Natural Language Processing

Sonja Aits

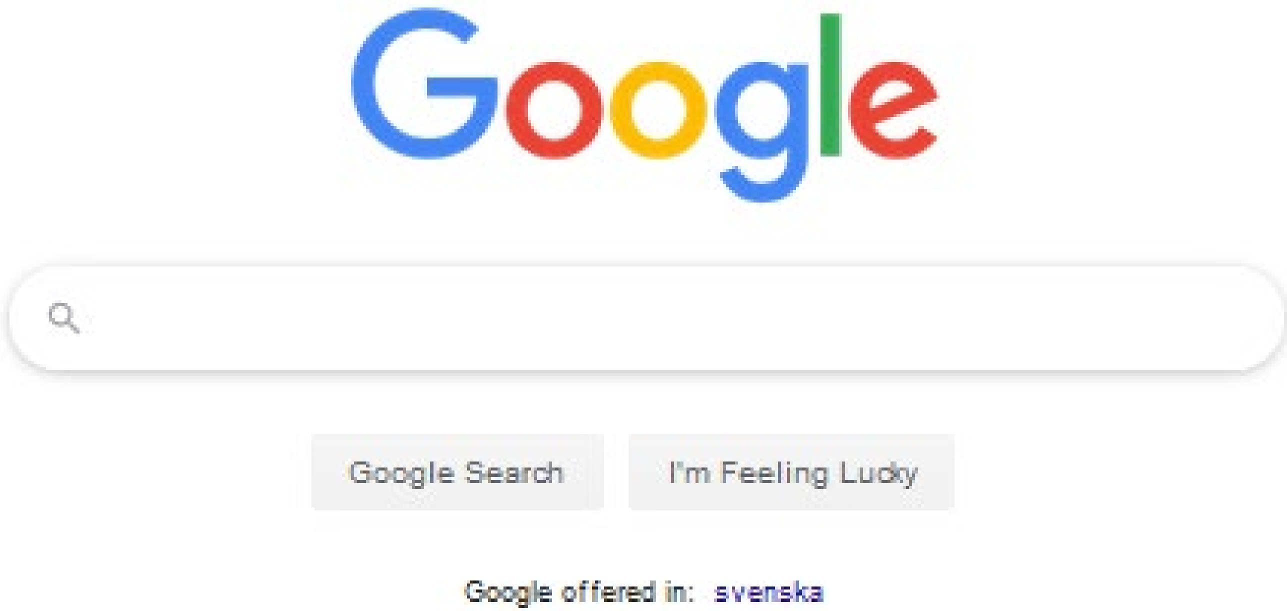
Lund University, 20231213

Natural Language Processing (NLP)

= Computational analysis and generation of natural language (text or speech)

BioNLP

= NLP related to medicine and life sciences


Google
Google Search
I'm Feeling Lucky
Google offered in: svenska

Computer vision tasks

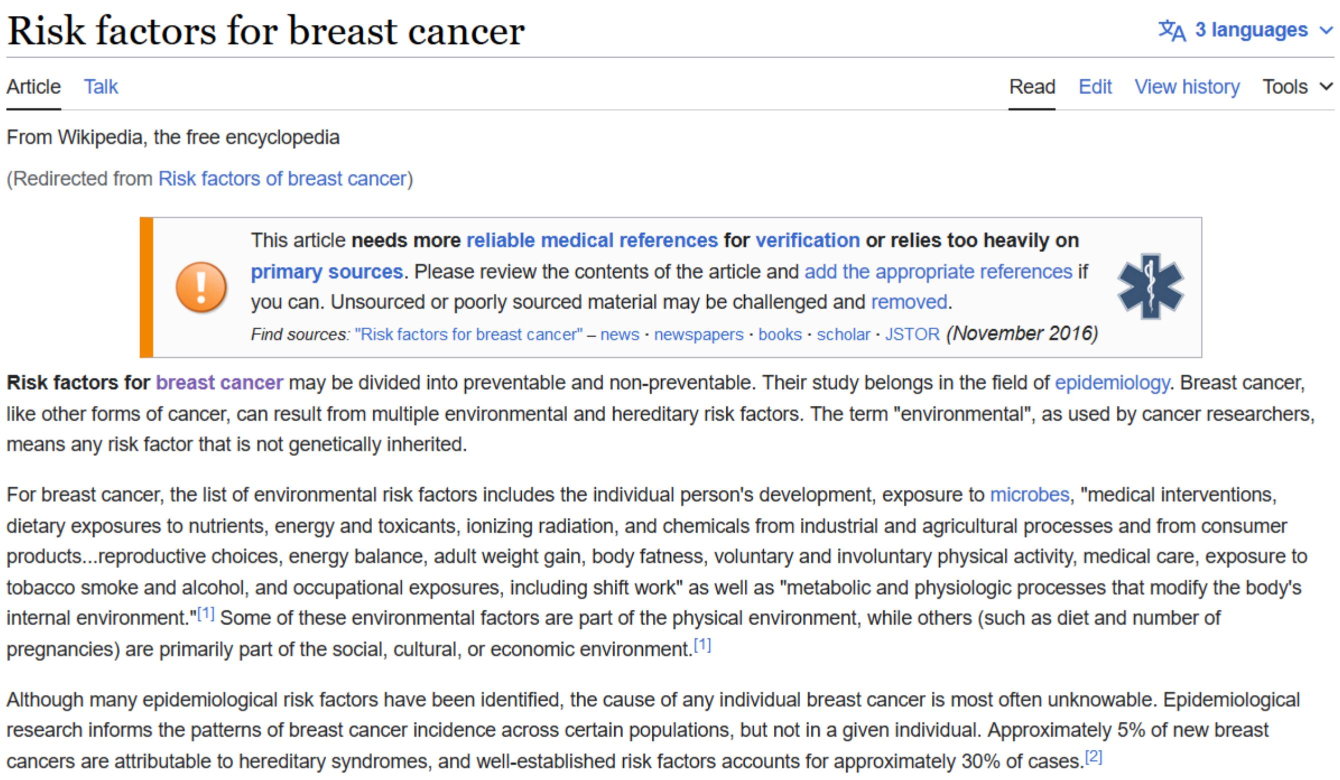
Examples of text and speech used in medicine and life sciences?

Language datasets

- Patient-doctor conversations
- Emergency calls
- Electronic health records
- Literature – abstracts or fulltext
- Reports
- Patents
- Social media posts

…

Summarization



Risk factors for breast cancer

From Wikipedia, the free encyclopedia

(Redirected from Risk factors of breast cancer)

This article **needs more reliable medical references for verification or relies too heavily on primary sources**. Please review the contents of the article and add the appropriate references if you can. Unsourced or poorly sourced material may be challenged and removed.

Find sources: "Risk factors for breast cancer" – news · newspapers · books · scholar · JSTOR *(November 2016)*

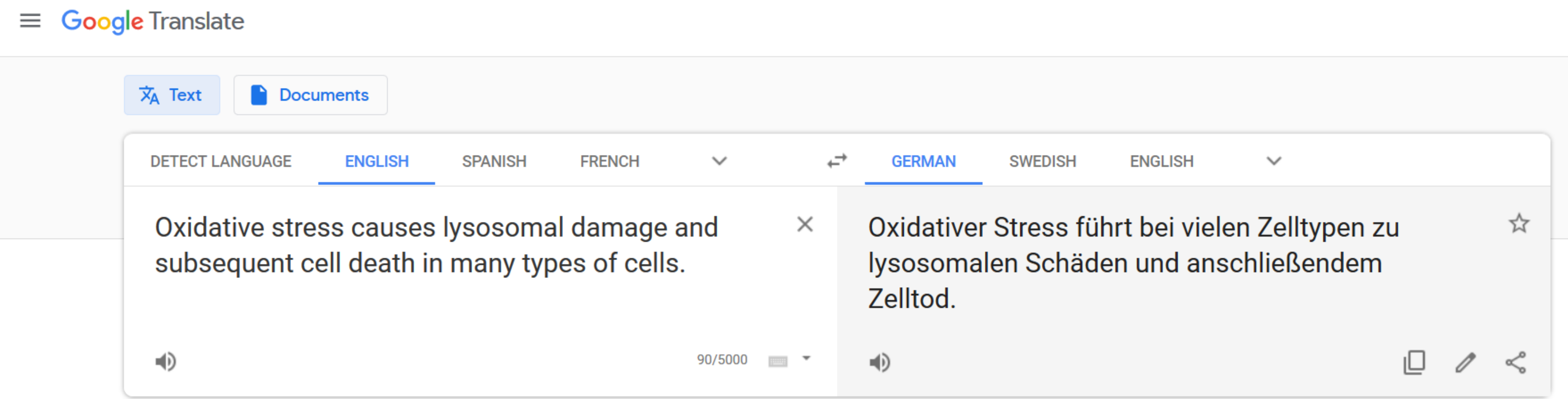
Risk factors for breast cancer may be divided into preventable and non-preventable. Their study belongs in the field of epidemiology. Breast cancer, like other forms of cancer, can result from multiple environmental and hereditary risk factors. The term "environmental", as used by cancer researchers, means any risk factor that is not genetically inherited.

For breast cancer, the list of environmental risk factors includes the individual person's development, exposure to microbes, "medical interventions, dietary exposures to nutrients, energy and toxicants, ionizing radiation, and chemicals from industrial and agricultural processes and from consumer products...reproductive choices, energy balance, adult weight gain, body fatness, voluntary and involuntary physical activity, medical care, exposure to tobacco smoke and alcohol, and occupational exposures, including shift work" as well as "metabolic and physiologic processes that modify the body's internal environment."[1] Some of these environmental factors are part of the physical environment, while others (such as diet and number of pregnancies) are primarily part of the social, cultural, or economic environment.[1]

Although many epidemiological risk factors have been identified, the cause of any individual breast cancer is most often unknowable. Epidemiological research informs the patterns of breast cancer incidence across certain populations, but not in a given individual. Approximately 5% of new breast cancers are attributable to hereditary syndromes, and well-established risk factors accounts for approximately 30% of cases.[2]

Common risk factors for breast cancer are age, female sex,hereditary factors (e.g. BRCA1/2 status) and life style factors (e.g. obesity and alcohol).

