A Journey from Transformers to Large Language Models: an Educational Perspective

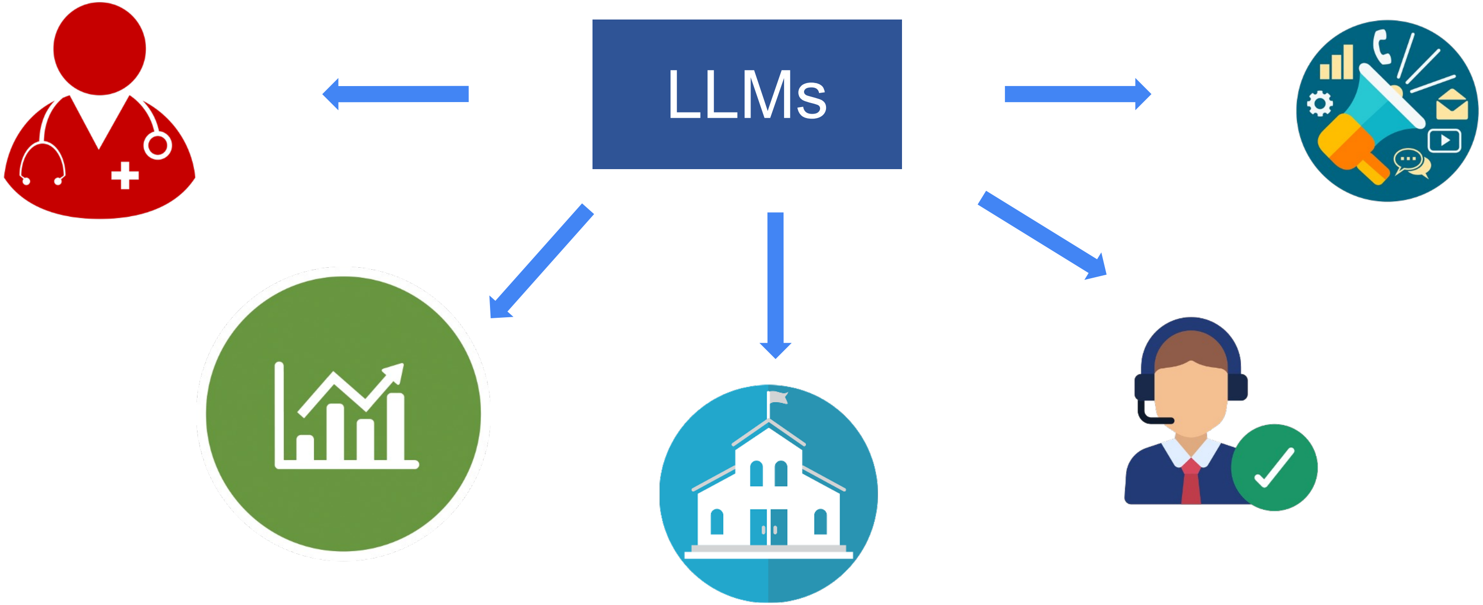
Irene Li
University of Tokyo

AIGC, 2023

Li Lab

東京大学
THE UNIVERSITY OF TOKYO

The New Foundation Model: Large Language Models

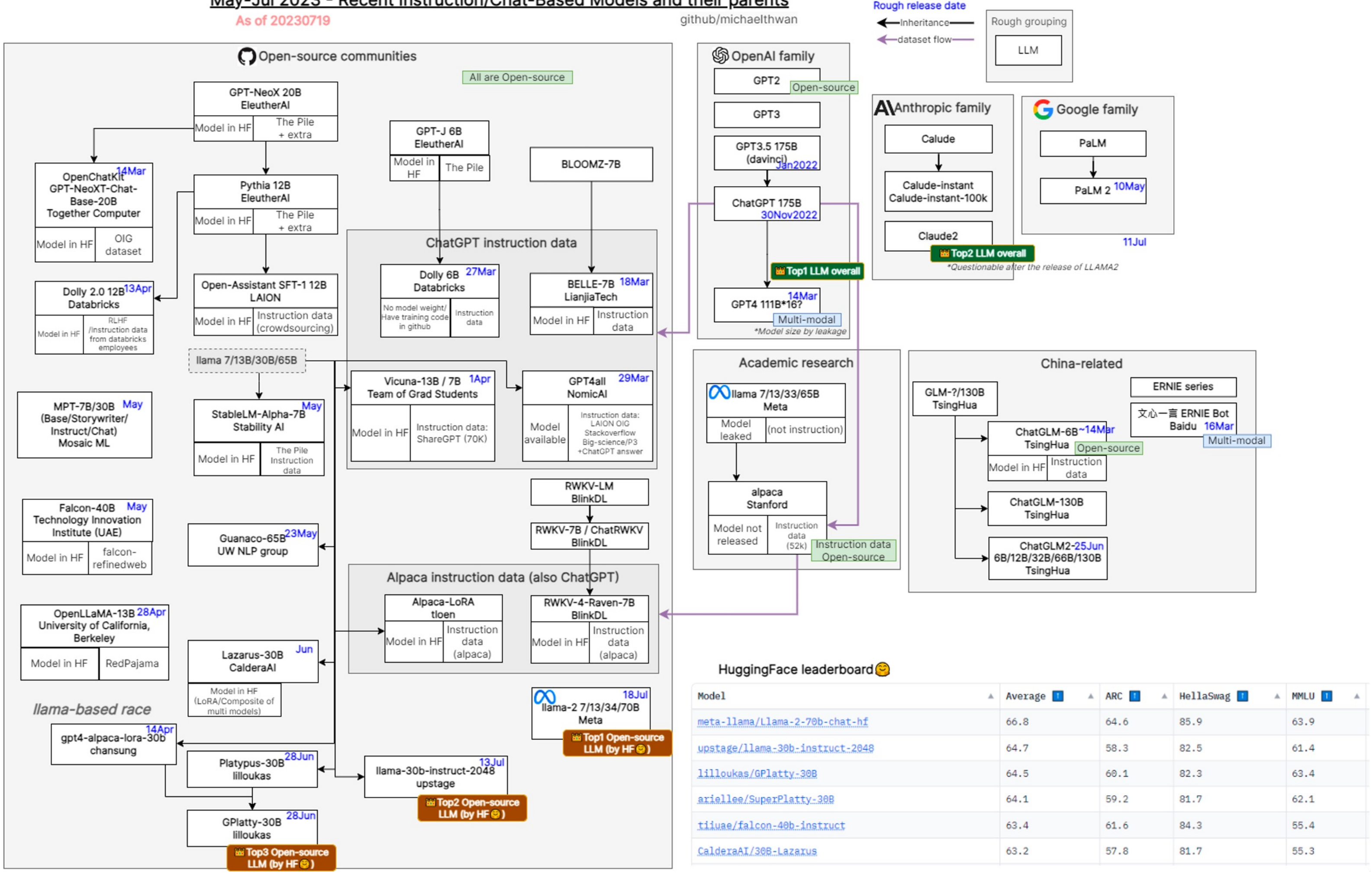


LLMs

May-Jul 2023 - Recent Instruction/Chat-Based Models and their parents

As of 20230719

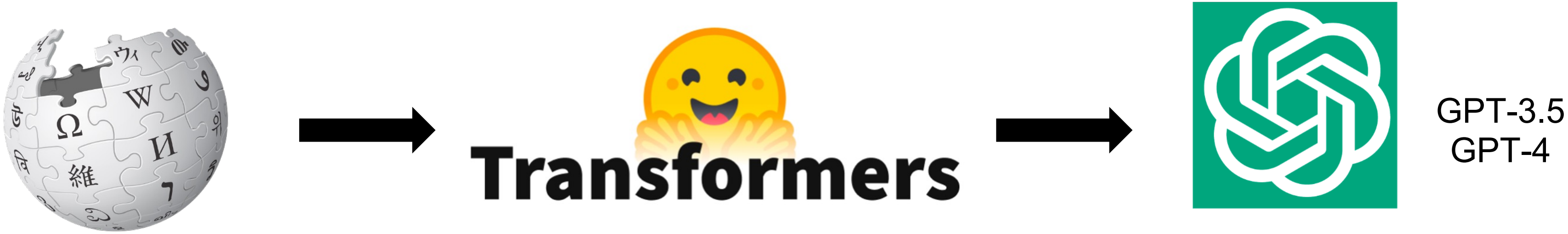
github/michaelthwan



