Q&A & Code Optimization

Intro to NLP

Rahul Mishra
IIIT-Hyderabad
Apr 23, 2024

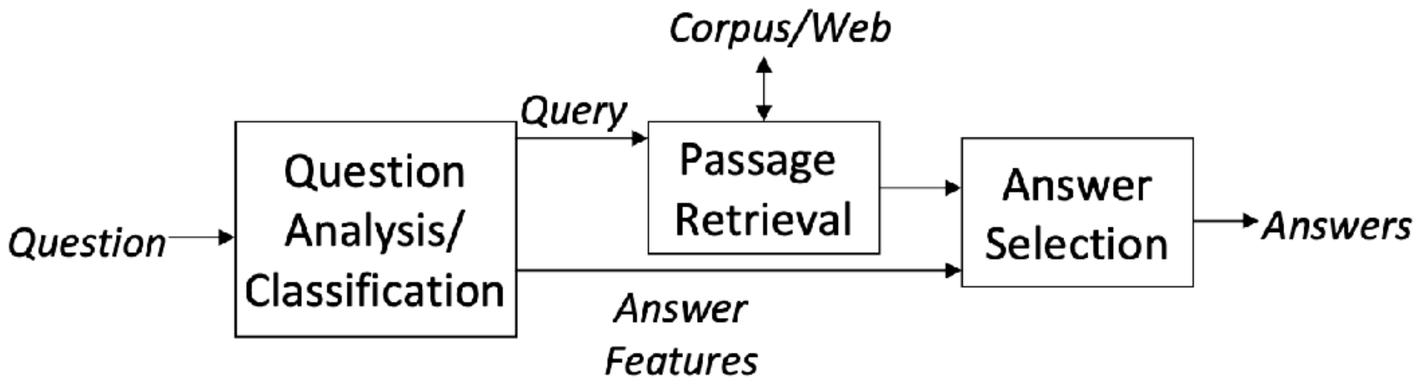
QnA

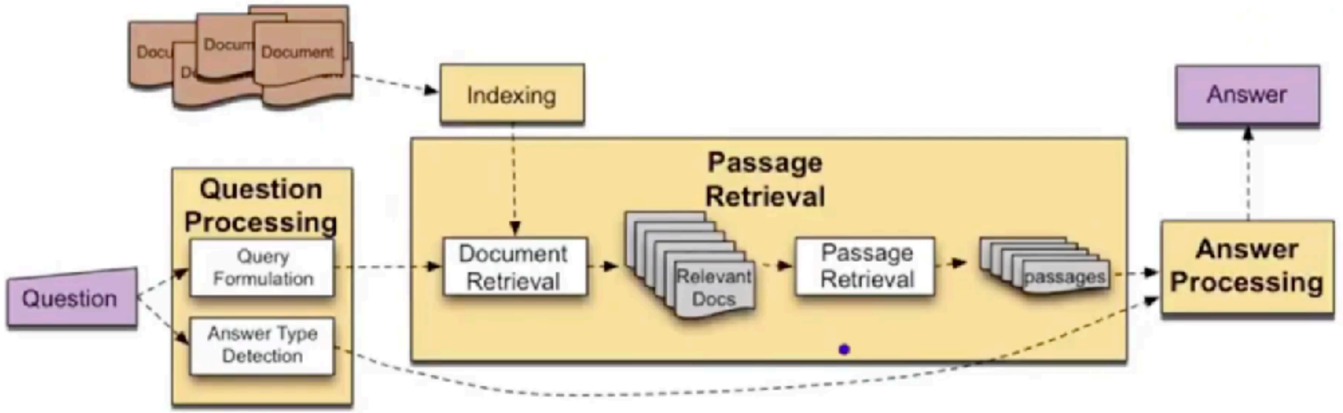
- Involves providing a specific answer to a user’s query, rather than a ranked list of documents.
- Answers can be extracted snippets from the documents or generated text snippet

- Answers can be derived from a single document or multiple documents.

The goal of question answering is to build systems that **automatically** answer questions posed by humans in a **natural language**

QnA


Corpus/Web
Query
Question
Question Analysis/ Classification
Passage Retrieval
Answer Selection
Answers
Answer Features

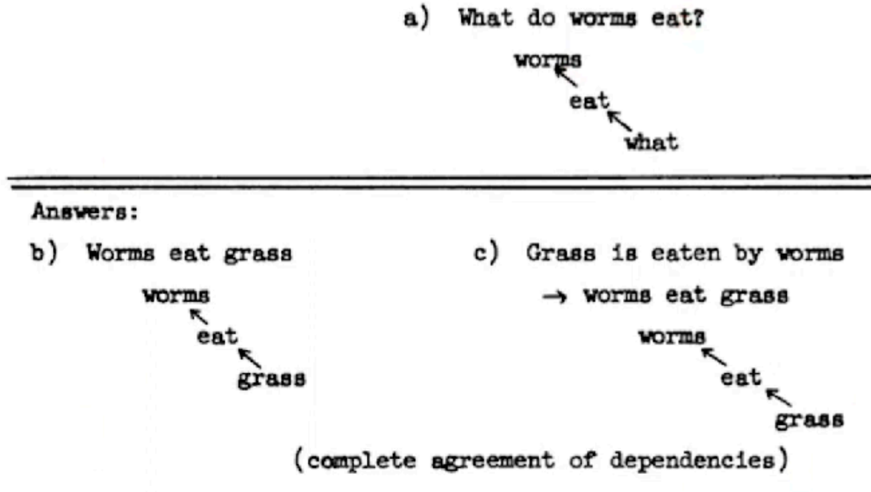

Document
Indexing
Answer
Question Processing
Query Formulation
Answer Type Detection
Question
Passage Retrieval
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

QnA: taxonomy

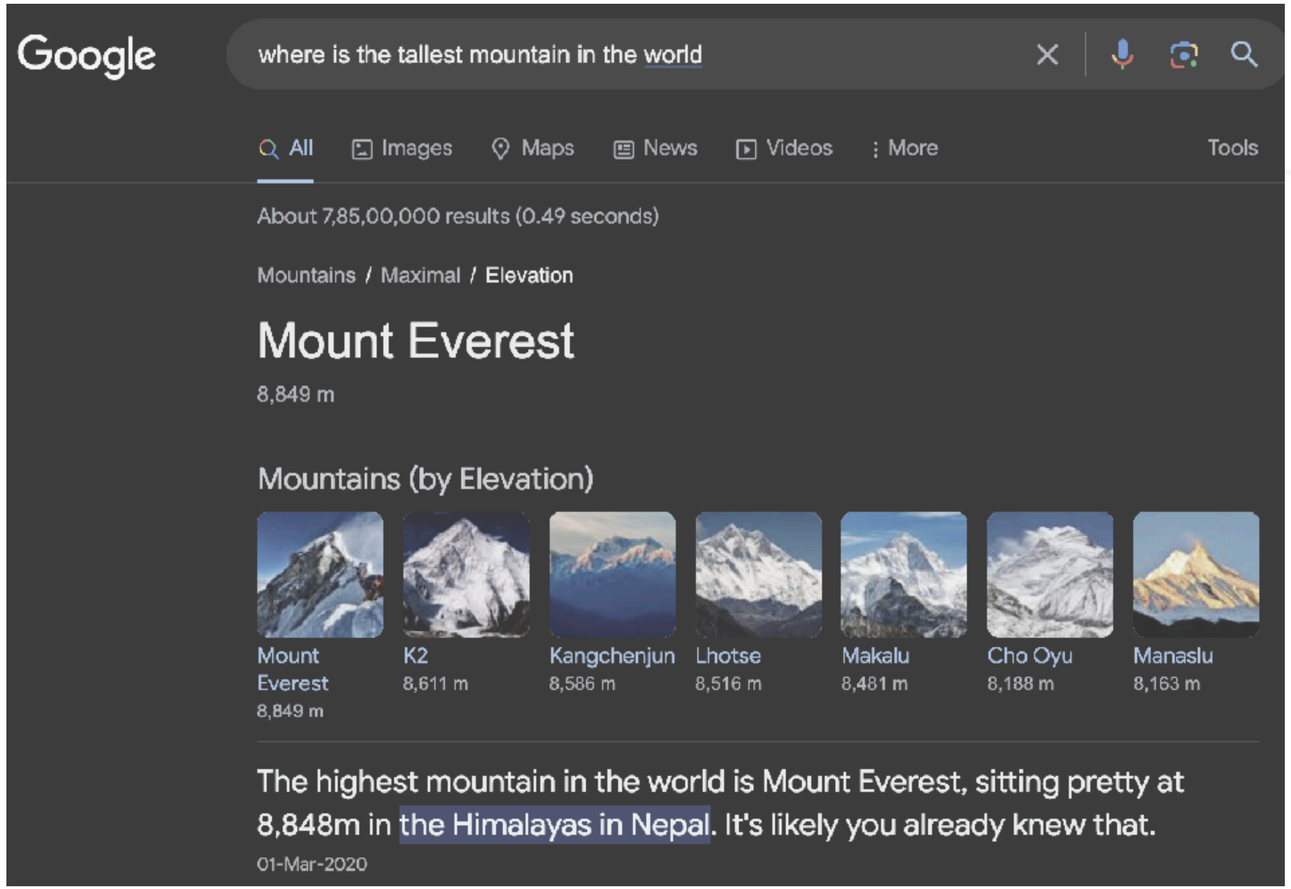
- What information source does a system build on?
  - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
  - Factoid vs non-factoid, open-domain vs closed-domain, simple .. compositional, ..
- Answer type
  - A short segment of text, a paragraph, a list, yes/no, …