Translation

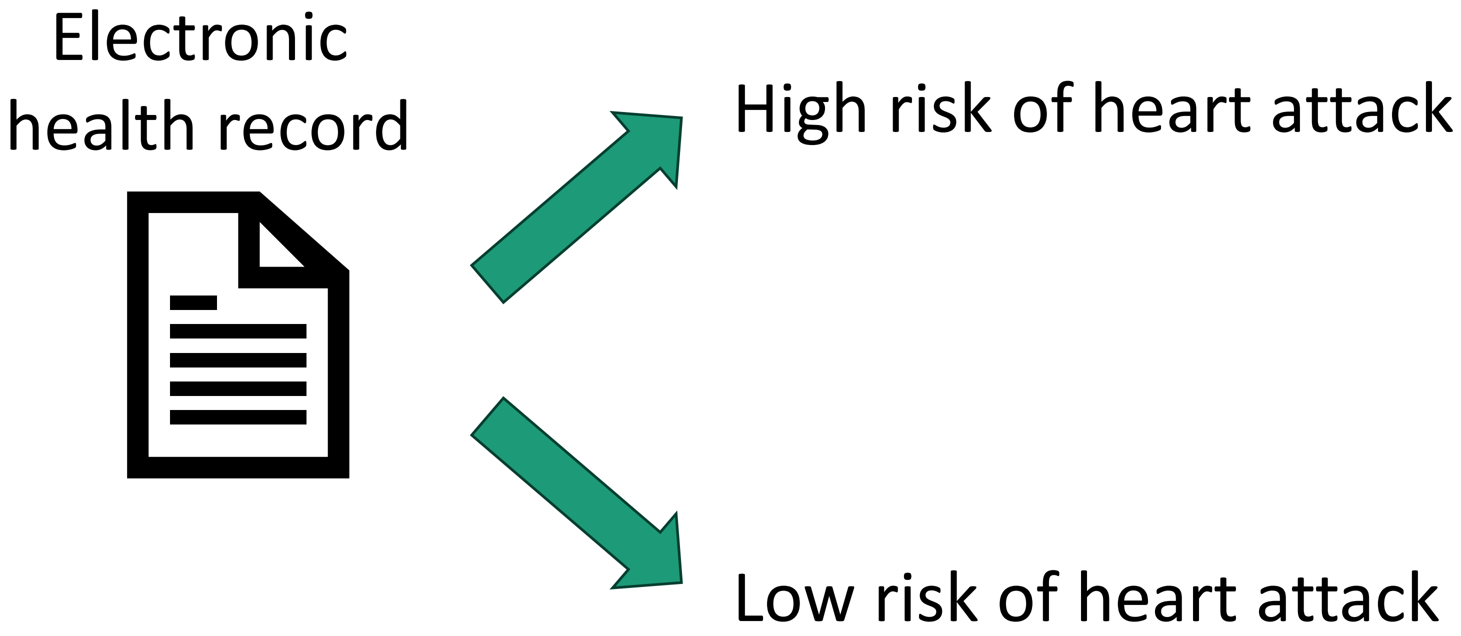


Oxidative stress causes lysosomal damage and subsequent cell death in many types of cells.

Oxidativer Stress führt bei vielen Zelltypen zu lysosomalen Schäden und anschließendem Zelltod.

Text clustering

Text classification

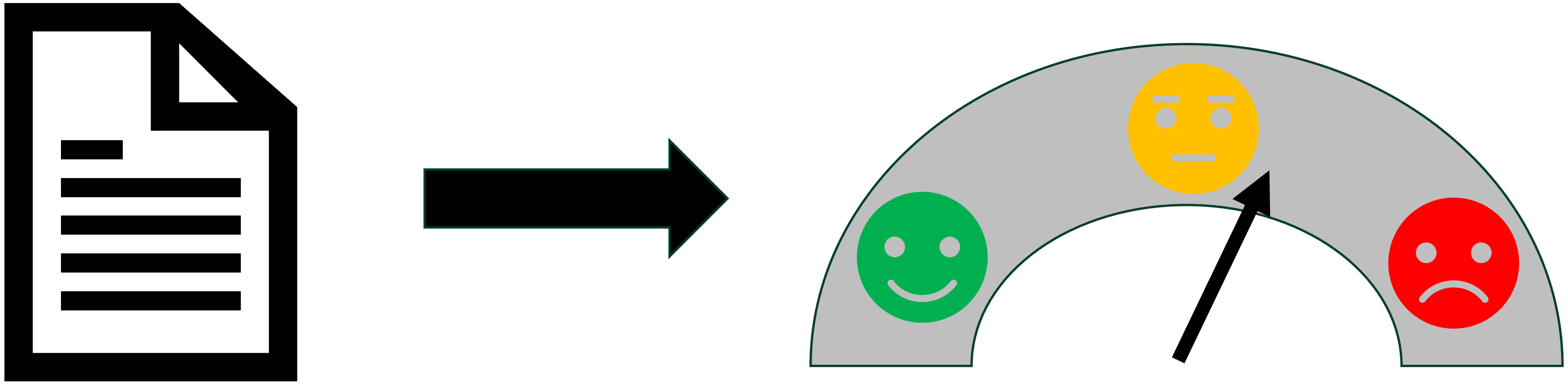


Electronic health record

High risk of heart attack

Low risk of heart attack

Sentiment analysis



Named entity recognition (token classification)

Disease **Drug/treatment** **Gene/protein** **Process/location** **Relation**

PLoS One. 2012;7(10):e45381. doi: 10.1371/journal.pone.0045381. Epub 2012 Oct 11.

Identification of cytoskeleton-associated proteins essential for lysosomal stability and survival of human cancer cells.

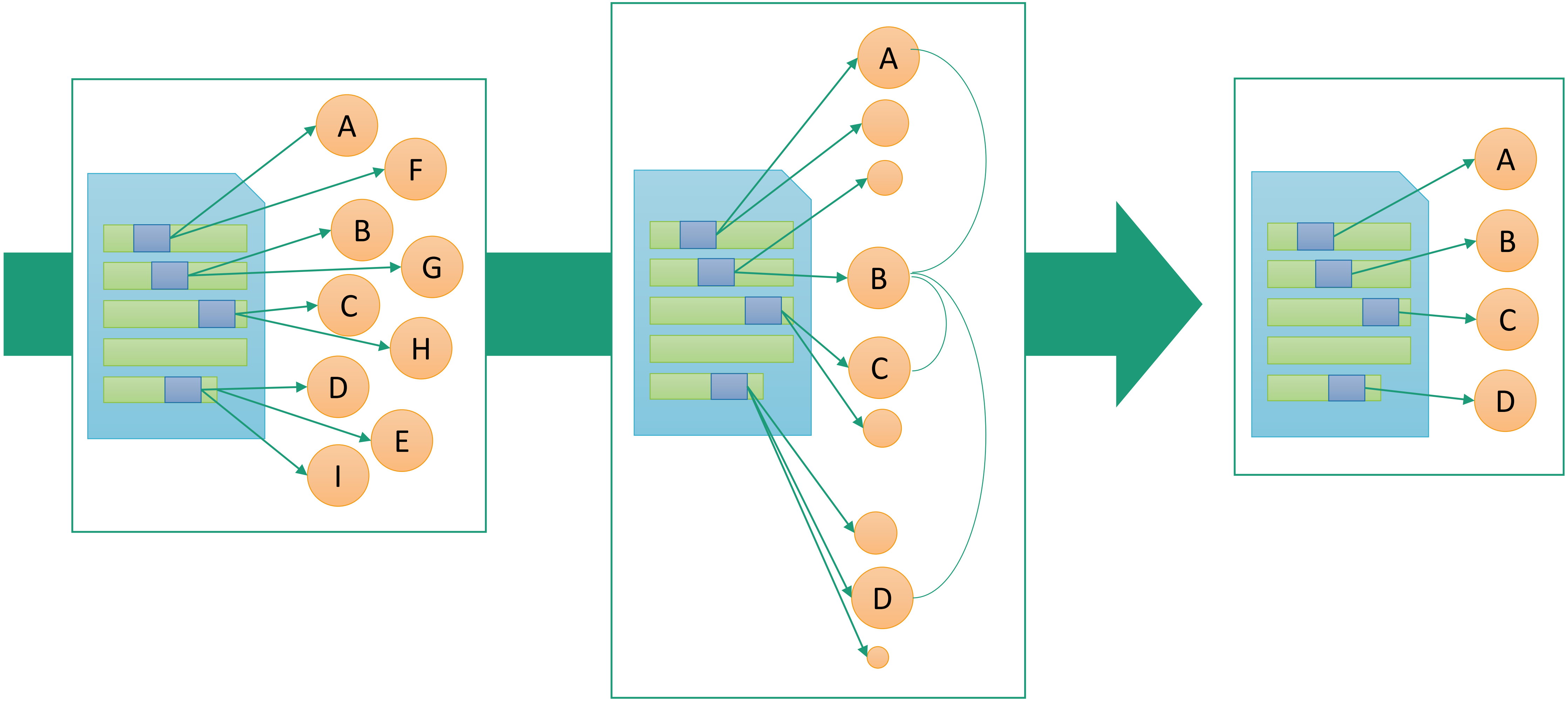
Groth-Pedersen L[1], Aits S, Corcelle-Termeau E, Petersen NH, Nylandsted J, Jäättelä M.

Author information

Abstract

Microtubule-disturbing drugs inhibit lysosomal trafficking and induce lysosomal membrane permeabilization followed by cathepsin-dependent cell death. To identify specific trafficking-related proteins that control cell survival and lysosomal stability, we screened a molecular motor siRNA library in human MCF7 breast cancer cells. SiRNAs targeting four kinesins (KIF11/Eg5, KIF20A, KIF21A, KIF25), myosin 1G (MYO1G), myosin heavy chain 1 (MYH1) and tropomyosin 2 (TPM2) were identified as effective inducers of non-apoptotic cell death. The cell death induced by KIF11, KIF21A, KIF25, MYH1 or TPM2 siRNAs was preceded by lysosomal membrane permeabilization, and all identified siRNAs induced several changes in the endo-lysosomal compartment, i.e. increased lysosomal volume (KIF11, KIF20A, KIF25, MYO1G, MYH1), increased cysteine cathepsin activity (KIF20A, KIF25), altered lysosomal localization (KIF25, MYH1, TPM2), increased dextran accumulation (KIF20A), or reduced autophagic flux (MYO1G, MYH1). Importantly, all seven siRNAs also killed human cervix cancer (HeLa) and osteosarcoma (U-2-OS) cells and sensitized cancer cells to other lysosome-destabilizing treatments, i.e. photo-oxidation, siramesine, etoposide or cisplatin.

Named entity linking



Entity candidates

Entity ranking

Linked entities

Relation extraction

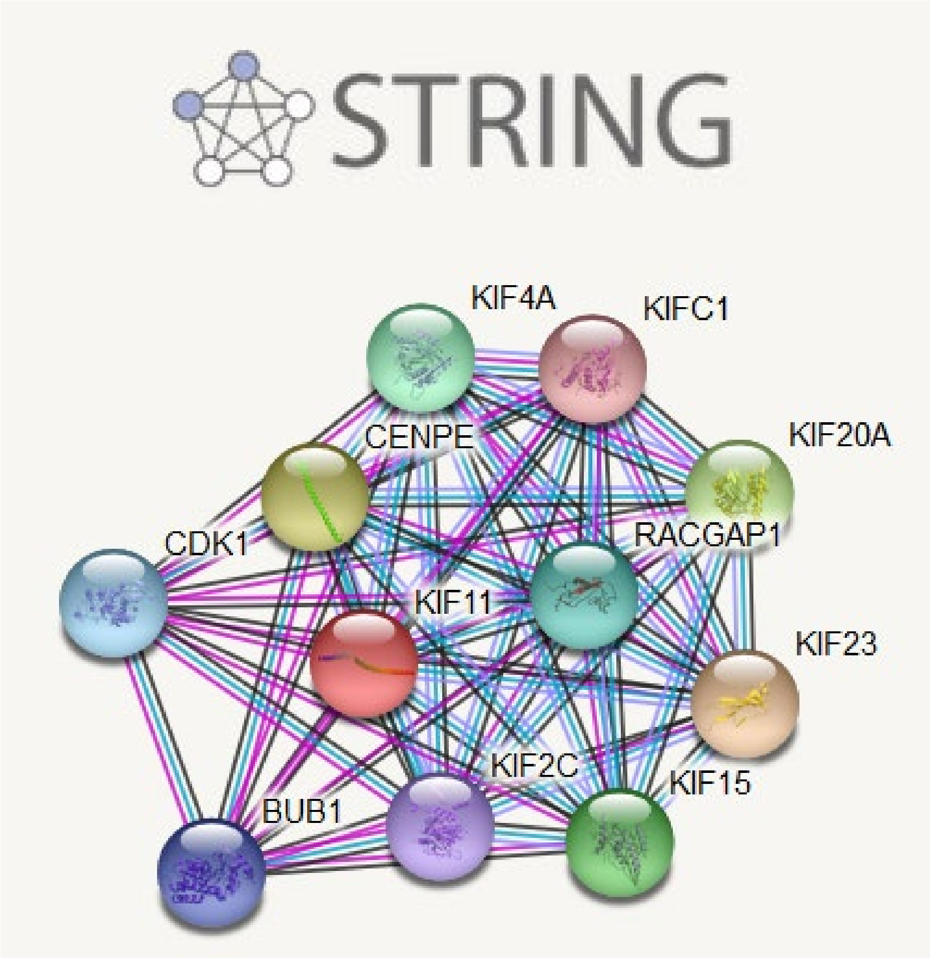
Rapamycin inhibits mTOR, and thereby affects autophagy.