Model	Average	ARC	HellaSwag	MMLU
meta-llama/Llama-2-70b-chat-hf	66.8	64.6	85.9	63.9
upstage/llama-30b-instruct-2048	64.7	58.3	82.5	61.4
lilloukas/GPlatty-30B	64.5	60.1	82.3	63.4
ariellee/SuperPlatty-30B	64.1	59.2	81.7	62.1
tiiuae/falcon-40b-instruct	63.4	61.6	84.3	55.4
CalderaAI/30B-Lazarus	63.2	57.8	81.7	55.3

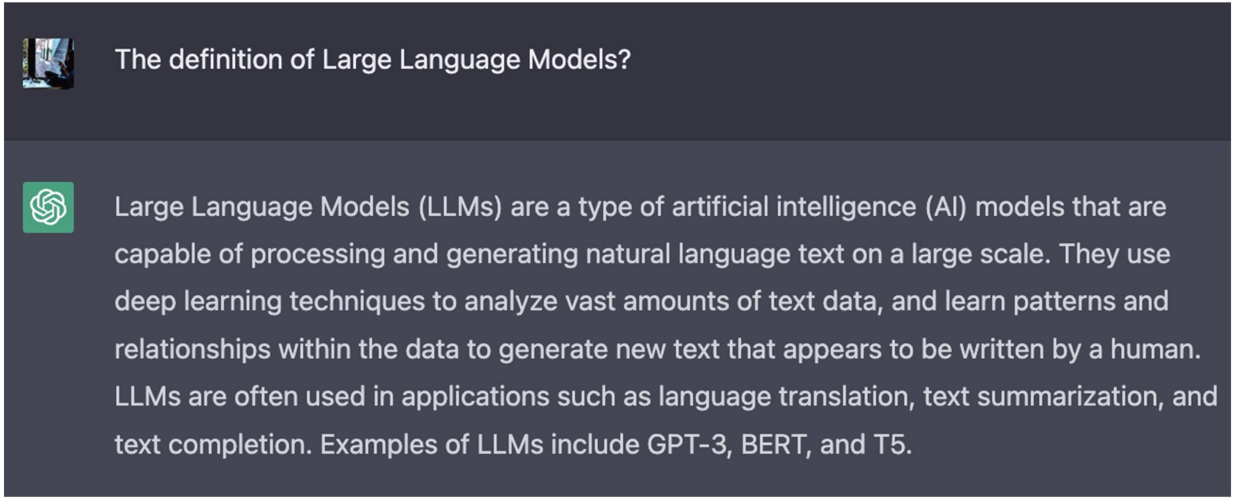
https://github.com/michaelthwan/llm_family_chart/blob/master/2023_LLMfamily.drawio.png