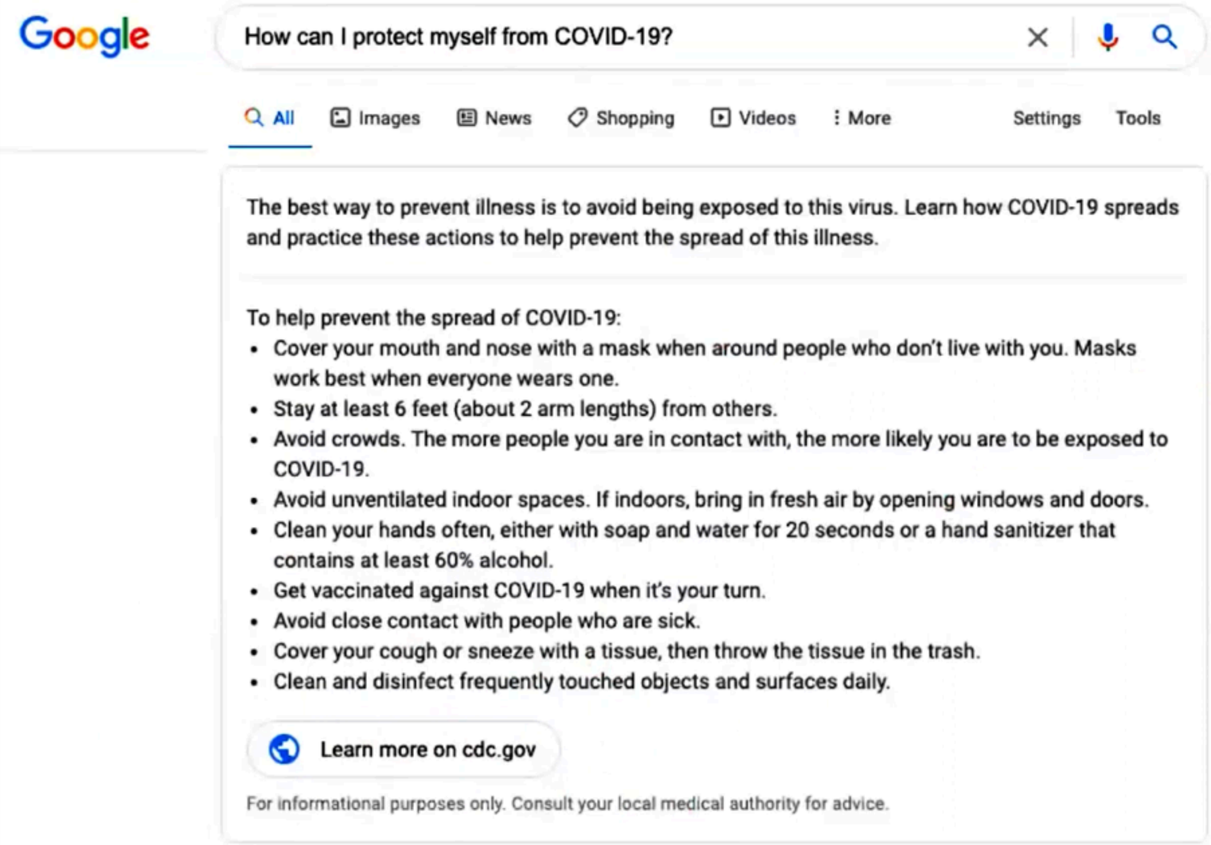
The earliest QA systems dated back to 1960s!

(Simmons et al., 1964)



Example


Google
where is the tallest mountain in the world
All
Images
Maps
News
Videos
More
Tools
About 7,85,00,000 results (0.49 seconds)
Mountains / Maximal / Elevation
Mount Everest
8,849 m
Mountains (by Elevation)
Mount Everest 8,849 m
K2 8,611 m
Kangchenjun 8,586 m
Lhotse 8,516 m
Makalu 8,481 m
Cho Oyu 8,188 m
Manaslu 8,163 m
The highest mountain in the world is Mount Everest, sitting pretty at 8,848m in the Himalayas in Nepal. It's likely you already knew that.
01-Mar-2020


Google
How can I protect myself from COVID-19?
All
Images
News
Shopping
Videos
More
Settings
Tools
The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.
To help prevent the spread of COVID-19:
Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
Stay at least 6 feet (about 2 arm lengths) from others.
Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
Get vaccinated against COVID-19 when it's your turn.
Avoid close contact with people who are sick.
Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
Clean and disinfect frequently touched objects and surfaces daily.
Learn more on cdc.gov
For informational purposes only. Consult your local medical authority for advice.

IBM Watson wins Jeopardy game

IBM Watson defeated two of Jeopardy's greatest champions in 2011

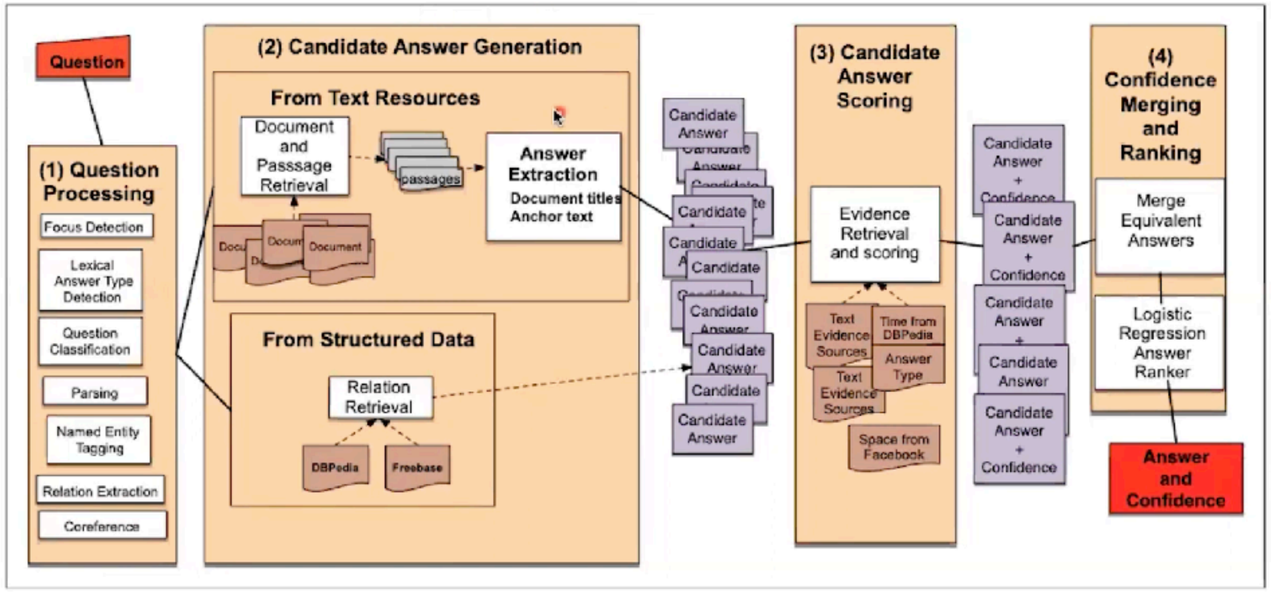


Image credit: J & M, edition 3

(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.