Entity 1: Rapamycin

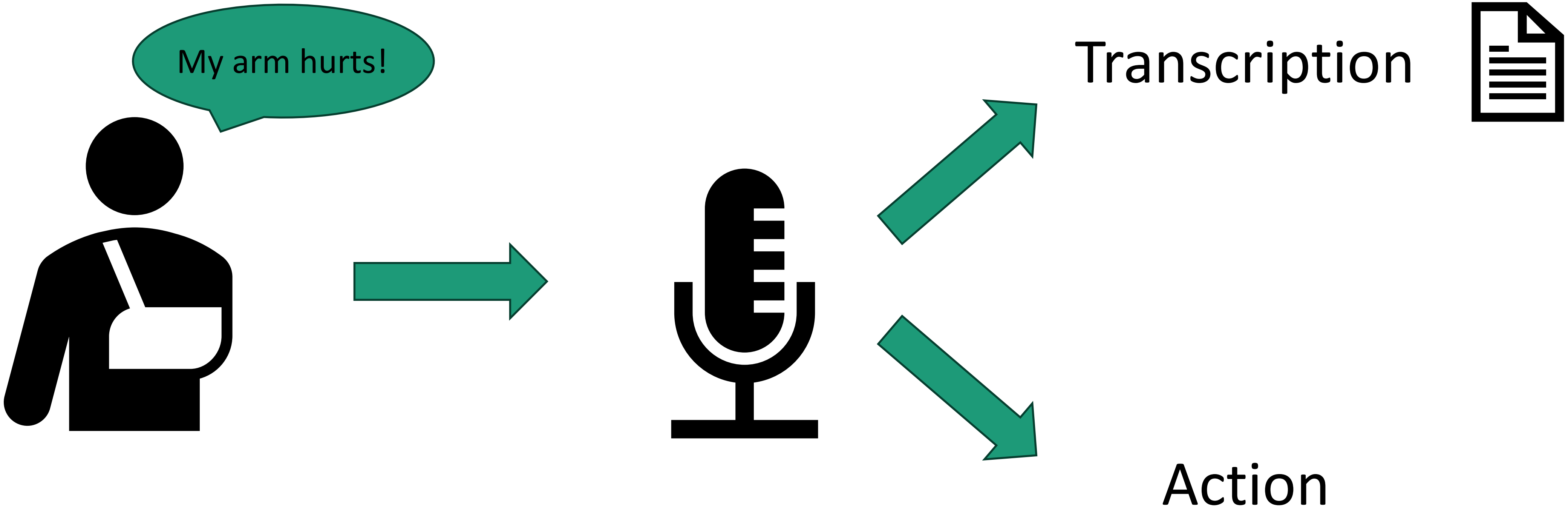
Entity 2: mTOR

Relation: inhibition

Relation extraction is used to populate protein-protein interaction databases



Speech recognition



My arm hurts!

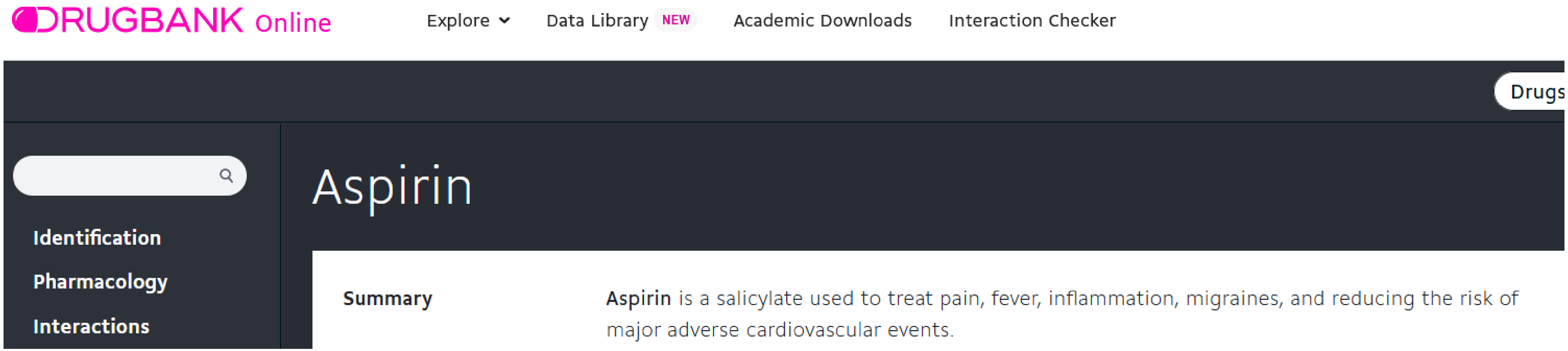
Transcription

Action

Question answering

Question: Which disease conditions are treated with aspirin?

Context

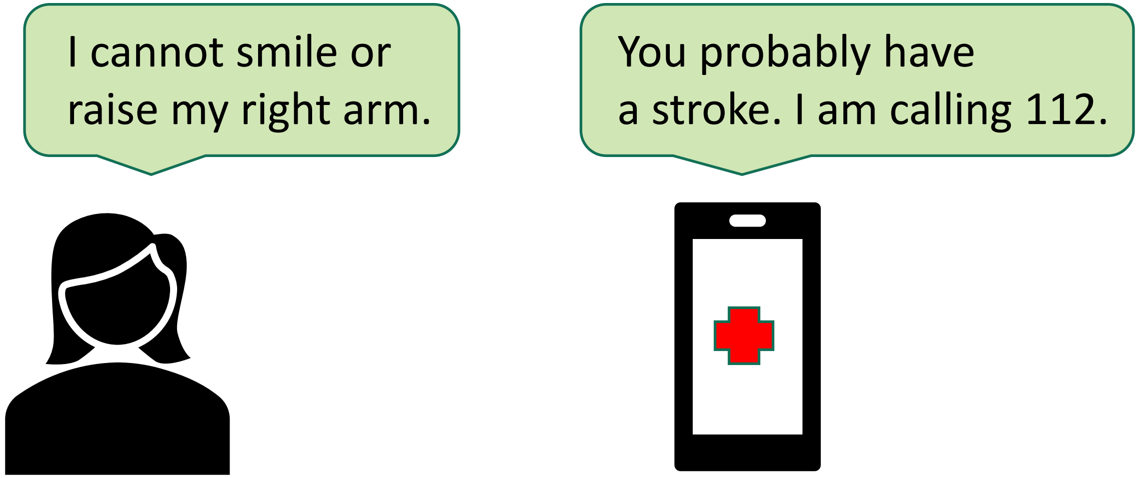


Aspirin

Summary	**Aspirin** is a salicylate used to treat pain, fever, inflammation, migraines, and reducing the risk of major adverse cardiovascular events.

Answer: Pain, fever, inflammation and migraines are treated with aspirin.

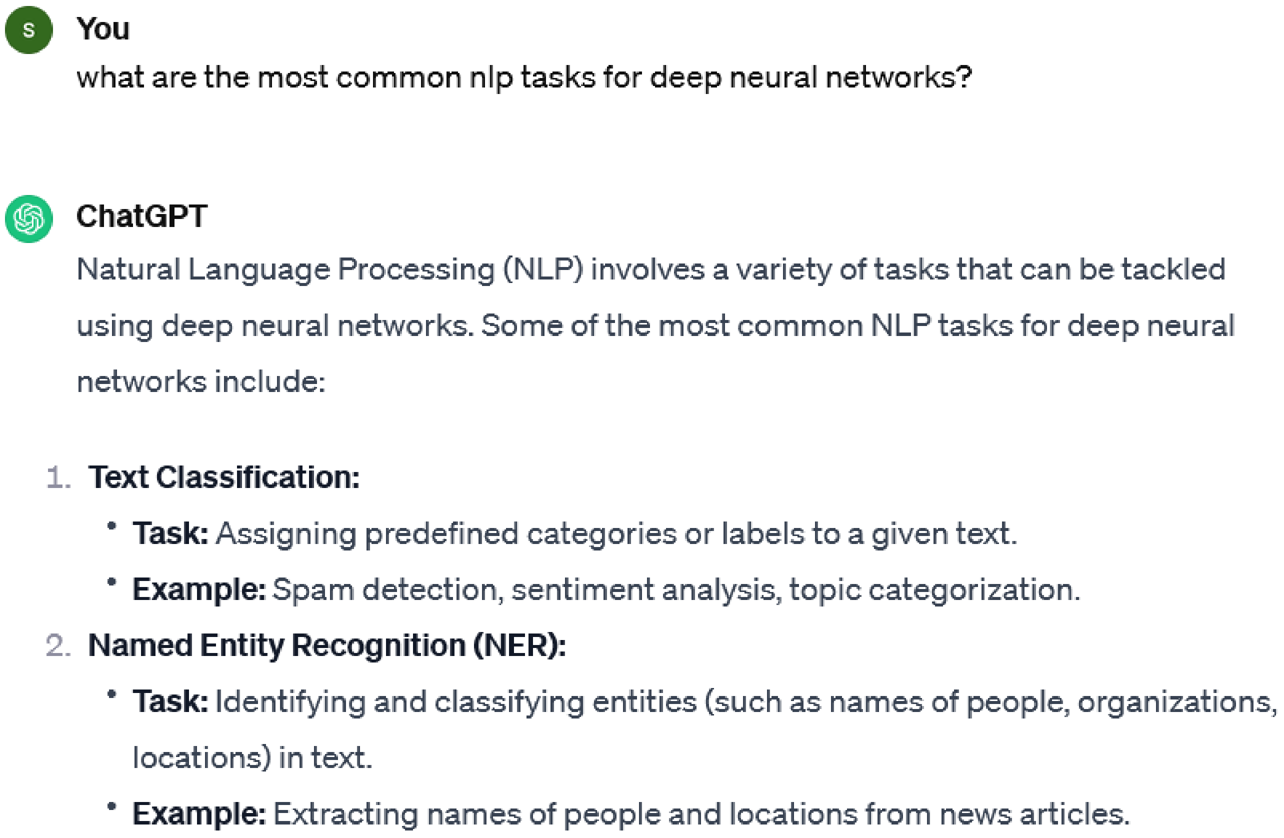
Conversational AI: Chat bots



I cannot smile or
raise my right arm.