Large Language Models



Transformers

GPT-3.5
GPT-4

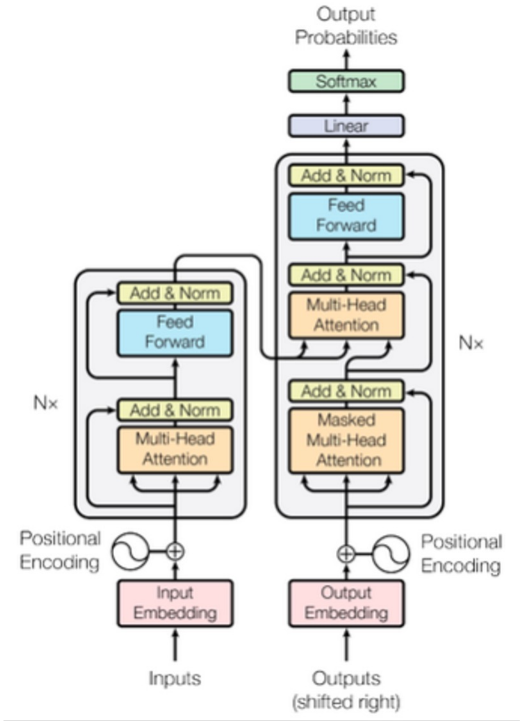


The definition of Large Language Models?

Large Language Models (LLMs) are a type of artificial intelligence (AI) models that are capable of processing and generating natural language text on a large scale. They use deep learning techniques to analyze vast amounts of text data, and learn patterns and relationships within the data to generate new text that appears to be written by a human. LLMs are often used in applications such as language translation, text summarization, and text completion. Examples of LLMs include GPT-3, BERT, and T5.

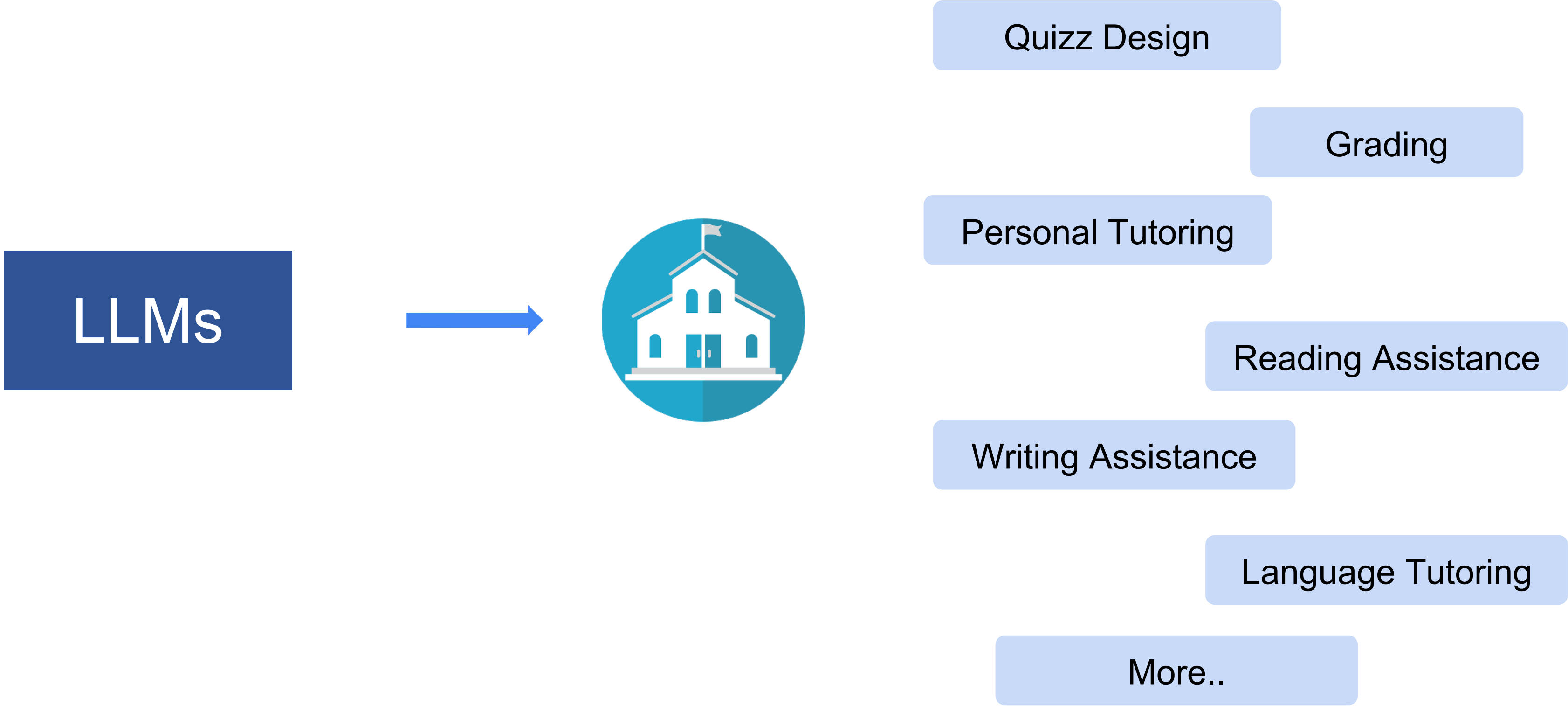
🤖Transformers

- A seq2seq model based on attention self-attention mechanism;
- Foundational architecture for later strong models;
- Variations in many other scenarios: BERT, RoBERTa, XLM, …



Vaswani, Ashish, et al. "Attention is All You Need." 2017

Education?



LLMs

Quizz Design

Grading

Personal Tutoring

Reading Assistance

Writing Assistance

Language Tutoring

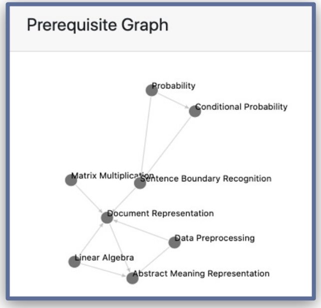
More..