Deep Learning Era

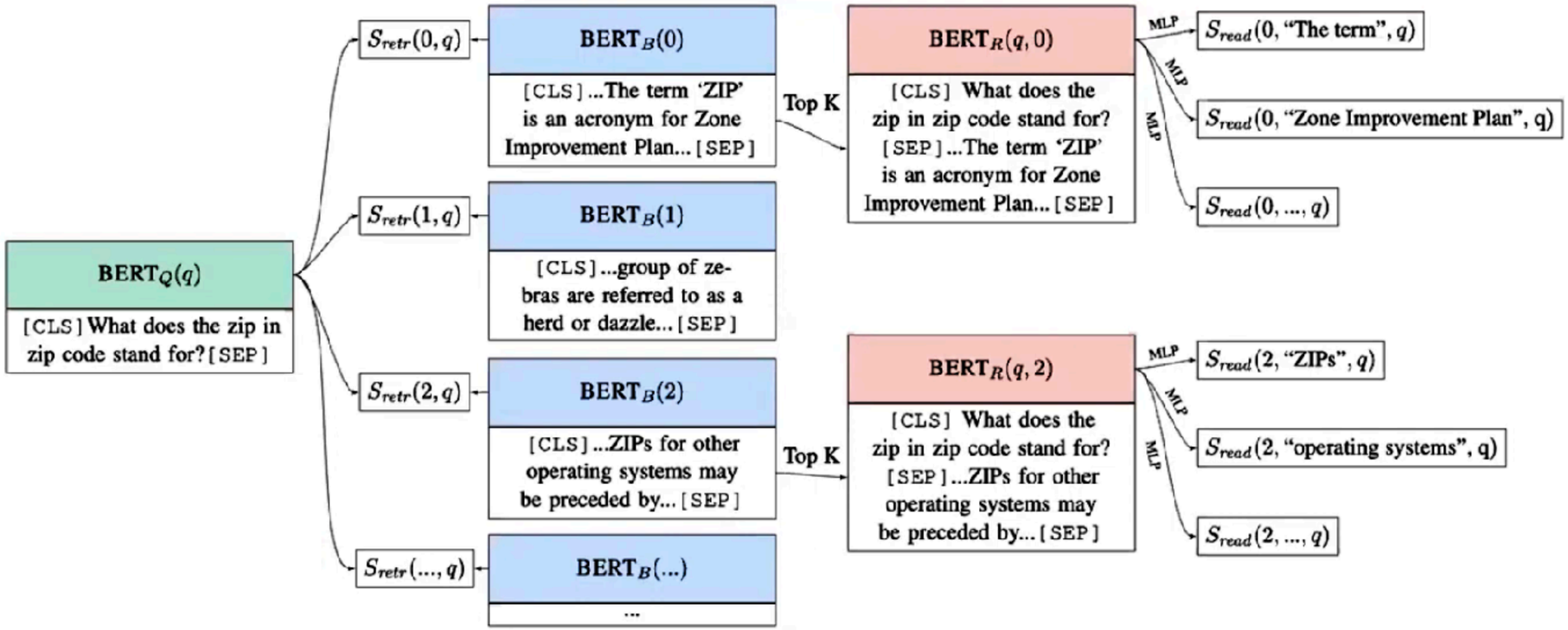
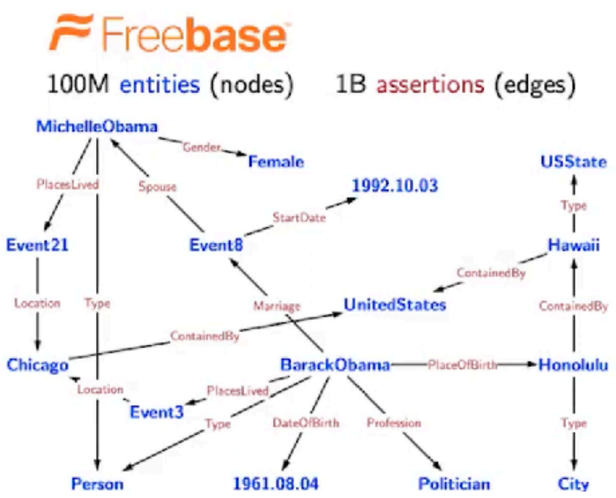


Image credit: (Lee et al., 2019)

Other kinds of QnA tasks

Knowledge based QA



Visual QA

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Reading Comprehension QnA

Reading comprehension: building systems to comprehend a passage of text and answer questions about its content $(P, Q) \longrightarrow A$

Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?

Information extraction

(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.

(Levy et al., 2017)

Semantic role labeling

UCD ***finished*** the 2006 championship as Dublin champions , by ***beating*** St Vincents in the final .

finished

- Who finished something? - UCD
- What did someone finish? - the 2006 championship
- What did someone finish something as? - Dublin champions
- How did someone finish something? - by beating St Vincents in the final

beating

- Who beat someone? - UCD
- When did someone beat someone? - in the final
- Who did someone beat? - St Vincents

(He et al., 2015)

Stanford QnA Dataset, SQuAD

- 100k annotated (passage, question, answer) triples

  Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!

- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.

  This is a limitation— not all the questions can be answered in this way!

- SQuAD still remains the most popular reading comprehension dataset; it is "almost solved" today and the state-of-the-art exceeds the estimated human performance.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Deep Learning for Comprehension QnA

LSTM-based vs BERT-based models

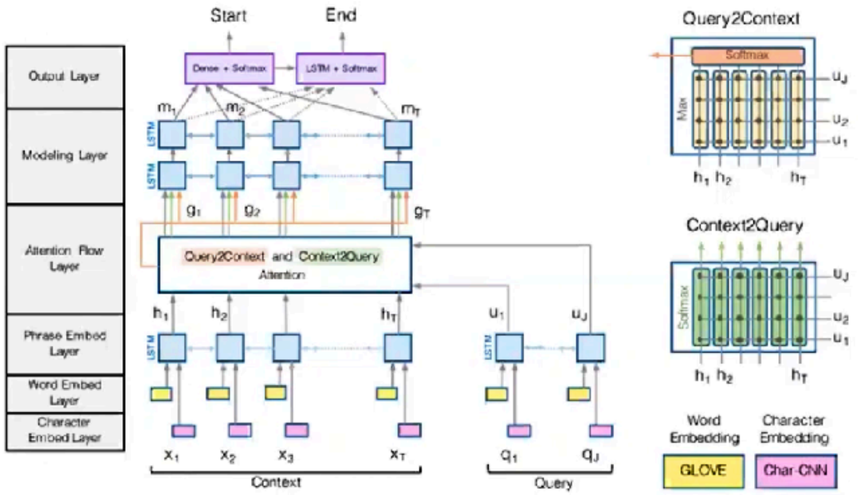


Image credit: (Seo et al, 2017)