You probably have
a stroke. I am calling 112.

Conversational AI: Chat bots



You

what are the most common nlp tasks for deep neural networks?

ChatGPT

Natural Language Processing (NLP) involves a variety of tasks that can be tackled using deep neural networks. Some of the most common NLP tasks for deep neural networks include:

1. **Text Classification:**
   - **Task:** Assigning predefined categories or labels to a given text.
   - **Example:** Spam detection, sentiment analysis, topic categorization.
2. **Named Entity Recognition (NER):**
   - **Task:** Identifying and classifying entities (such as names of people, organizations, locations) in text.
   - **Example:** Extracting names of people and locations from news articles.

Mini quiz: NLP tasks

- Matching electronic health records to ICD-10 disease codes
- Identifying the major types of adverse effects of a new vaccine from reports submitted by doctors
- Finding all links between mutations and diseases reported in the literature
- Flagging high priority patients in emergency calls
- Building a treatment recommender that suggests drugs based on mutation status in sequencing reports

The technical side…

NLP is a challenging task!

She killed the man with the tie.

Why is NLP challenging?

- **Ambiguity**
  - Homographs: We ran a Western blot to measure RAN levels.
  - Homophones: to, two, too
- **Co-reference**

  Cancer is one of the most common causes of death. This disease…
- **Synonymous expressions**
  - This caused cell death./This led to cellular demise./This killed the cells./The viability was greatly reduced./The cells were eradicated.
- **Abbreviations**

State-of-the-art NLP models encode words based on their context

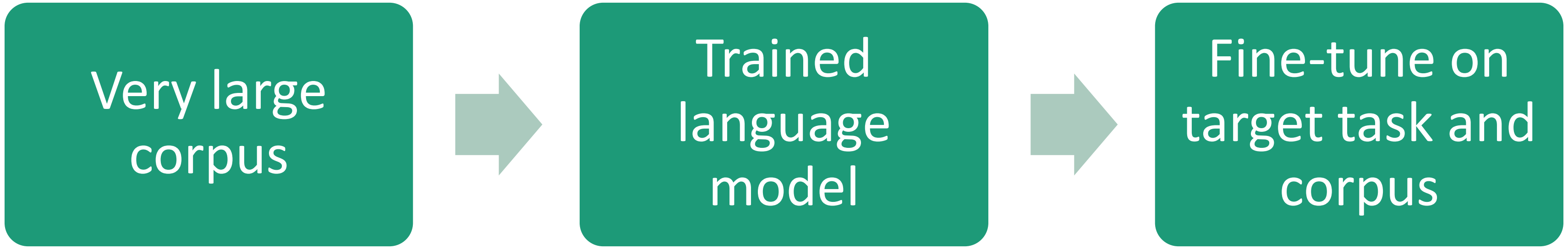
Occupation groups who spend many hours doing computer works often suffer from wrist pain related to **mouse** use.

mouse → vector a

This study used a transgenic **mouse** model to evaluate the impact of TP53 mutations on lung cancer survival.

mouse → vector b

Transfer learning with large language models (LLMs)



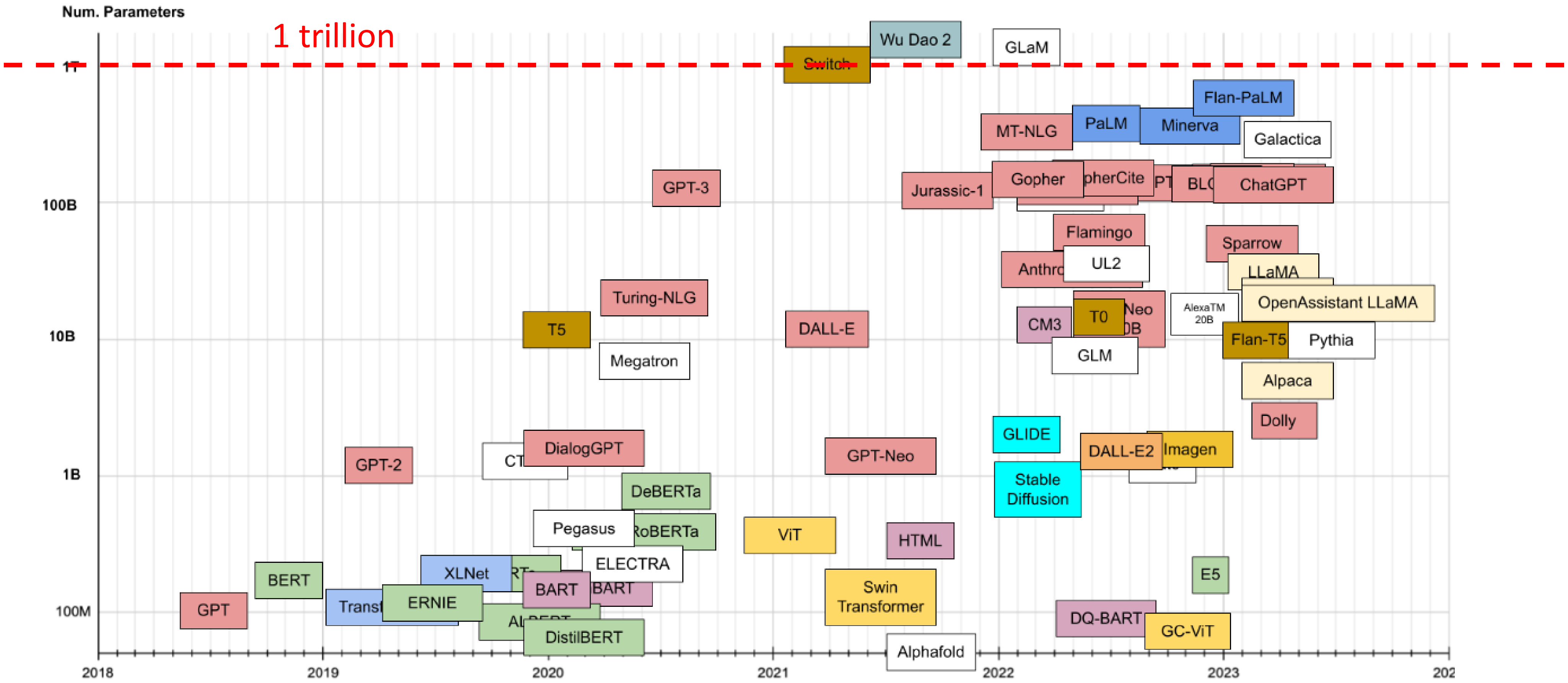
Very large corpus

Trained language model

Fine-tune on target task and corpus

BERT
GPT

LLMs are becoming really large



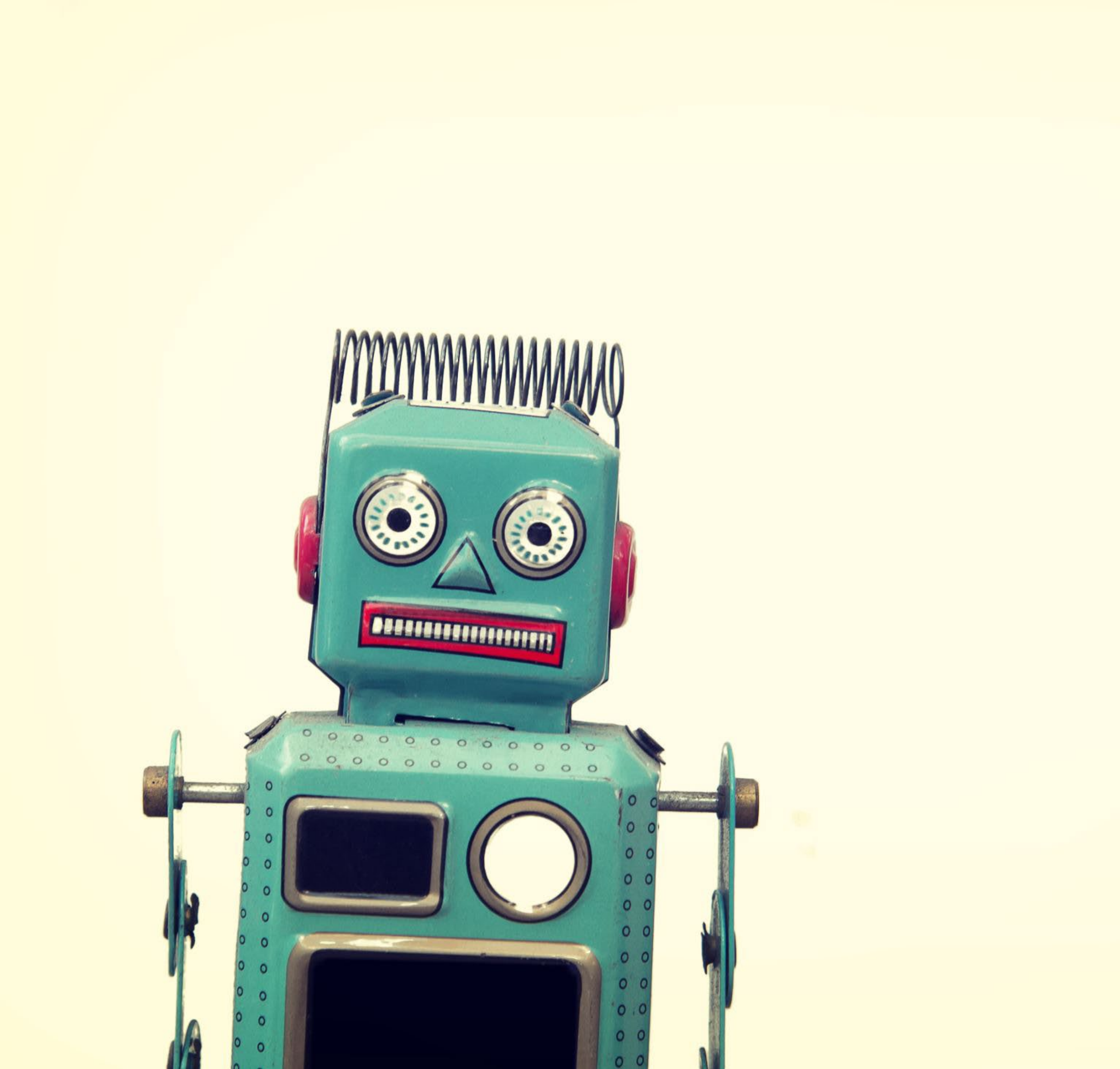
Amatriain X et al. arxiv 2023

Summary: common tasks for NLP

- Summarization
- Translation
- Clustering
- Text classification
- Sentiment analysis
- Named entity recognition
- Named entity linking
- Speech recognition
- Question answering
- Chat bots