Days Before LLMs

We typically follow a pipeline…

Educational Resources:
Textbooks, papers,
videos, blog posts,...
Unstructured Texts

Generic
Tasks

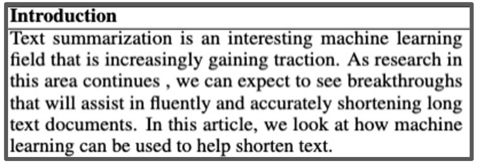


Prerequisites: updated with
new concepts
Dynamic Knowledge Graph:
structured

Support
Downstream
Applications

Learning Path

Resource
Recommendation



Introduction
Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues , we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. In this article, we look at how machine learning can be used to help shorten text.

Machine-generated Survey

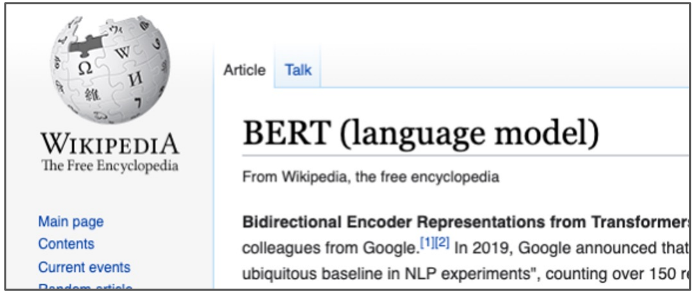
Irene Li, Fabbri A, Kawamura R, Liu Y, Tang X, Tae J, Shen C, Ma S, Mizutani T, Radev D. Surfer100: Generating Surveys From Web Resources, Wikipedia-style. LREC 2022

Irene Li, Alexander Fabbri, Robert Tung and Dragomir Radev. What Should I Learn First: Introducing LectureBank for NLP Education and Prerequisite Chain Learning. AAAI, 2019

Survey Generation

Motivation: summarizing educational web resources

What’s the best way to get an idea on a **novel** concept? Survey paper? Blog post? Wikipedia article?



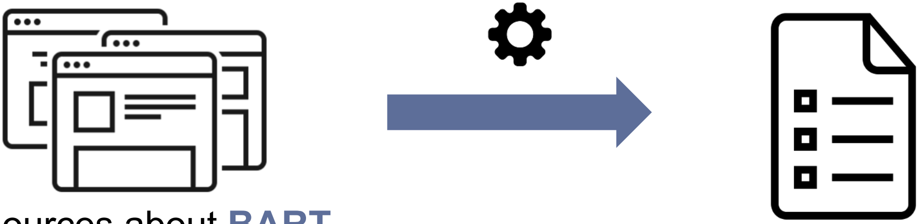
BERT (language model)

From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformer
colleagues from Google.[1][2] In 2019, Google announced that
ubiquitous baseline in NLP experiments", counting over 150 r

“Where can I find BART?”
“Where can I find **recent new concepts** and topics? ”

A way to do **automatic** summarization for novel concepts….



Resources about BART

Wikipedia-like survey about BART

Introduction: survey generation

Comprehensive survey generation:

A simple paragraph for introduction ×

A Wikipedia-style survey includes multiple sections. ✅

Problem Definition:

Abstractive summarization from a list of related input documents;

Generate short summary for **EACH** individual section.

Surfer100:

Manually selected 100 scientific topics, mainly NLP topics.

Based on web data, 8 annotators, each survey requires 45-60 minutes.

Each section contains 50-150 words.

Generating Wikipedia-style surveys from web resources:

- *Introduction*
- *History*
- *Main Idea*
- *Variations*
- *Applications*

Survey Structure

Key Idea:
The A* algorithm has 3 parameters. The first, g, is the cost of moving from the initial cell to the current cell, i.e. it is the sum of all the cells that have been visited since leaving the first cell. h, also known as the heuristic value, is the estimated cost of moving from the current cell to the final cell. The actual cost cannot be calculated until the final cell is reached, therefore, it is an estimated value. …

Example of A* Search.

Methodology: a two-stage method

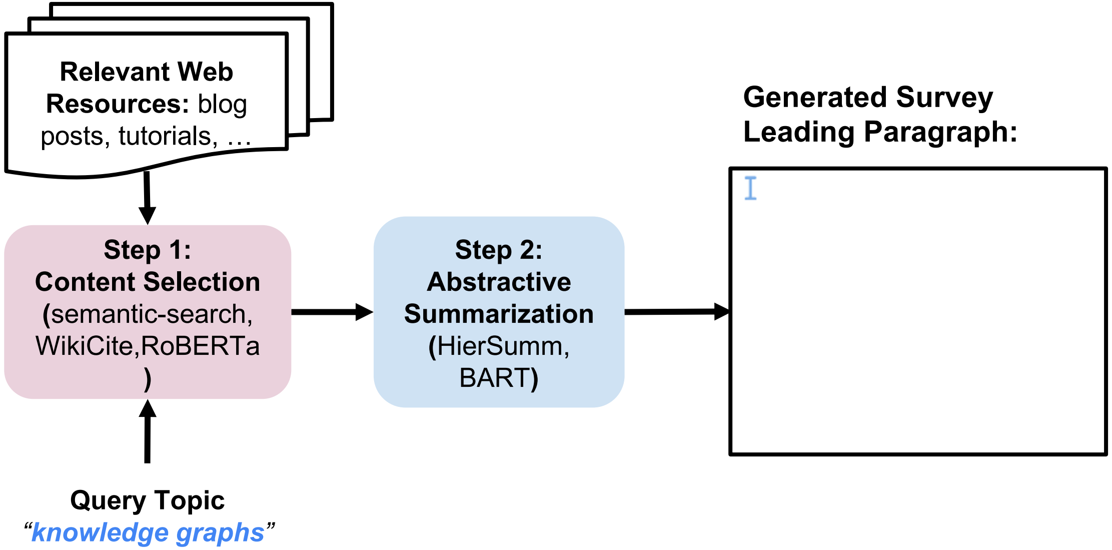
Step 1: Content selection

Not every single sentence is considered to be relevant.