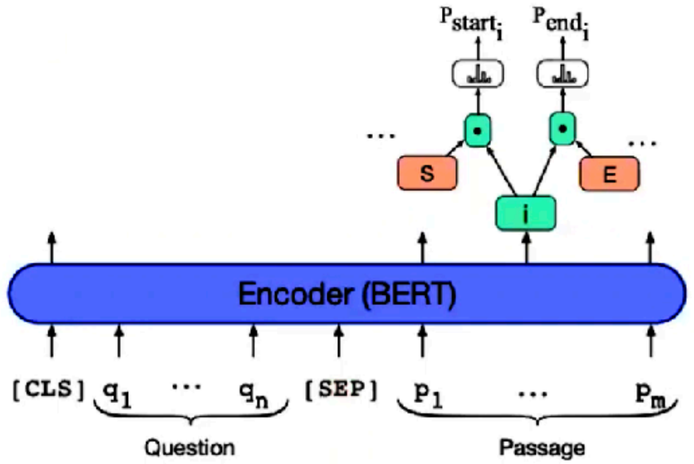
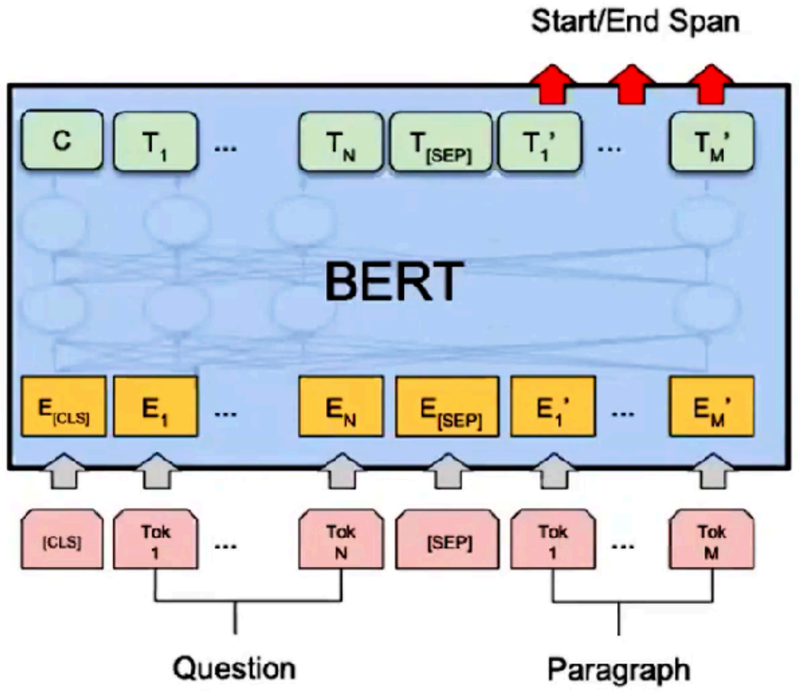


Image credit: J & M, edition 3

Pre 2019 vs Post 2019

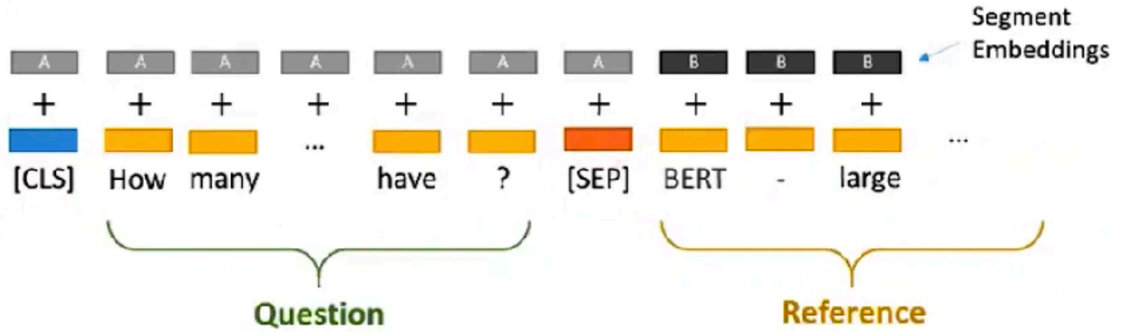
BERT for comprehension QnA



Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\mathsf{T}}\mathbf{h}_i)$$

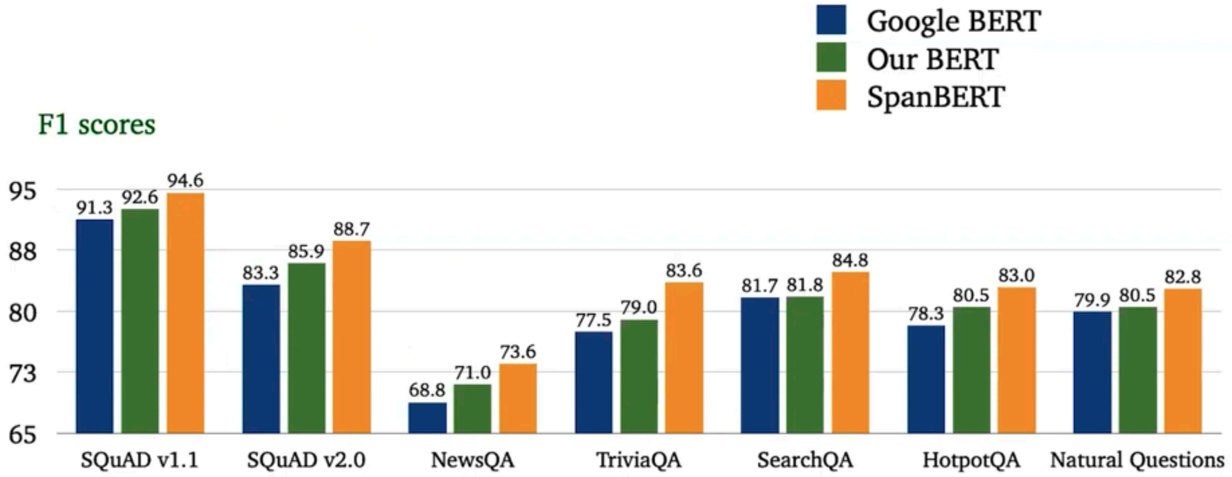
$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\mathsf{T}}\mathbf{h}_i)$$

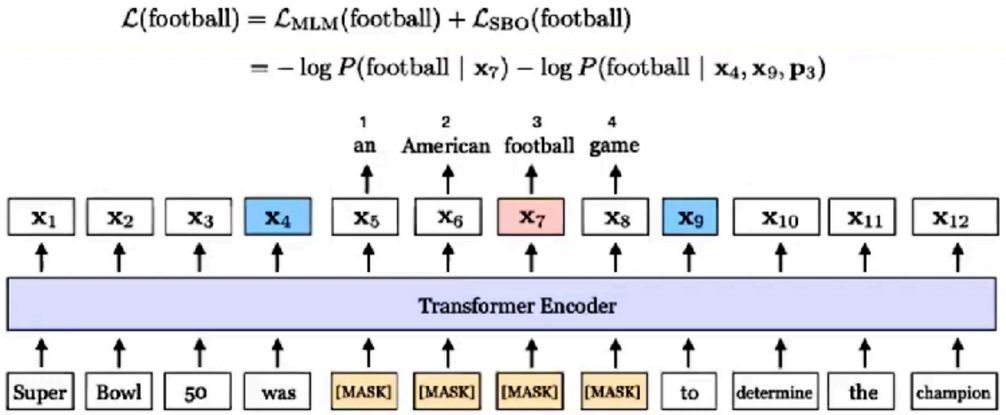
where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

Performance of LLMs

	F1	EM
Human performance	91.2*	82.3*
BiDAF	77.3	67.7
BERT-base	88.5	80.8
BERT-large	90.9	84.1
XLNet	94.5	89.0
RoBERTa	94.6	88.9
ALBERT	94.8	89.3

SpanBERT





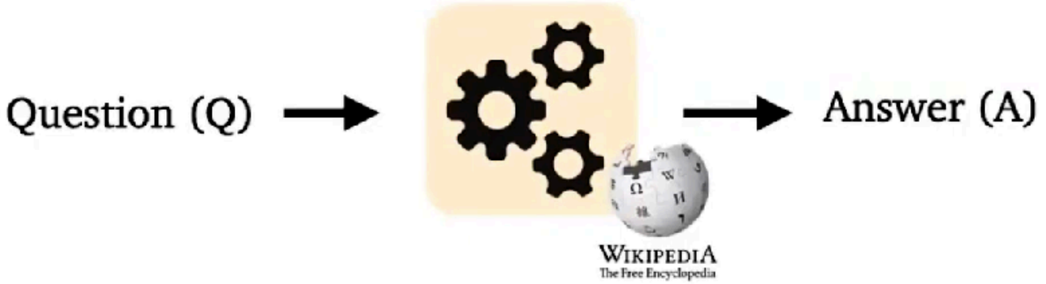
Two ideas:

1) masking contiguous spans of words instead of 15% random words

2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

$$\mathbf{y}_i = f(\mathbf{x}_{s-1}, \mathbf{x}_{e+1}, \mathbf{p}_{i-s+1})$$

Open Domain QnA



Different from reading comprehension, we don't assume a given passage.

Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.

Much more challenging but a more practical problem!

Retriever Reader Model

- Input: a large collection of documents $\mathscr{D} = D_1, D_2, \ldots, D_N$ and Q
- Output: an answer string A

- Retriever: $f(\mathscr{D}, Q) \longrightarrow P_1, \ldots, P_K$ K is pre-defined (e.g., 100)
- Reader: $g(Q, \{P_1, \ldots, P_K\}) \longrightarrow A$ A reading comprehension problem!

