes over the datasets, and automatically reinitiating the script.

General observations

Åbo Akademi University

- ChatGPT (ADA) enabled interpretations of the results and performing and finalizing the interdependent tasks. i.e., a shift to less human involvement can be observed, whereas it is still crucial.

- Currently time spent by professionals: data preparation may constitute 50-90%

- GPT can change this by exempting from regular routine tasks, and professionals can focus on logical thinking and creativity

Åbo Akademi University

Thank You!