Long input issue.

Step 2: Abstractive Summarization

Pre-trained models for generating abstractive summarization for each section.



Relevant Web Resources: blog posts, tutorials, …

Step 1: Content Selection (semantic-search, WikiCite,RoBERTa)

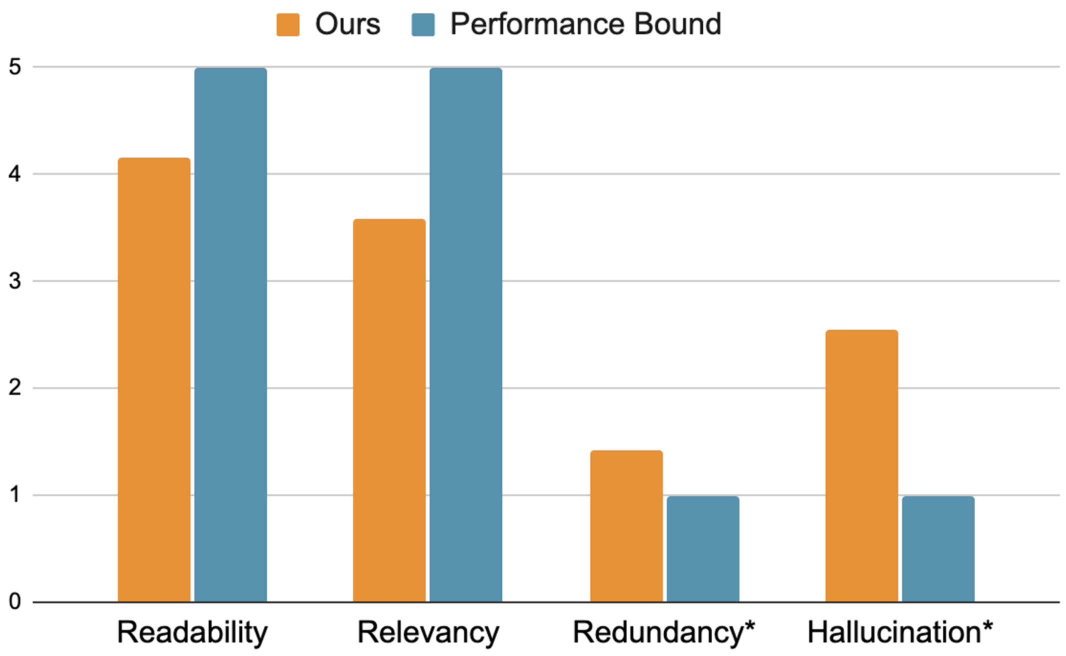
Step 2: Abstractive Summarization (HierSumm, BART)

Generated Survey Leading Paragraph:

Query Topic
“***knowledge graphs***”

A Two-stage Method for Survey Generation

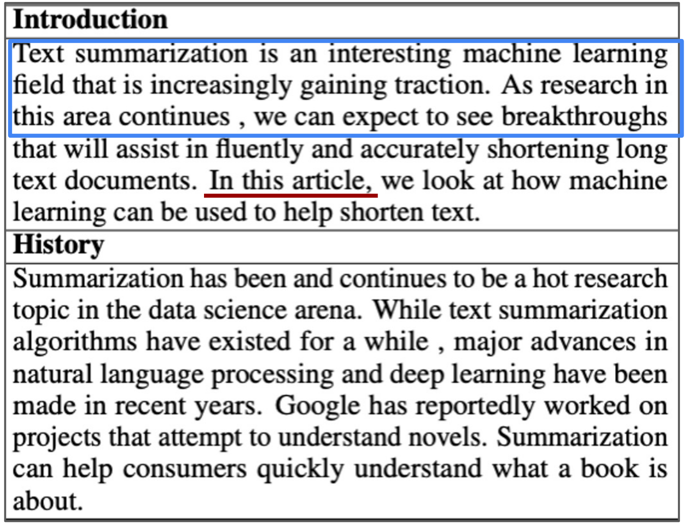
Human Evaluation on 20 concepts



Randomly select 20 concepts and ask two human judges to give scores (range 1-5) on the following four perspectives: *readability, relevancy, redundancy* and *hallucination*.

Case Study

Generated Survey on the topic
"*text summarization*"

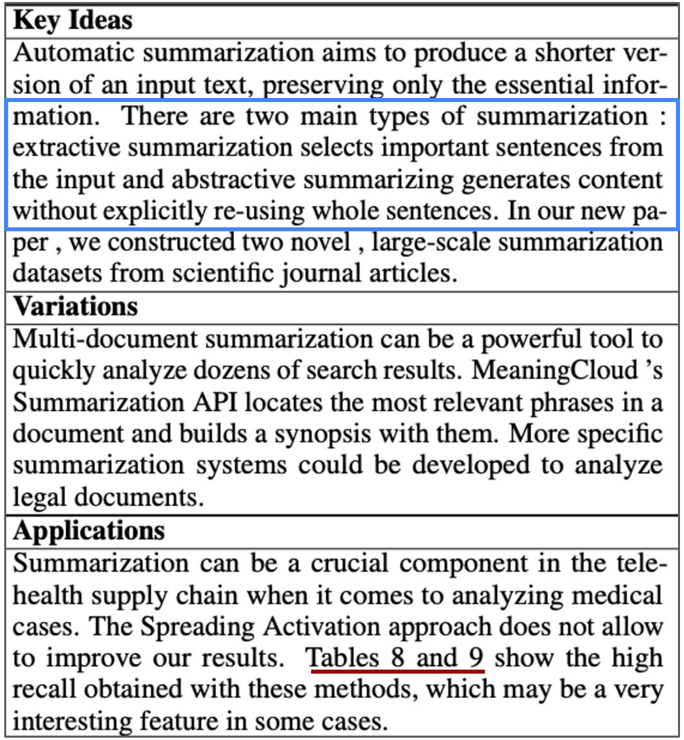


Introduction

Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues , we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. In this article, we look at how machine learning can be used to help shorten text.