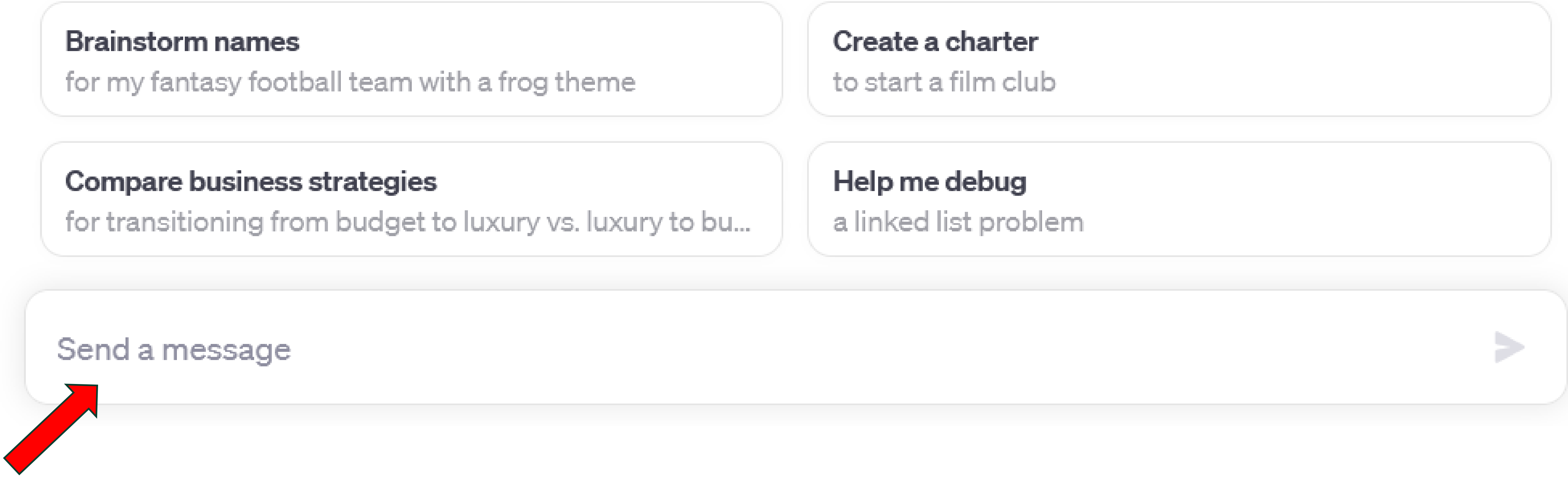
AI chatbots

ChatGPT
Bard
HuggingChat
Bing Chat



ChatGPT

https://chat.openai.com/



Brainstorm names
for my fantasy football team with a frog theme

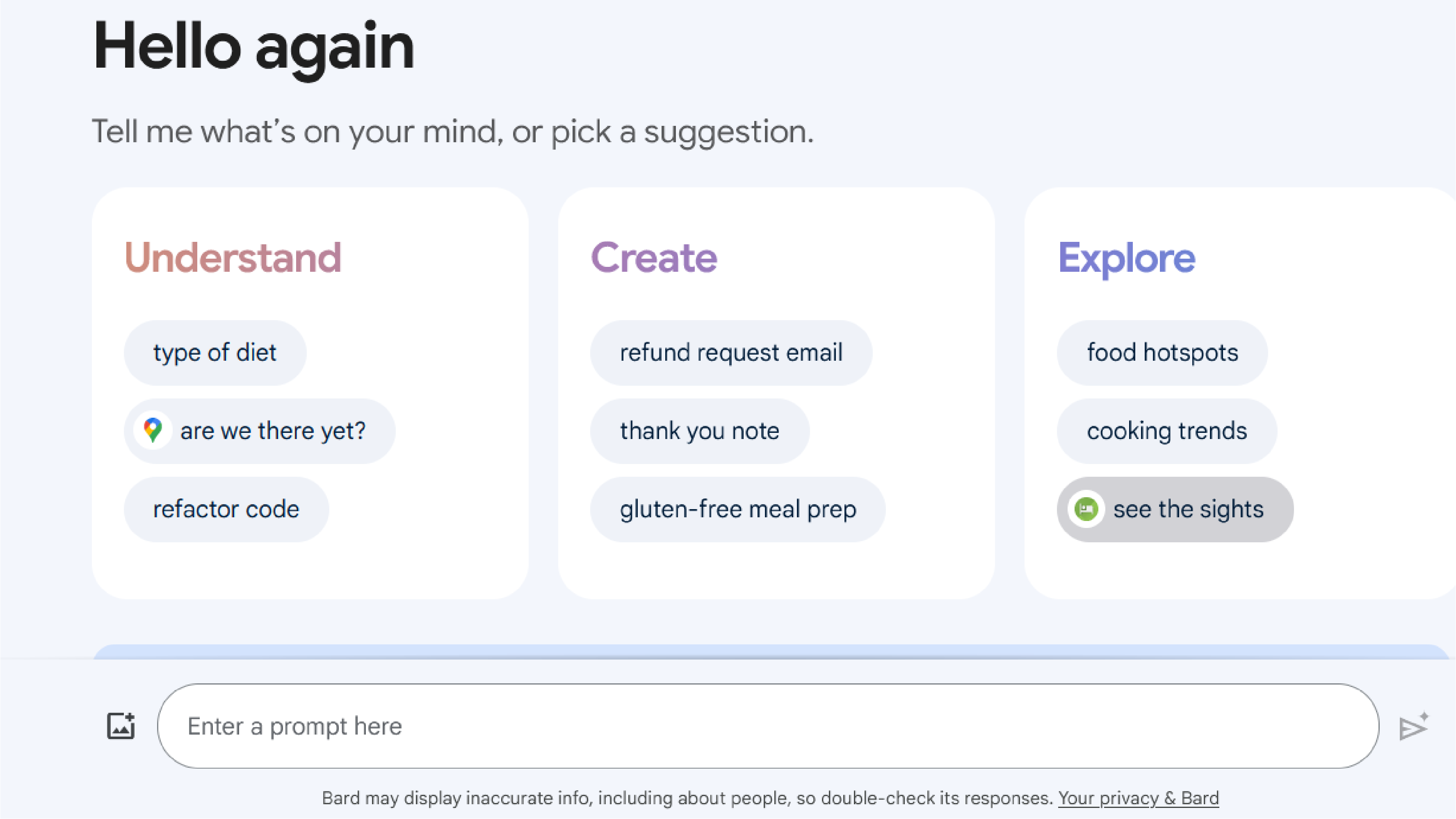
Create a charter
to start a film club

Compare business strategies
for transitioning from budget to luxury vs. luxury to bu...

Help me debug
a linked list problem

Bard

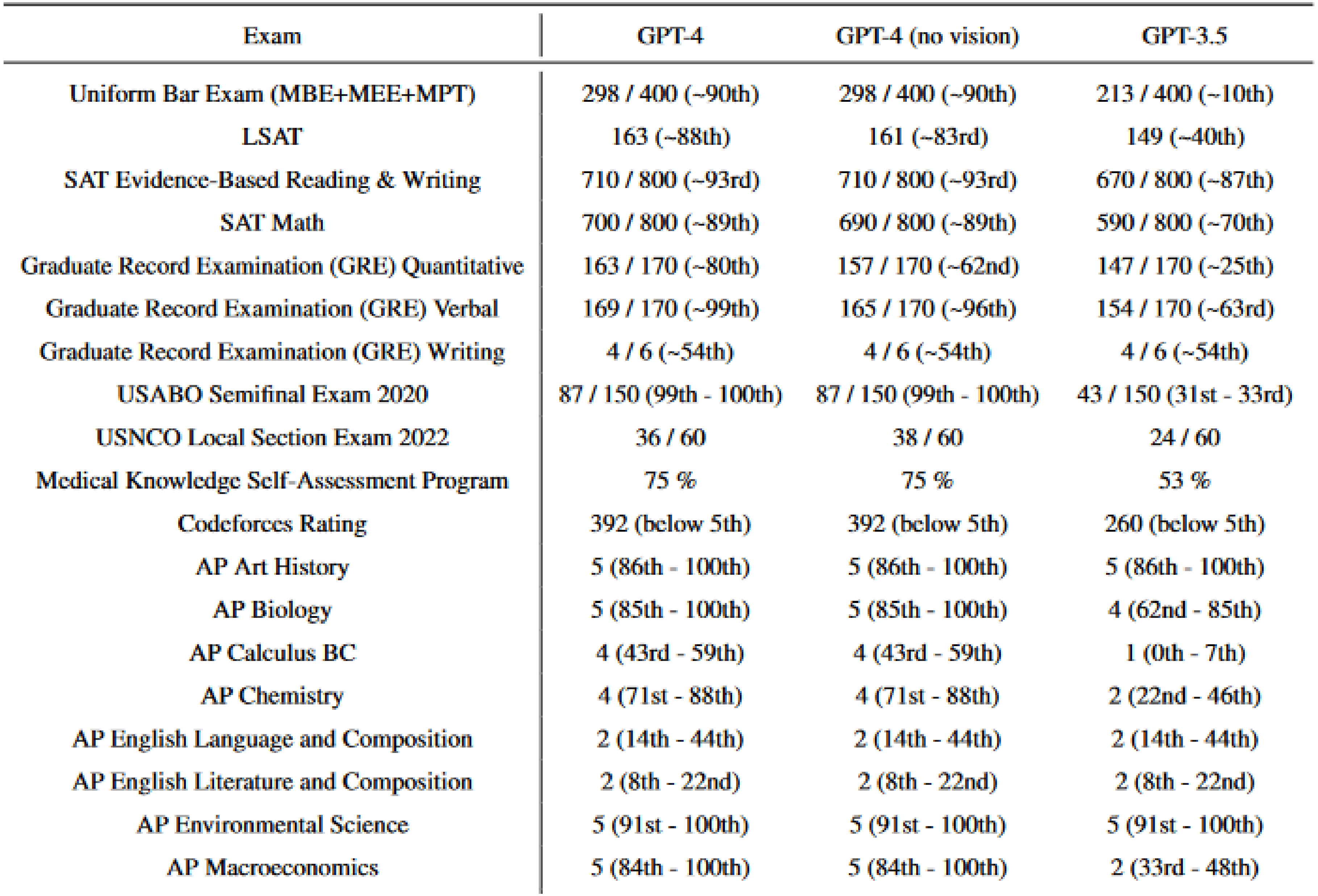
https://bard.google.com/



Hello again

Tell me what's on your mind, or pick a suggestion.