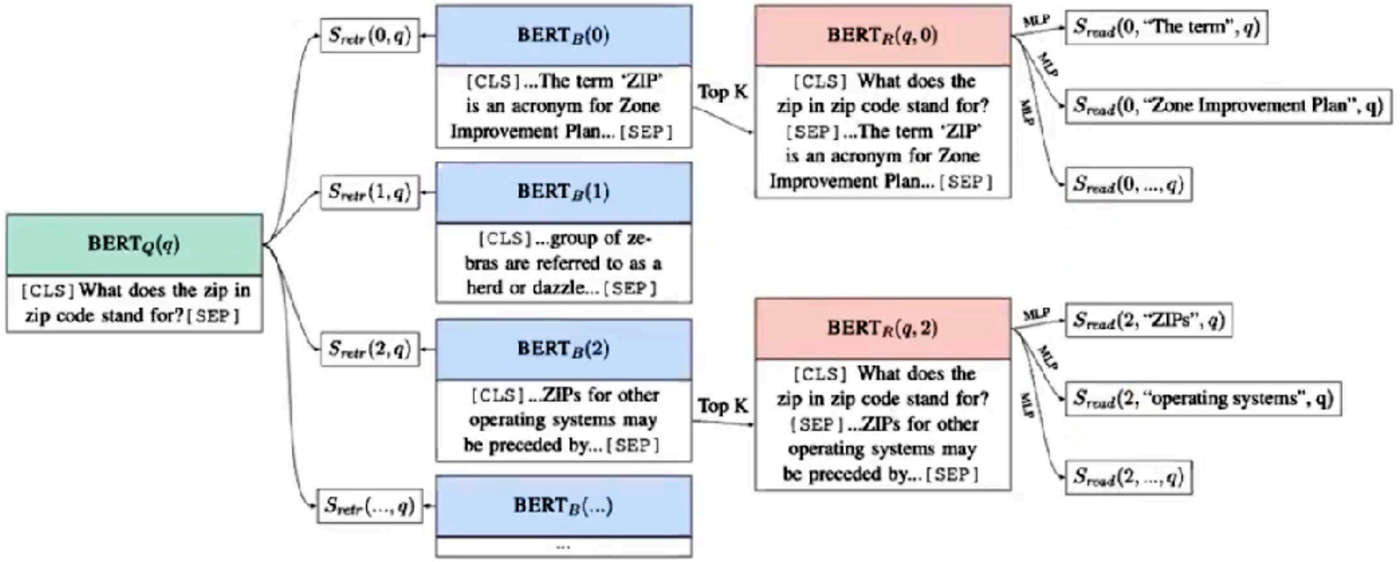
In DrQA,

- Retriever = A standard TF-IDF information-retrieval sparse model (a fixed module)
- Reader = a neural reading comprehension model that we just learned
  - Trained on SQuAD and other distantly-supervised QA datasets

Distantly-supervised examples: (Q, A) $\longrightarrow$ (P, Q, A)

Retriever Reader Model

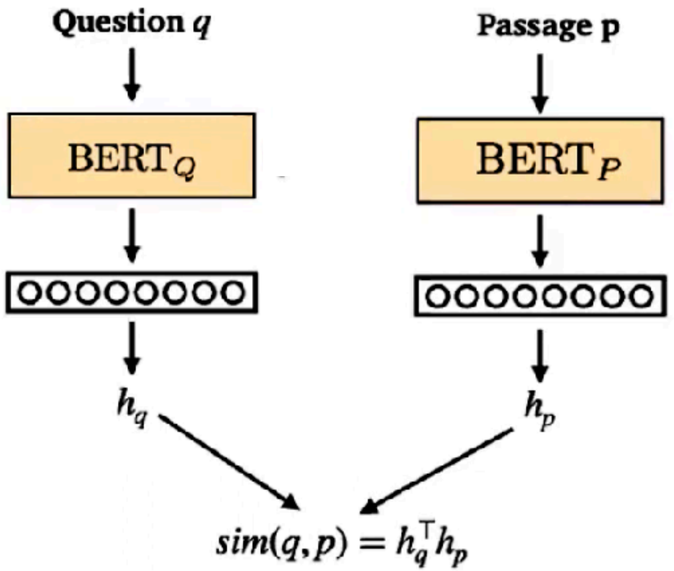
- **Joint training** of retriever and reader

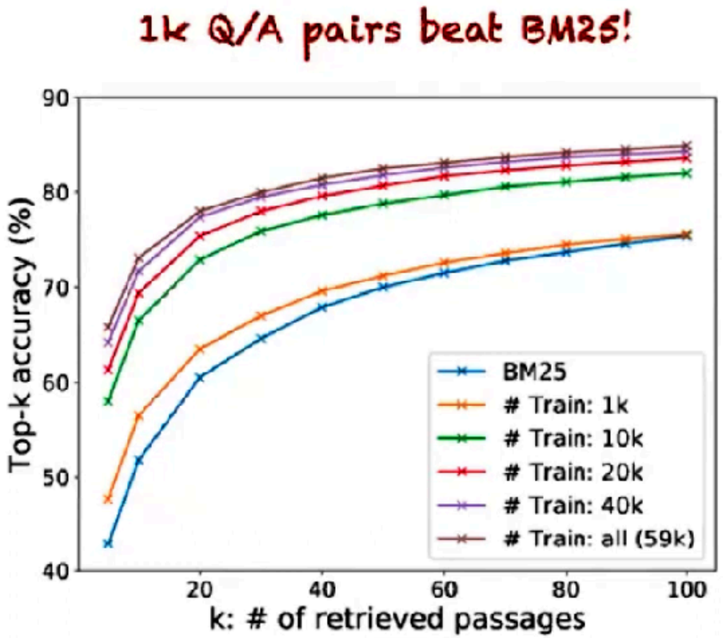


- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.

Training the retriever also

- Dense passage retrieval (DPR) - We can also just train the retriever using question-answer pairs!



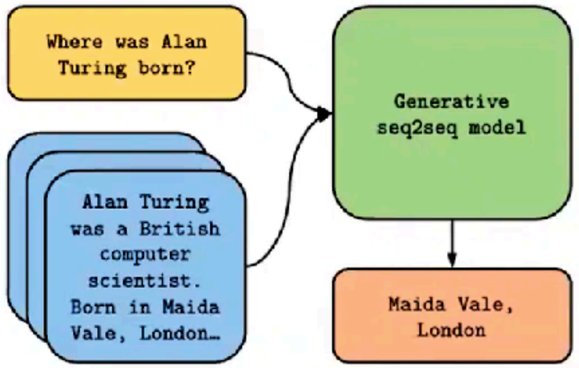


- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

Dense Retrieval + Generative Models

it is beneficial to generate answers instead of to extract answers.