History

Summarization has been and continues to be a hot research topic in the data science arena. While text summarization algorithms have existed for a while , major advances in natural language processing and deep learning have been made in recent years. Google has reportedly worked on projects that attempt to understand novels. Summarization can help consumers quickly understand what a book is about.



Key Ideas

Automatic summarization aims to produce a shorter version of an input text, preserving only the essential information. There are two main types of summarization : extractive summarization selects important sentences from the input and abstractive summarizing generates content without explicitly re-using whole sentences. In our new paper , we constructed two novel , large-scale summarization datasets from scientific journal articles.

Variations

Multi-document summarization can be a powerful tool to quickly analyze dozens of search results. MeaningCloud 's Summarization API locates the most relevant phrases in a document and builds a synopsis with them. More specific summarization systems could be developed to analyze legal documents.

Applications

Summarization can be a crucial component in the telehealth supply chain when it comes to analyzing medical cases. The Spreading Activation approach does not allow to improve our results. Tables 8 and 9 show the high recall obtained with these methods, which may be a very interesting feature in some cases.

Now with LLMs…

[Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts]

Fan Gao, Hang Jiang, Moritz Blum, Jinghui Lu, Yuang Jiang and Irene Li;

Preprint

GPT-3.5 and GPT 4: prompts

We ask GPTs to generate 20 structured survey articles, and provide 3 types: **zero-shot**, **one-shot**, and **with description prompt**.

zero-shot

Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words.

20 Topics:

- BERT
- Autoencoders
- Clustering
- Decision Trees
- Ensemble Learning
- Gaussian Mixture Model
- Generative Adversarial Network
- Gradient Boosting
- Hidden Markov Models
- Knowledge Graphs
- Language Modeling
- Long Short-Term Memory Network
- Maximum Marginal Relevance
- Meta Learning
- Multilingual BERT
- Perceptron
- Relation Extraction
- Residual Neural Network
- RMSprop Optimizer
- Sentiment Analysis

one-shot

Example survey:

<INTRODUCTION>
Word2Vec is one of the most popular tools to learn word embeddings using shallow neural networks. It first constructs a vocabulary from the training text data and then learns word embeddings....

<HISTORY>
Word2vec was developed by a group of researchers headed by Tomas Mikolov at Google. Machine learning models take vectors as input, ...

<KEY IDEAS>
Word2Vec converts words into vector forms such that similar meaning words appear together and dissimilar words are located far away…

<USES/APPLICATIONS>
Gensim provides the Word2Vec class for working with a Word2Vec model. Training your own word vectors can take a long time and uses lots of memory…

<VARIATIONS>
Word embeddings is an active research area trying to figure out better word representations than the existing ones...

Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words.

with prompt

Generate a survey about <Topic>. There should be five sub-sections: Introduction, History, Key Ideas, Variations and Applications. Each sub-section should contain 50-150 words.

The following is the guideline for each section:

SECTION 1: INTRODUCTION
Describe what the topic is (a method, a model, a task, a dataset), which field/subfield it is part of, quick overview of applications and motivation behind concept and related ideas)

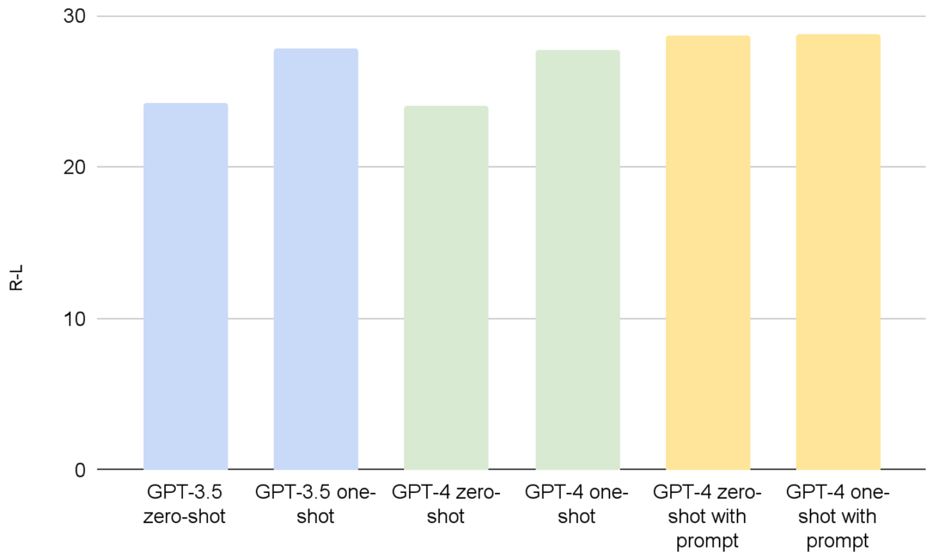
SECTION 2: HISTORY
Describe when or by who the topic was introduced, in what context, what problems it addresses.

SECTION 3: KEY IDEAS
Describe in greater depth (could provide some mathematical context or explain core concepts).

SECTION 4: USES/APPLICATIONS
Describe for what tasks this model/data is used.

SECTION 5: VARIATIONS
What variations or similar models, datasets, tasks exist and how does this topic fit into a bigger picture.

ROUGE scores



One-shot improves on zero-shot setting;

With description prompt, there is no large difference.

Human Evaluation: 4 human judges, 6 perspectives

📖 **Readability:** The text is easy to read, well-structured, and flows naturally.