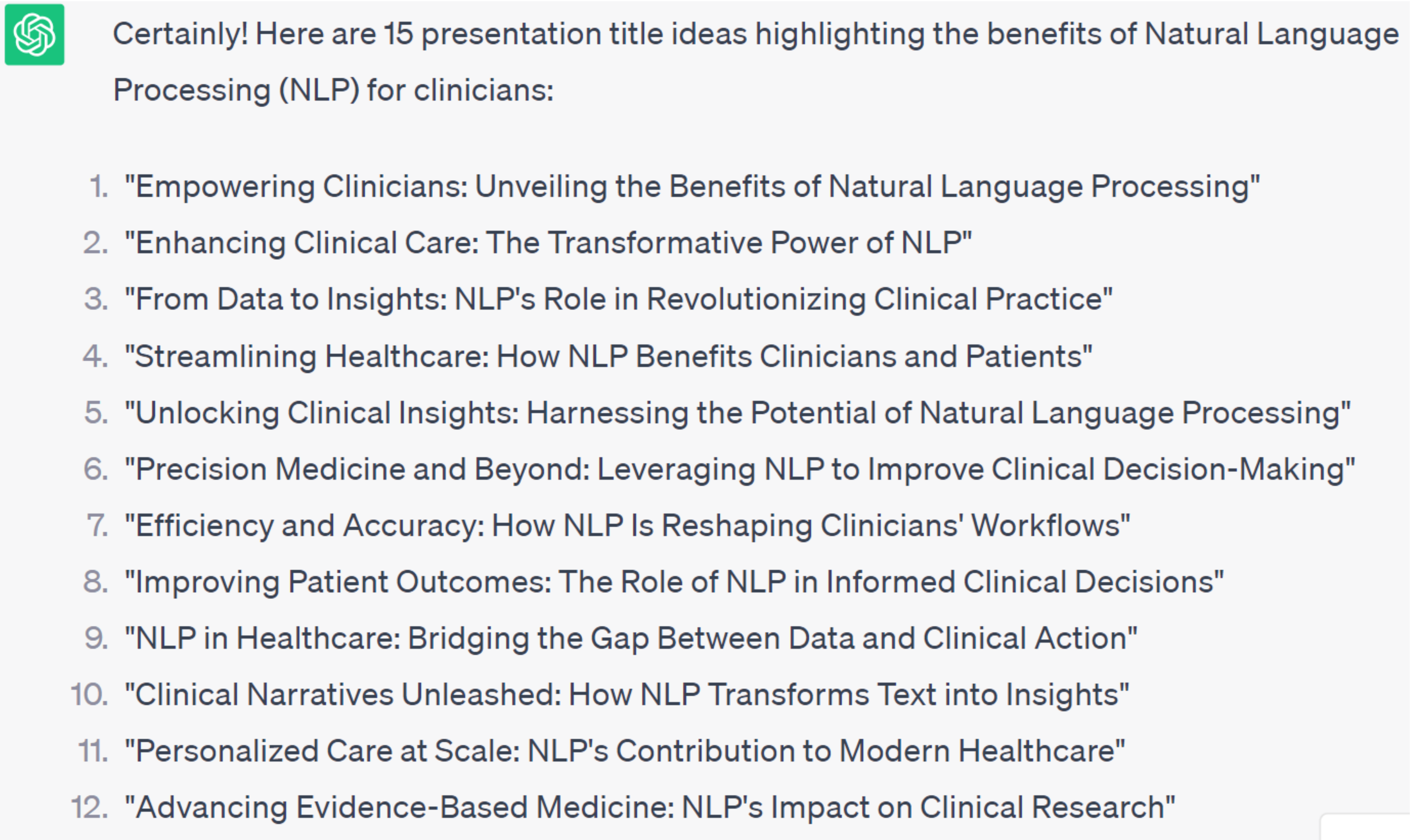
Bard may display inaccurate info, including about people, so double-check its responses. Your privacy & Bard



Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)

GPT-4 Technical report https://cdn.openai.com/papers/gpt-4.pdf

Writing assistant: Title suggestions



Certainly! Here are 15 presentation title ideas highlighting the benefits of Natural Language Processing (NLP) for clinicians:

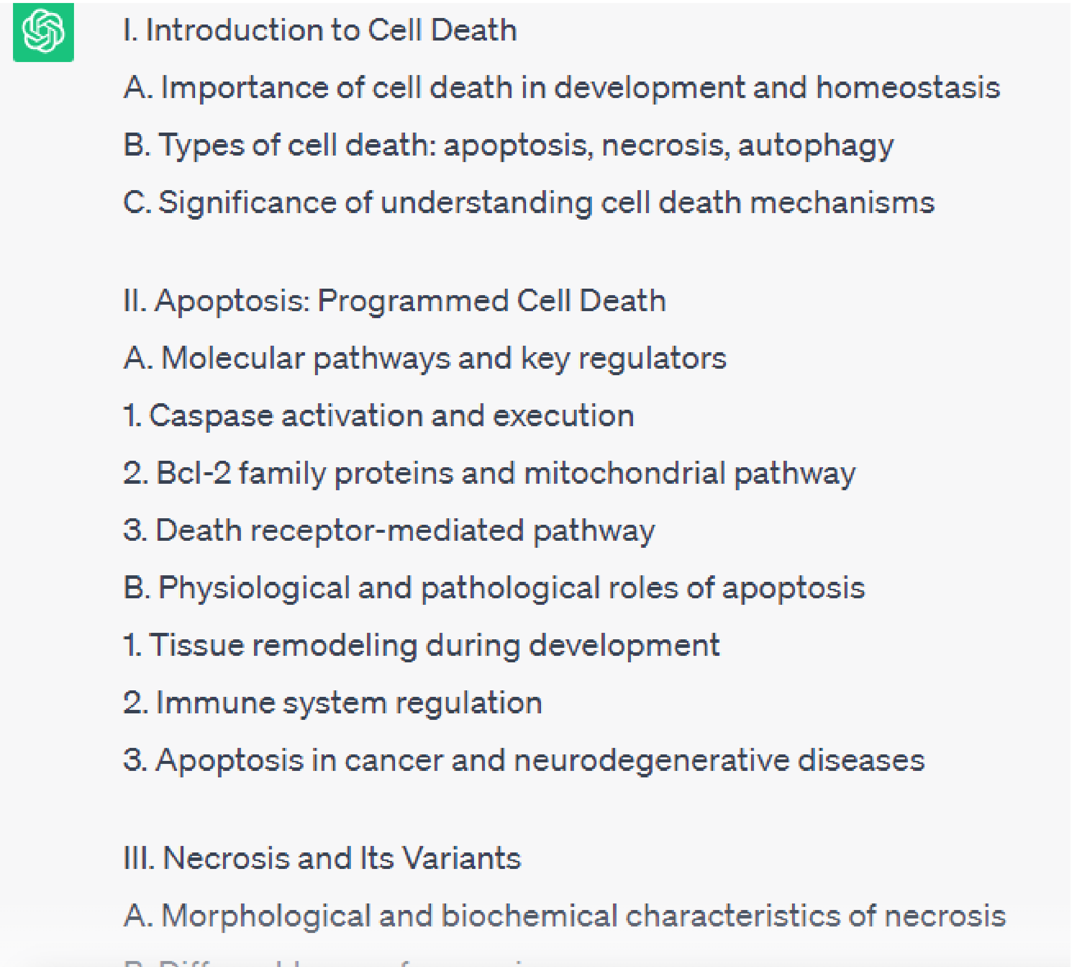
1. "Empowering Clinicians: Unveiling the Benefits of Natural Language Processing"
2. "Enhancing Clinical Care: The Transformative Power of NLP"
3. "From Data to Insights: NLP's Role in Revolutionizing Clinical Practice"
4. "Streamlining Healthcare: How NLP Benefits Clinicians and Patients"
5. "Unlocking Clinical Insights: Harnessing the Potential of Natural Language Processing"
6. "Precision Medicine and Beyond: Leveraging NLP to Improve Clinical Decision-Making"
7. "Efficiency and Accuracy: How NLP Is Reshaping Clinicians' Workflows"
8. "Improving Patient Outcomes: The Role of NLP in Informed Clinical Decisions"
9. "NLP in Healthcare: Bridging the Gap Between Data and Clinical Action"
10. "Clinical Narratives Unleashed: How NLP Transforms Text into Insights"
11. "Personalized Care at Scale: NLP's Contribution to Modern Healthcare"
12. "Advancing Evidence-Based Medicine: NLP's Impact on Clinical Research"

ChatGPT

Writing assistant: Outline generation

Prompt: “write an outline for a scientific review on cell death”

ChatGPT



I. Introduction to Cell Death

A. Importance of cell death in development and homeostasis

B. Types of cell death: apoptosis, necrosis, autophagy

C. Significance of understanding cell death mechanisms

II. Apoptosis: Programmed Cell Death

A. Molecular pathways and key regulators

1. Caspase activation and execution

2. Bcl-2 family proteins and mitochondrial pathway

3. Death receptor-mediated pathway

B. Physiological and pathological roles of apoptosis

1. Tissue remodeling during development

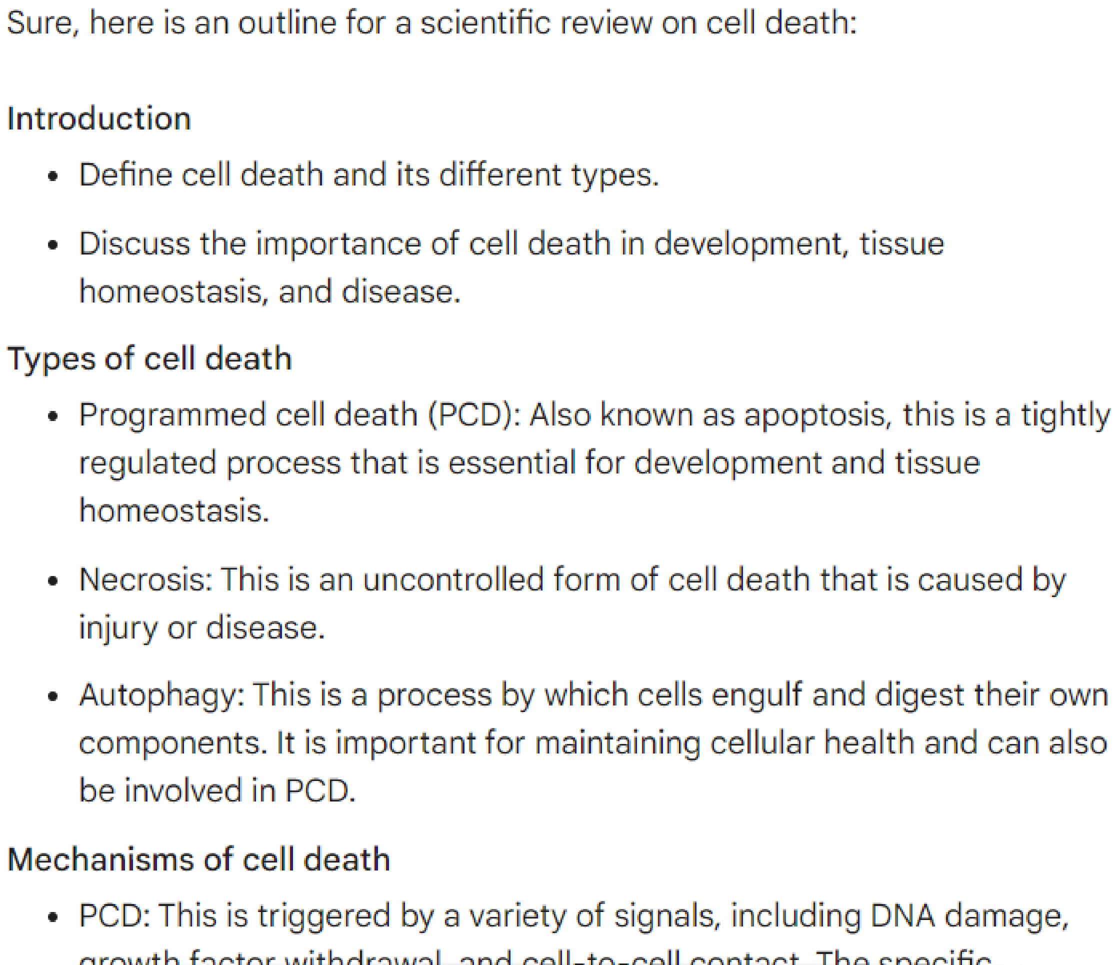
2. Immune system regulation

3. Apoptosis in cancer and neurodegenerative diseases

III. Necrosis and Its Variants

A. Morphological and biochemical characteristics of necrosis

Bard



Sure, here is an outline for a scientific review on cell death:

Introduction

- Define cell death and its different types.
- Discuss the importance of cell death in development, tissue homeostasis, and disease.