Fusion-in-decoder (FID) = DPR + T5



Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

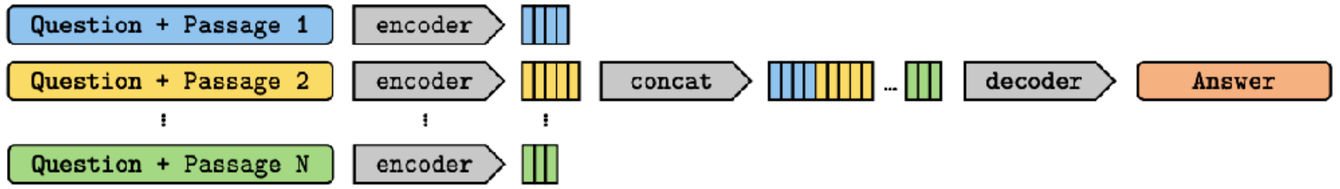
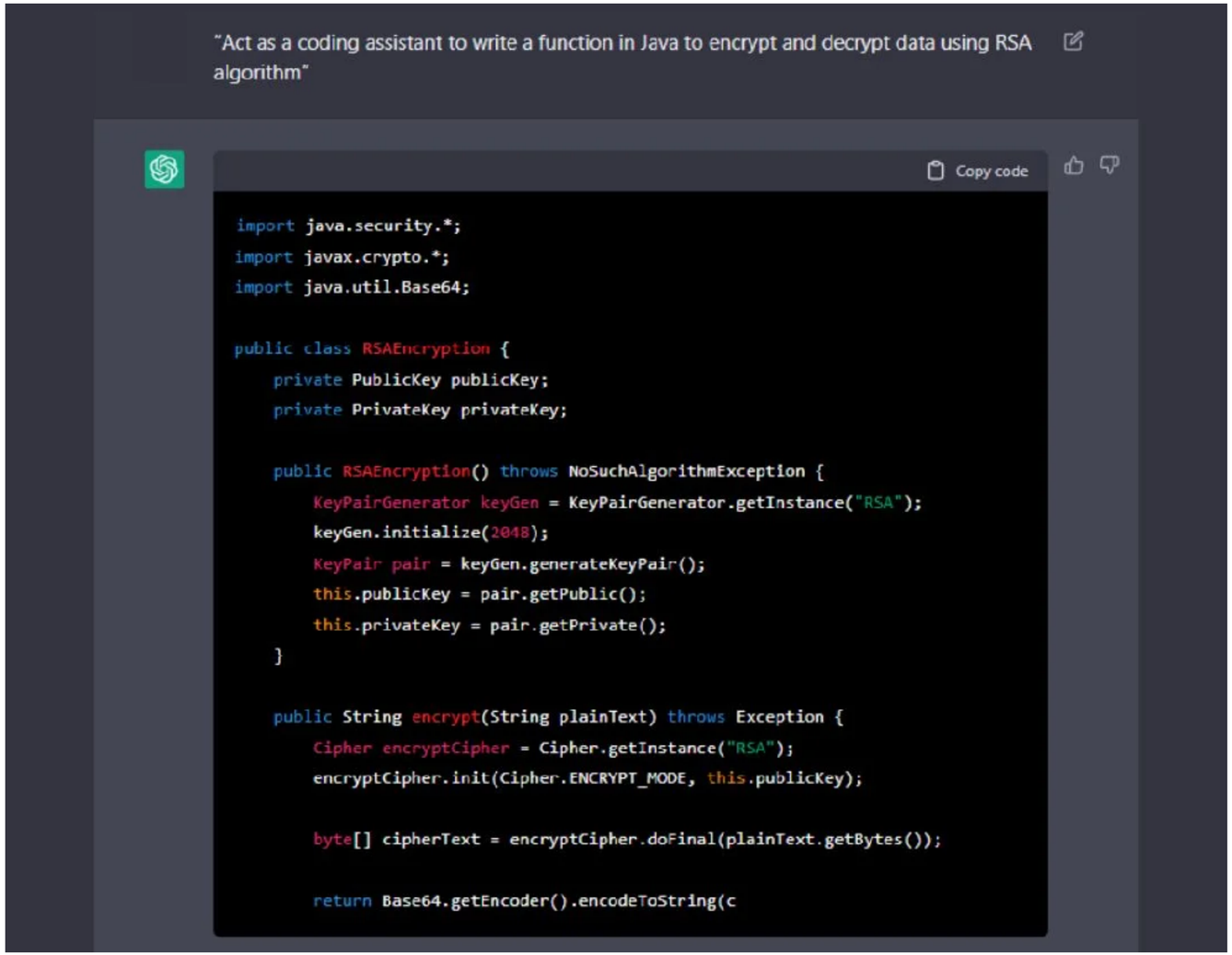


Figure 2: Architecture of the Fusion-in-Decoder method.

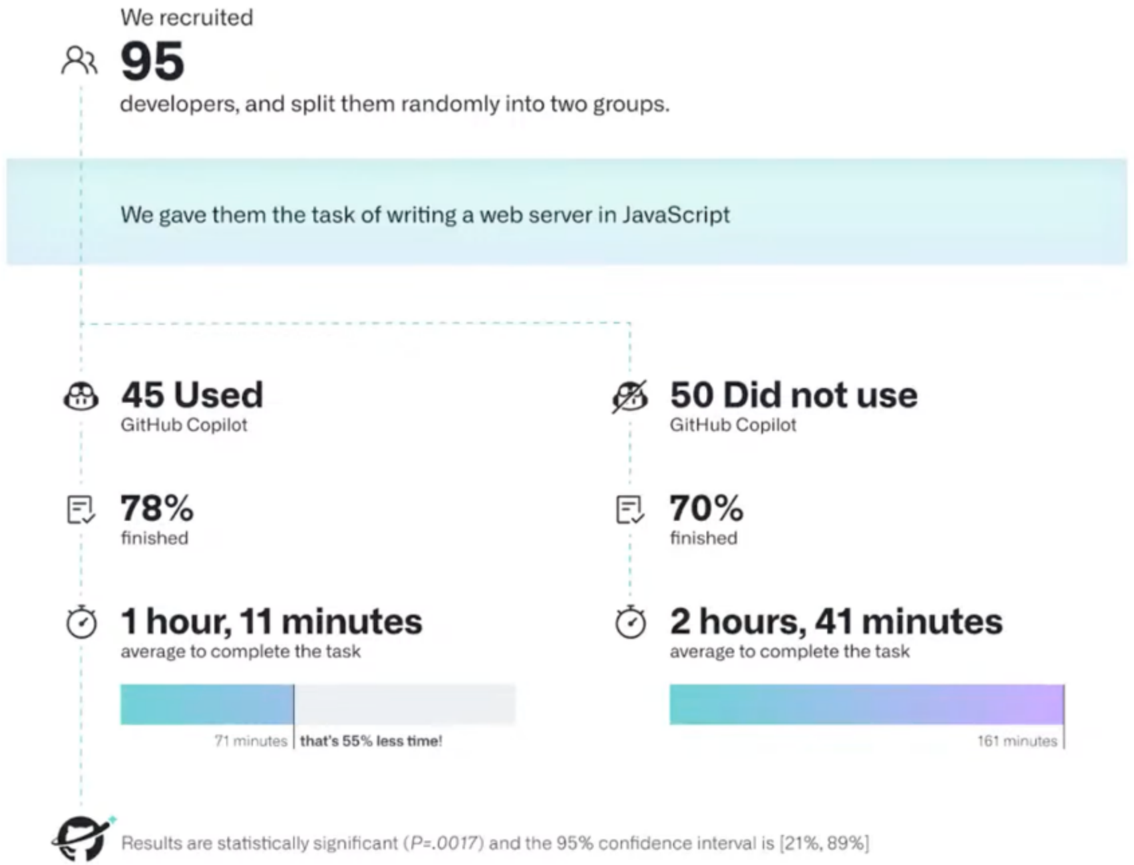
Program Optimization

LLMs can code

Surprisingly well !



Program Optimization



https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness

Program Optimization

Automated Unit Test Improvement using Large Language Models at Meta

Nadia Alshahwan*
Jubin Chheda
Anastasia Finegenova
Beliz Gokkaya
Mark Harman
Inna Harper
Alexandru Marginean
Shubho Sengupta
Eddy Wang
Meta Platforms Inc.,
Menlo Park, California, USA

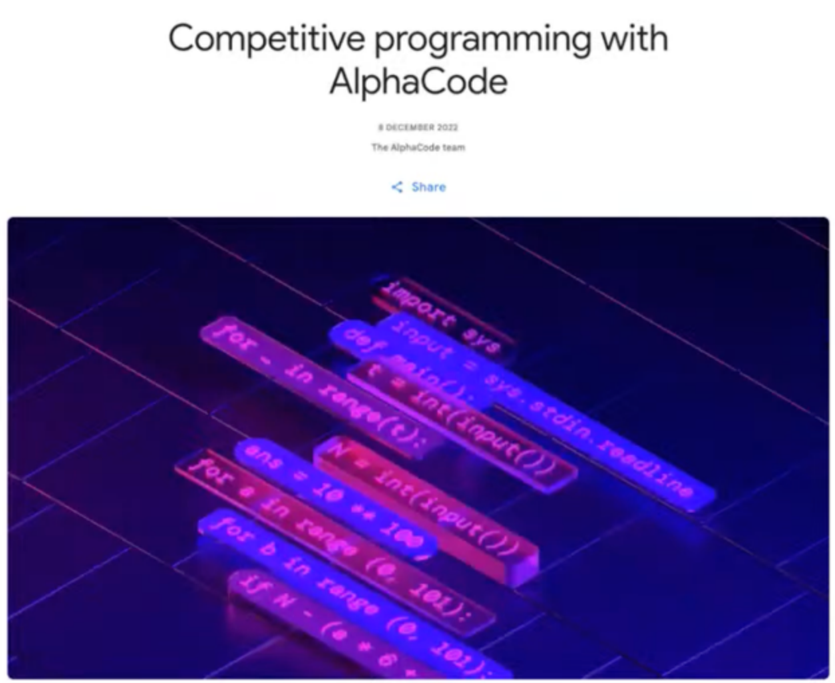
25% of TestGen-LLM's test cases increased coverage