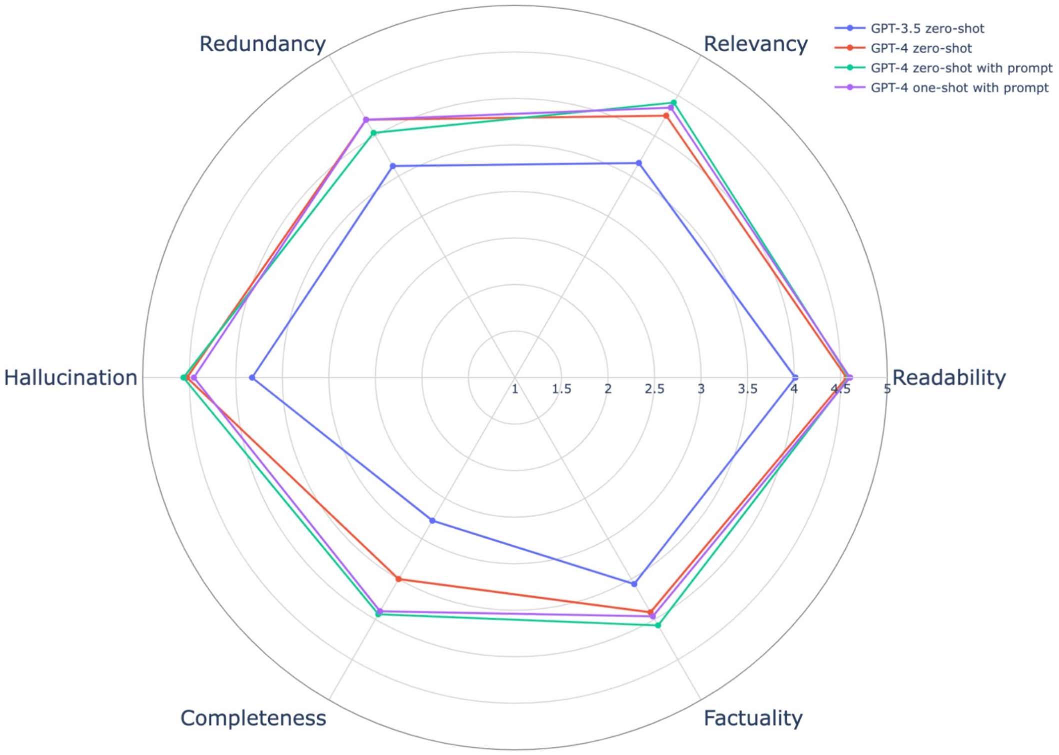
🎯 **Relevancy:** The generated text is highly relevant and directly addresses the given context or prompt.

🔍 **Redundancy:** The text is concise and free from redundancy, providing only essential information.

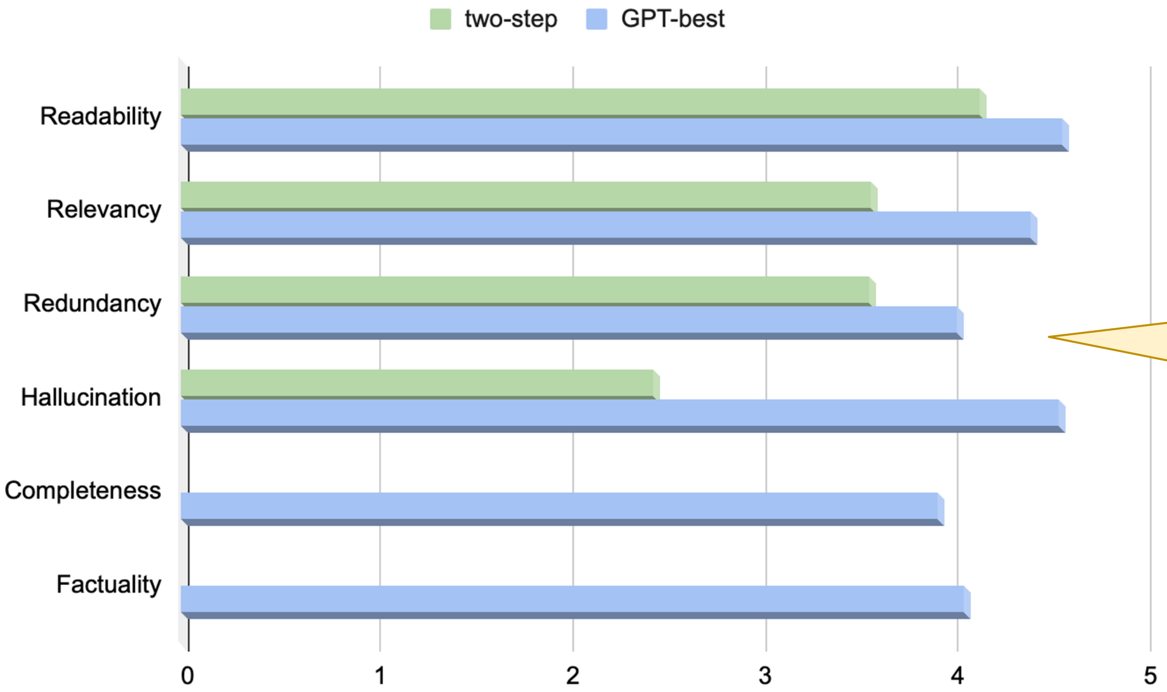
👻 **Hallucination:** The generated text is free from hallucinations and provides accurate and contextually appropriate information.

📚 **Completeness:** (compared with the ground truth) The generated text is comprehensive, accurate, and includes all relevant information.

✅ **Factuality:** The text is factually accurate, supported by evidence, and free from misinformation.



Compare with the Two-step Method



(normalized to 1-5 range)
reduced Hallucination;
improved Relevancy.