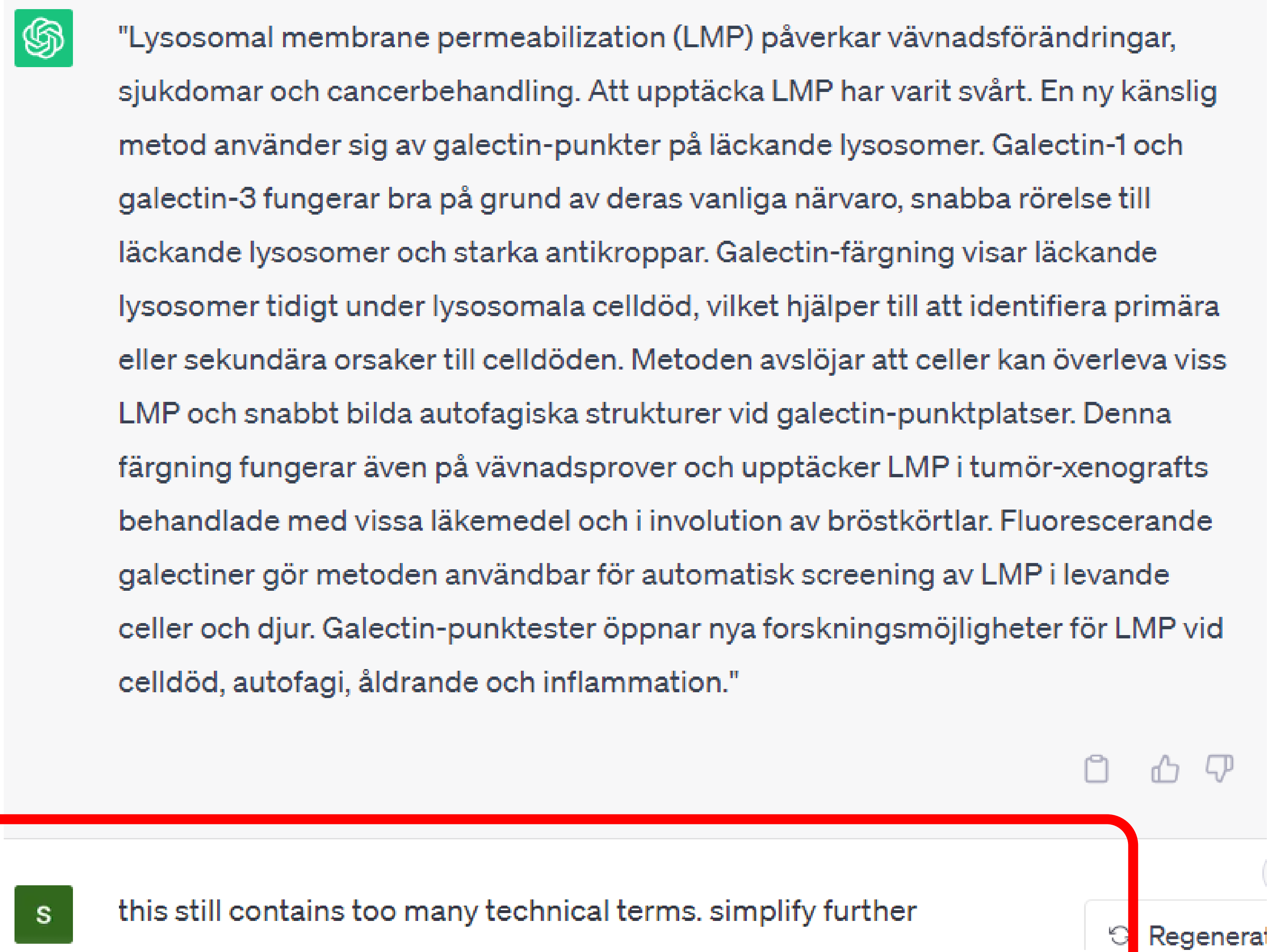
Types of cell death

- Programmed cell death (PCD): Also known as apoptosis, this is a tightly regulated process that is essential for development and tissue homeostasis.
- Necrosis: This is an uncontrolled form of cell death that is caused by injury or disease.
- Autophagy: This is a process by which cells engulf and digest their own components. It is important for maintaining cellular health and can also be involved in PCD.

Mechanisms of cell death

- PCD: This is triggered by a variety of signals, including DNA damage, growth factor withdrawal, and cell-to-cell contact. The specific

Writing assistant:
Summarization + style change + translation



"Lysosomal membrane permeabilization (LMP) påverkar vävnadsförändringar, sjukdomar och cancerbehandling. Att upptäcka LMP har varit svårt. En ny känslig metod använder sig av galectin-punkter på läckande lysosomer. Galectin-1 och galectin-3 fungerar bra på grund av deras vanliga närvaro, snabba rörelse till läckande lysosomer och starka antikroppar. Galectin-färgning visar läckande lysosomer tidigt under lysosomala celldöd, vilket hjälper till att identifiera primära eller sekundära orsaker till celldöden. Metoden avslöjar att celler kan överleva viss LMP och snabbt bilda autofagiska strukturer vid galectin-punktplatser. Denna färgning fungerar även på vävnadsprover och upptäcker LMP i tumör-xenografts behandlade med vissa läkemedel och i involution av bröstkörtlar. Fluorescerande galectiner gör metoden användbar för automatisk screening av LMP i levande celler och djur. Galectin-punktester öppnar nya forskningsmöjligheter för LMP vid celldöd, autofagi, åldrande och inflammation."

S

this still contains too many technical terms. simplify further

ChatGPT

Writing assistant:
Production of complete text

Prompt:
"write a script for a 2 min video that explains the process of gene expression in a clear and concise way to biomedical university students "

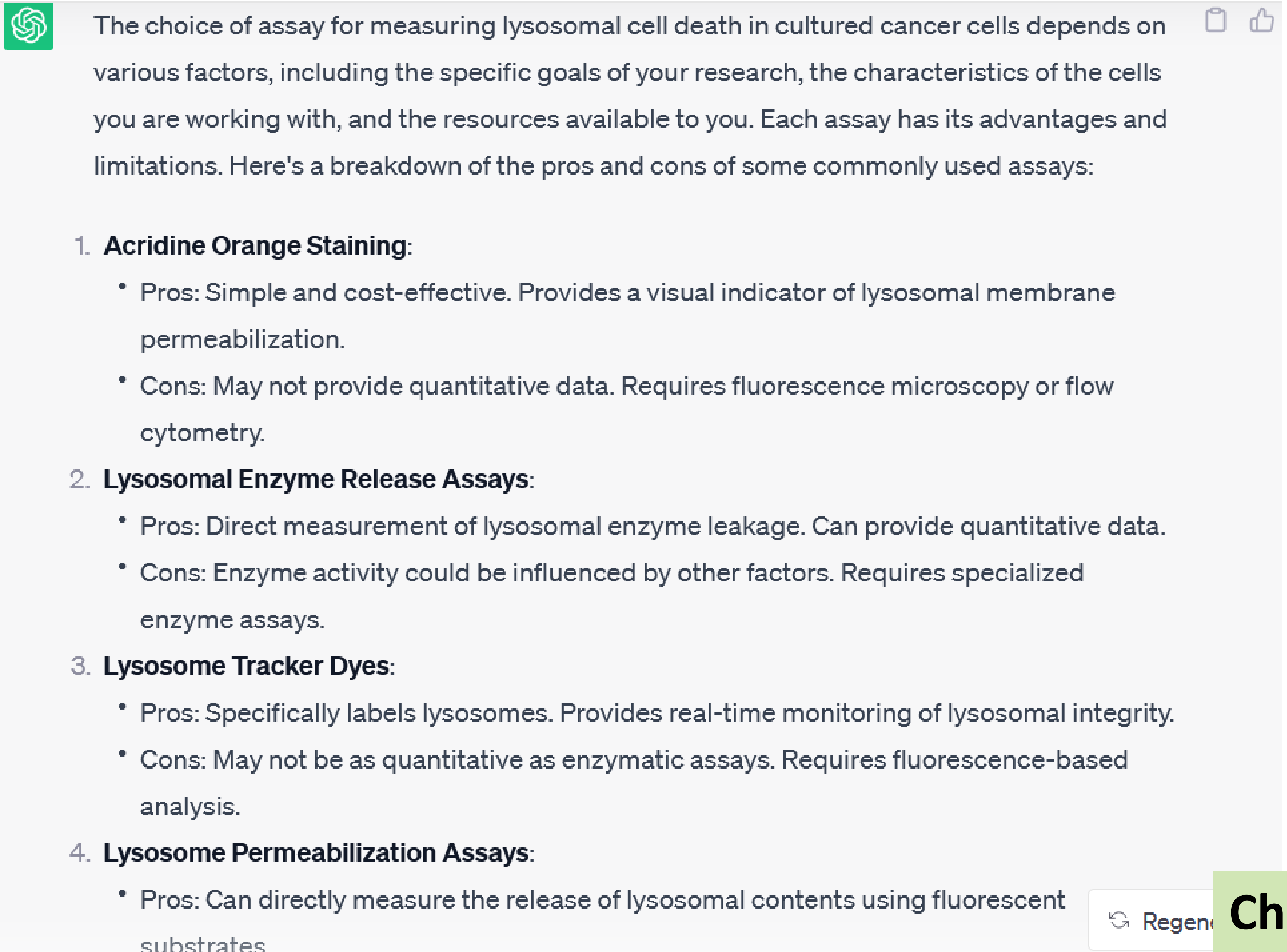
Hi everyone, and welcome to this video about gene expression. Gene expression is the process by which DNA is converted into proteins. Proteins are the building blocks of cells and tissues, and they play a role in almost every function of the body.

The first step in gene expression is transcription…

Bard

"how can lysosomal cell death be measured in cultured cancer cells?"
"which of these assays is best?"