Specialized vs Generic foundation Model

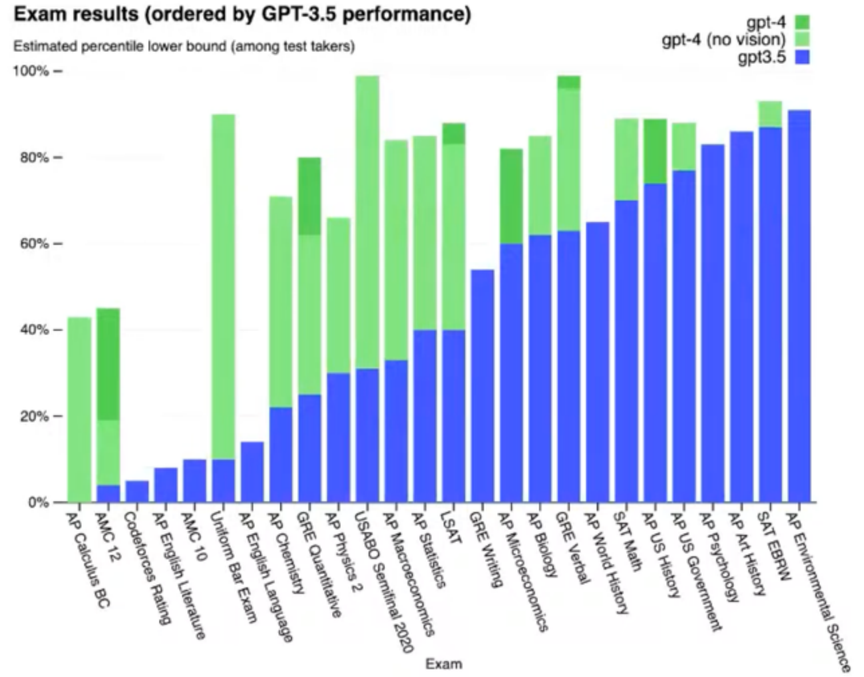
- AlphaCode training requires 2149 petaflops-days and 175 megawatt-hours (≈ 16× average household yearly energy consumption)

Can we use foundation models instead?

Specialized vs Generic foundation Model



Specialized system: 45th percentile



General purpose system: 5th percentile

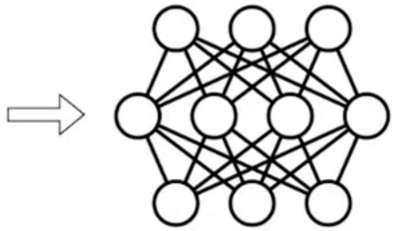
Generic models need adaptation

Code Optimization

```cpp
#include <iostream>
using namespace std;

int main(){
    int n;
    cin >> n;
    int sum = 0;
    for (int i = 1; i <= n; i++) {
        sum += i;
    }
    cout << sum << endl;
    return 0;
}
```

(a) Slower Code.



```cpp
#include <iostream>
using namespace std;

int main(){
    int n;
    cin >> n;
    cout << n*(n+1)/2 << endl;
    return 0;
}
```

(b) Faster Code.

Code Optimization

AlphaDev discovers faster sorting algorithms

7 JUNE 2023

Daniel J. Mankowitz and Andrea Michi

Share

Original

```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = min(A, D)
cmovl T S  // S = max(A, D)
cmp R P
mov P T
cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C, D)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C, D)
cmovl T Q  // Q = max(B, min(A, C, D))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C, D))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

AlphaDev

```
Memory[0] = A
Memory[1] = B
Memory[2] = C
Memory[3] = D

mov Memory[0] P  // P = A
mov Memory[1] Q  // Q = B
mov Memory[2] R  // R = C
mov Memory[3] S  // S = D

cmp S P
mov P T
cmovl S P  // P = max(C, min(A, D))
cmovl T S  // S = max(A, D)
cmp R P

cmovg R P  // P = max(C, min(A, D))
cmovl R T  // T = min(A, C)
cmp Q T
mov T U
cmovl Q U  // U = min(A, B, C)
cmovl T Q  // Q = max(B, min(A, C))

mov U Memory[0]  // = min(A, B, C, D)
mov Q Memory[1]  // = max(B, min(A, C))
mov P Memory[2]  // = max(C, min(A, D))
mov S Memory[3]  // = max(A, D)
```

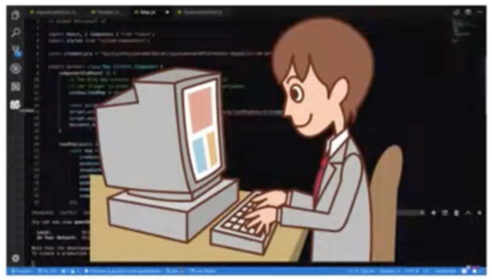
Challenges

- **Need good datasets**
  - Reinforcement learning is very computationally expensive
  - Much more efficient if we can use supervised learning

- **Unreliable execution time**
  - Stochasticity due to operating system
  - If we benchmark the same program against itself, obtain a mean speedup of $1.12\times$, with the top 5% exhibiting a speedup of $1.91\times$

Dataset

- **Performance Improving Code Edits (PIE)**



```
// Code for to_be_optimized goes here
#include <bits/stdc++.h>
using namespace std;

int main() {
    int N, len;
    cin >> N;
    string s;
    cin >> s;
    len = s.size();
    if (len > N) {
        for (int i = len; i > N; i--) {
            s.pop_back();
        }
        for (int j = 0; j < 3; j++) {
            s.push_back('.');
        }
        cout << s;
    } else {
        cout << s;
    }
    return 0;
}
```

```
// Retrieved 1-nearest prompt, slower
    src_code
#include <iostream>
#include <stack>
using namespace std;

stack<char> s;

int main() {
    int n;
    cin >> n;
    for (int i = 0; i < n; ++i) {
        char t;
        cin >> t;
        if (s.empty())
            s.push(t);
        else if (t == s.top())
            ;
        else
            s.push(t);
    }
    cout << s.size();
    return 0;
}
```

```
// Retrieved 1-nearest prompt, faster
    tgt_code
#include <cstdio>

int n, ans;
char ch1, ch2;

int main() {
    scanf("%d", &n);
    ch1 = getchar();
    ch1 = getchar();
    ans = 1;
    for (int i = 1; i < n; i++) {
        ch2 = getchar();
        if (ch2 != ch1) ans++;
        ch1 = ch2;
    }
    printf("%d", ans);
}
```

Instruction Prompting

```
Given the program below, improve its performance:

### Program:
{src_code}

### Optimized Version:
```

Raffel et al., Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Method	Model	Best@1 %Opt	Best@1 Speedup	Best@1 %Correct	Best@8 %Opt	Best@8 Speedup	Best@8 %Correct
Instruction-Only	CODELLAMA 7B	0.92%	1.01×	23.52%	5.21%	1.06×	68.30%
Instruction-Only	CODELLAMA 13B	0.41%	1.00×	10.02%	2.45%	1.03×	40.49%
Instruction-Only	CODELLAMA 34B	2.86%	1.05×	44.27%	18.92%	1.26×	84.97%
Instruction-Only	GPT-3.5	16.26%	1.20×	80.67%	39.16%	1.54×	98.77%
Instruction-Only	GPT-4	8.49%	1.15×	**93.25%**	21.17%	1.31×	98.77%

Few-shot Prompting