Observations: failed cases

GPT-3.5 zero-shot

Topic: Hidden Markov Models

Introduction:

This section aims to introduce the concept of Hidden Markov Models (HMMs) to the participants. HMMs are powerful statistical models used to represent systems that undergo successive probabilistic transitions. This section will briefly explain the basics of HMMs and provide the necessary background information for the subsequent sections.

Key Ideas:

This section will delve into the fundamental concepts and principles behind Hidden Markov Models.Participants will learn about the core elements of HMMs, such as states, transitions, and observations. They will also be introduced to the notion of hidden states and the probabilistic framework that characterizes HMMs.

Variations:

In this section, participants will explore the different variations and extensions of Hidden Markov Models. The section will cover topics such as continuous HMMs, time-inhomogeneous HMMs, and higher-order HMMs. Participants will gain insights into the modifications made to basic HMMs, enabling them to handle more complex real-world scenarios.

Observations: incompleteness

GPT-4 one-shot with prompt

Knowledge Graph/History:
Debuted by Google in 2012, Knowledge Graph is a synthesis of graph theory and semantic searching that targets more precise search results by understanding intent and meaning of search queries. Google introduced it to refine their search engine, understanding user queries' context and delivering relevant results. The Knowledge Graph evolution saw ontology and machine learning playing significant roles, with schema verification, information extraction and data alignment offering more precise data interpretations and integration.

Ground Truth

Knowledge Graph/History:
In 1960, Semantic Networks were invented as a knowledge representation framework. The main idea of Semantic Networks was to capture a wide range of entities including real-world objects, events, situations or abstract concepts. In 2001, Tim Berners-Lee coined the term Semantic Web which is an application of Semantic Network combined with the Web. **In 2012, Google popularized the term knowledge graph with the announcement of Knowledge Graph.** There are very few technical details about its organization, coverage and size, but it is said that it contained 70 billion connected facts by 2016. Today, knowledge graphs are used by companies from a wide range of industries.

Going Beyond the Ground Truth

GPT-4 one-shot with prompt

HMM/Introduction:

Hidden Markov Models (HMMs) are a statistical model used primarily in the field of machine learning and data analysis. Part of the broader category of Markov chain models (present category), HMMs offer a way to describe the evolution of observable events that are driven by an internal state which is not directly accessible (present features). The model's effectiveness has found it a important role in numerous applications, ranging from image recognition, speech recognition, genomics to finance (present applications). The primary motivation behind HMMs is to provide a framework to infer the missing information from observable data, where every state from the hidden sequence generates an output according to a certain probability distribution related to that state (present motivation).

Ground Truth

HMM/Introduction:

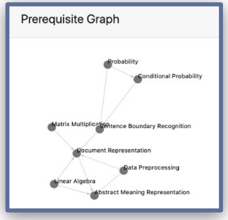
Hidden Markov models (HMMs) are a way of relating a sequence of observations to a sequence of hidden classes or hidden states that explain the observations. They are a class of a full probabilistic model—the model parameters and the overall sequence 'scores' are all probabilities. They form the foundation for creating probabilistic models of linear sequence 'labeling' problems. From just drawing an intuitive picture, HMMs offer fundamental concepts for building a complex model and are the core of many algorithms in computational sequence analysis, including genefinding, profile searches, multiple sequence alignment, and regulatory site identification.

Domain-specific LLMs ?

Better way to learn new things?

Educational Resources:
Textbooks, papers,
videos, blog posts,...
Unstructured Texts

Generic Tasks

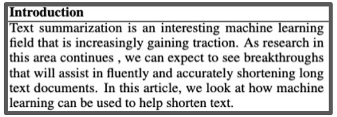


Prerequisites: updated with new concepts
Dynamic Knowledge Graph: structured

Support Downstream Applications

Learning Path

Resource Recommendation



Introduction
Text summarization is an interesting machine learning field that is increasingly gaining traction. As research in this area continues , we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. In this article, we look at how machine learning can be used to help shorten text.

Machine-generated Survey

Generate a survey on the topic "text summarization".

Text summarization is a vital area of natural language processing, focusing on condensing…

What are the main methods?

The basic idea behind LexRank is to represent sentences…

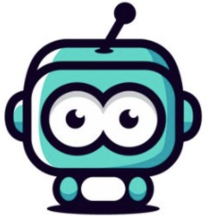
LLM-generated Survey and QA

LLMs for Domain-specific Tasks

Why?

Safety?

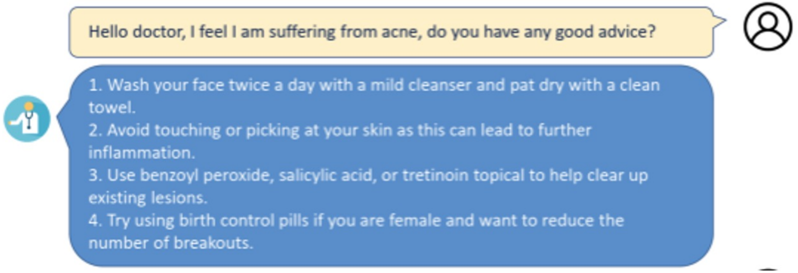
Privacy?



Personalized

(LLM) ChatGPT-like models for medical/educational applications:

- Utilizing specific domain knowledge
- Possibility to be open source
- Personalized/local LLM support



Hello doctor, I feel I am suffering from acne, do you have any good advice?

1. Wash your face twice a day with a mild cleanser and pat dry with a clean towel.
2. Avoid touching or picking at your skin as this can lead to further inflammation.
3. Use benzoyl peroxide, salicylic acid, or tretinoin topical to help clear up existing lesions.
4. Try using birth control pills if you are female and want to reduce the number of breakouts.

Example from ChatDoctor

Thank You!

http://li-lab.me

Li Lab

東京大学
THE UNIVERSITY OF TOKYO