Scientific discussion partner

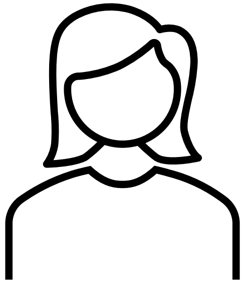


The choice of assay for measuring lysosomal cell death in cultured cancer cells depends on various factors, including the specific goals of your research, the characteristics of the cells you are working with, and the resources available to you. Each assay has its advantages and limitations. Here's a breakdown of the pros and cons of some commonly used assays:

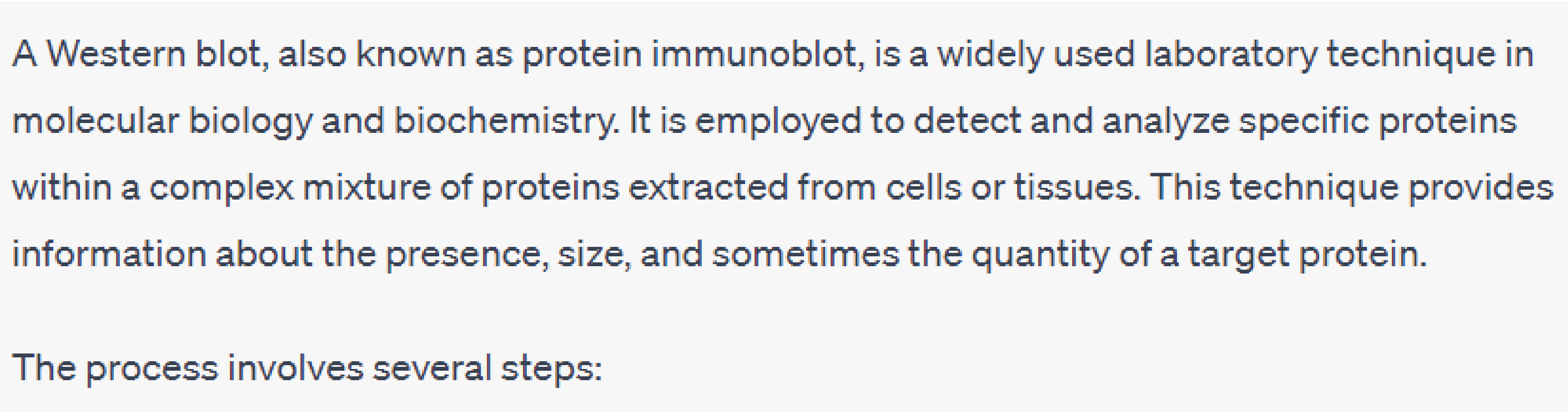
1. **Acridine Orange Staining**:
   - Pros: Simple and cost-effective. Provides a visual indicator of lysosomal membrane permeabilization.
   - Cons: May not provide quantitative data. Requires fluorescence microscopy or flow cytometry.
2. **Lysosomal Enzyme Release Assays**:
   - Pros: Direct measurement of lysosomal enzyme leakage. Can provide quantitative data.
   - Cons: Enzyme activity could be influenced by other factors. Requires specialized enzyme assays.
3. **Lysosome Tracker Dyes**:
   - Pros: Specifically labels lysosomes. Provides real-time monitoring of lysosomal integrity.
   - Cons: May not be as quantitative as enzymatic assays. Requires fluorescence-based analysis.
4. **Lysosome Permeabilization Assays**:
   - Pros: Can directly measure the release of lysosomal contents using fluorescent substrates

ChatGPT

Personal teacher



“explain what a Western Blot is”

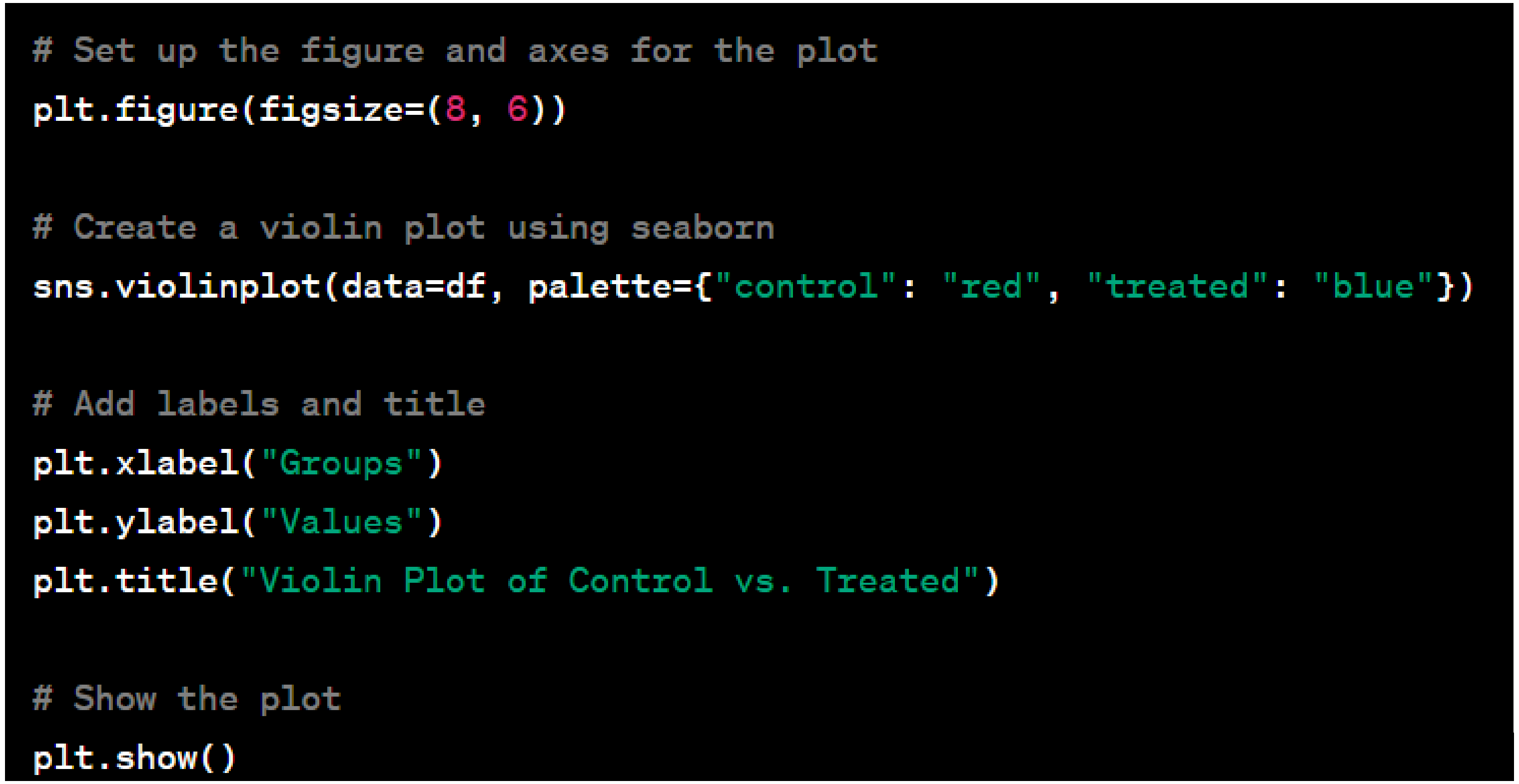


A Western blot, also known as protein immunoblot, is a widely used laboratory technique in molecular biology and biochemistry. It is employed to detect and analyze specific proteins within a complex mixture of proteins extracted from cells or tissues. This technique provides information about the presence, size, and sometimes the quantity of a target protein.

The process involves several steps:

Programming assistant

"generate a script that does the following: 1. loads the values from a pandas dataframe called df which has two columns called "control" and "treated". 2. plots the two columns as two data series in a violin plot with the control column values in red and the treated column values in blue"

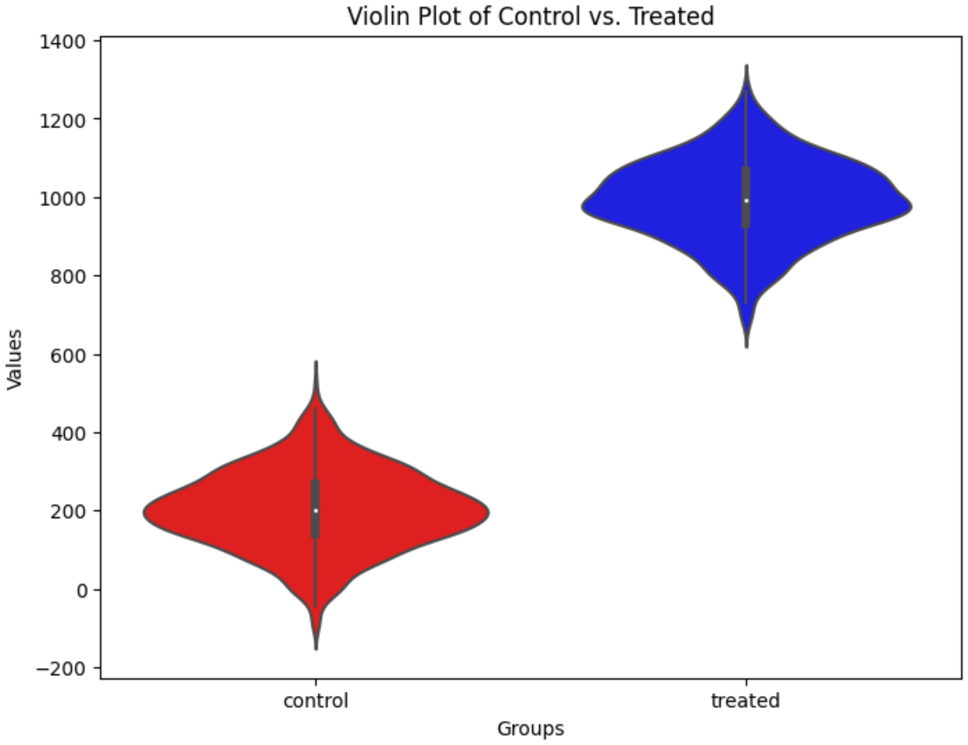


```
# Set up the figure and axes for the plot
plt.figure(figsize=(8, 6))

# Create a violin plot using seaborn
sns.violinplot(data=df, palette={"control": "red", "treated": "blue"})

# Add labels and title
plt.xlabel("Groups")
plt.ylabel("Values")
plt.title("Violin Plot of Control vs. Treated")

# Show the plot
plt.show()
```



ChatGPT

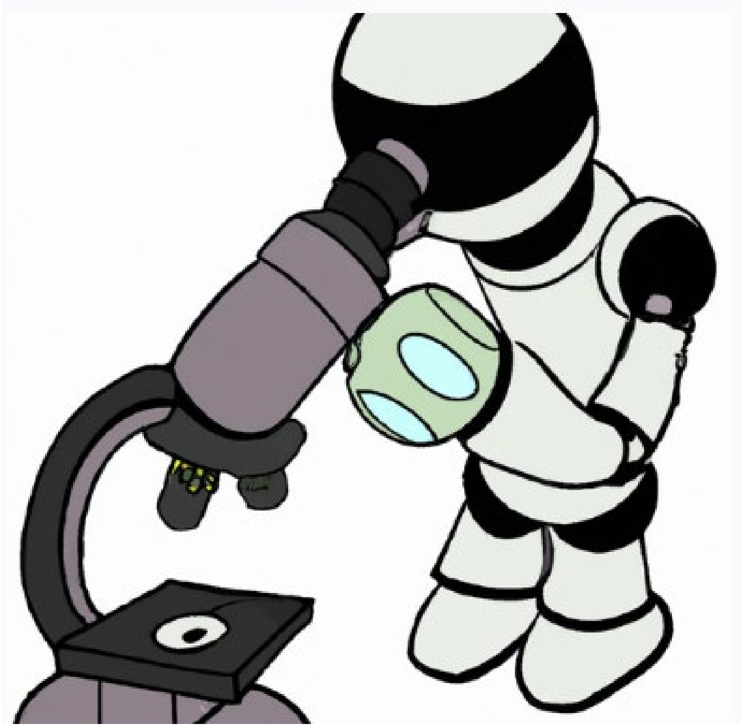
AI chatbots can support you in many ways

- Writing assistant
  - Title suggestions
  - Content suggestions
  - Outline generation
  - Rephrasing/Correction/Shortening
  - Summarization/Style change/Translation
  - Production of complete text
- Discussion partner
- Personal teacher
- Programming assistant

AI tools for image generation

“a cartoon image of a white robot looking into a microscope”

DALL-E2



A few words of caution…

- Incorrect/biased results
- Prompts are stored – do not provide sensitive information!
- Check journal policies for AI
- Ethical/legal/societal issues