```
slow1 → fast1 || slow2 → fast2 || slow3 → fast3 || ... || slowN →
↪  fastN

### Program:
{src_code}

### Optimized Version:
```

Few-shot Prompting

Method	Model	Best@1 %Opt	Best@1 Speedup	Best@1 %Correct	Best@8 %Opt	Best@8 Speedup	Best@8 %Correct
Instruction-Only	CODELLAMA 7B	0.92%	1.01×	23.52%	5.21%	1.06×	68.30%
Instruction-Only	CODELLAMA 13B	0.41%	1.00×	10.02%	2.45%	1.03×	40.49%
Instruction-Only	CODELLAMA 34B	2.86%	1.05×	44.27%	18.92%	1.26×	84.97%
Instruction-Only	GPT-3.5	16.26%	1.20×	80.67%	39.16%	1.54×	98.77%
Instruction-Only	GPT-4	8.49%	1.15×	**93.25%**	21.17%	1.31×	98.77%
Few-Shot	CODELLAMA 7B	2.15%	1.02×	43.46%	9.51%	1.15×	85.07%
Few-Shot	CODELLAMA 13B	2.25%	1.02×	40.29%	13.70%	1.21×	83.03%
Few-Shot	CODELLAMA 34B	2.66%	1.02×	43.97%	13.70%	1.16×	82.62%
Few-Shot	GPT-3.5	11.45%	1.13×	80.98%	29.04%	1.38×	95.91%
Few-Shot	GPT-4	18.92%	1.25×	82.82%	36.40%	1.44×	**98.98%**

Chain-of—thought Prompting

```
### slower program:
{src_code}
Give a potential strategy improving the program:
```

LLM reasoning: The current program calculates the sum of integers from 1 to n-1 using a loop. This iterative approach results in a time complexity of O(n). The sum of the first n natural numbers follows an arithmetic progression, and by leveraging Gauss's formula, n(n-1)/2, we can compute the sum directly. This formula-based approach allows us to calculate the sum in constant time, O(1), eliminating the need for iteration.

```
### optimized version of the same code with the strategy mentioned
↪  above:
```

LLM Response: optimized code

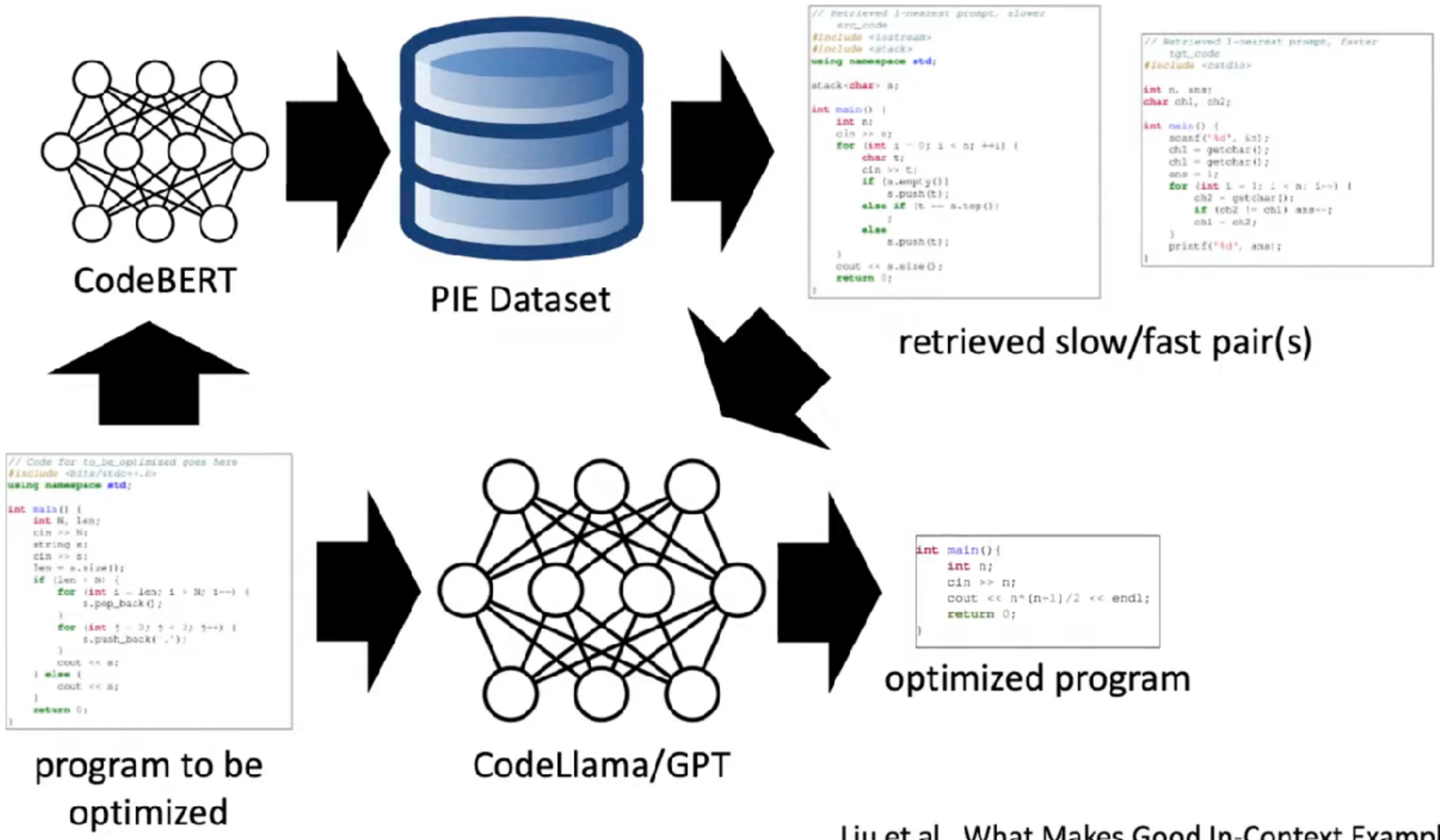
Chain-of—thought Prompting

Method	Model	Best@1 %Opt	Best@1 Speedup	Best@1 %Correct	Best@8 %Opt	Best@8 Speedup	Best@8 %Correct
Instruction-Only	CODELLAMA 7B	0.92%	1.01×	23.52%	5.21%	1.06×	68.30%
Instruction-Only	CODELLAMA 13B	0.41%	1.00×	10.02%	2.45%	1.03×	40.49%
Instruction-Only	CODELLAMA 34B	2.86%	1.05×	44.27%	18.92%	1.26×	84.97%
Instruction-Only	GPT-3.5	16.26%	1.20×	80.67%	39.16%	1.54×	98.77%
Instruction-Only	GPT-4	8.49%	1.15×	**93.25%**	21.17%	1.31×	98.77%
Few-Shot	CODELLAMA 7B	2.15%	1.02×	43.46%	9.51%	1.15×	85.07%
Few-Shot	CODELLAMA 13B	2.25%	1.02×	40.29%	13.70%	1.21×	83.03%
Few-Shot	CODELLAMA 34B	2.66%	1.02×	43.97%	13.70%	1.16×	82.62%
Few-Shot	GPT-3.5	11.45%	1.13×	80.98%	29.04%	1.38×	95.91%
Few-Shot	GPT-4	18.92%	1.25×	82.82%	36.40%	1.44×	**98.98%**
COT	CODELLAMA 7B	0.82%	1.01×	27.40%	7.46%	1.13×	73.31%
COT	CODELLAMA 13B	2.25%	1.04×	32.92%	11.15%	1.20×	79.24%
COT	CODELLAMA 34B	3.99%	1.08×	30.27%	19.63%	1.30×	78.73%
COT	GPT-3.5	21.37%	1.25×	65.95%	**43.05%**	**1.60×**	91.72%
COT	GPT-4	**26.99%**	**1.32×**	63.09%	42.74%	1.58×	84.87%

What is the problem?

- Limited knowledge about how to write faster code in pretraining data
- Few in-context examples may not be relevant
- Can we leverage the PIE dataset?

Solution



Liu et al., What Makes Good In-Context Examples for GPT-3?

Results

Model	Best@1 %Opt	Best@1 Speedup	Best@1 %Correct	Best@8 %Opt	Best@8 Speedup	Best@8 %Correct
CODELLAMA 7B	4.40%	1.13×	20.55%	16.87%	1.51×	55.32%
CODELLAMA 13B	9.10%	1.35×	28.73%	28.02%	1.97×	64.72%
CODELLAMA 34B	10.22%	1.27×	25.87%	34.25%	2.28×	63.19%
GPT-3.5	26.18%	1.58×	80.37%	48.06%	2.14×	**97.85%**
GPT-4	**50.00%**	**2.61×**	**80.57%**	**74.74%**	**3.95×**	**97.85%**