

FROM TRANSFORMERS TO LARGE LANGUAGE MODELS: A REVIEW OF THE LAST SIX YEARS

Dr. Adrian M.P. Brașoveanu
Modul University Vienna &
Modul Technology GmbH

How many generations of LLMs?

A

What are the generations of large language models since the discovery of the Transformer architecture?



Since the discovery of the Transformer architecture in 2017, there have been several generations of large language models that have advanced the field of natural language processing (NLP). These generations include:



1. First Generation (2017-2018):

- The original Transformer model was introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017. While this model laid the foundation for subsequent developments, it was not initially used for very large language models.

2. Second Generation (2018-2019):

- The second generation of large language models began with models like ELMo (Embeddings from Language Models) and OpenAI's GPT-1 (Generative Pre-trained Transformer 1), both from 2018. These models demonstrated the power of pre-training on large text corpora to improve NLP tasks.

3. Third Generation (2019-2020):

- The third generation saw the emergence of models like BERT (Bidirectional Encoder Representations from Transformers) in 2018 and GPT-2 (Generative Pre-trained Transformer 2) in 2019. BERT introduced bidirectional context to pre-training, and GPT-2 demonstrated the capabilities of large-scale generative models.

4. Fourth Generation (2020-present):

- The fourth generation includes models like GPT-3 (Generative Pre-trained Transformer 3), released by OpenAI in 2020. GPT-3 is particularly notable for its enormous size, with 175 billion parameters, and its ability to perform a wide range of NLP tasks with few-shot learning.

These generations represent key milestones in the development of large language models based on the Transformer architecture. Research in NLP continues to evolve, and there have been ongoing efforts to create even more advanced models beyond my last known update in September 2021.

Regenerate

**ChatGPT:
Three or Four!**

**The answer is
prompt-sensitive!**

GENERATIONS

How many generations of LLMs?

A

what are the 3 generations of large language models since the inception of Transformers in 2017?



Since the inception of Transformers in 2017, there have been three generations of large language models:



1. First Generation (2017-2018):

- The original Transformer model was introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017. This model laid the foundation for subsequent developments in natural language processing (NLP) and machine learning. However, the first-generation Transformers were not as large as later models and had fewer parameters.

2. Second Generation (2019-2020):

- The second generation of large language models includes models like BERT (Bidirectional Encoder Representations from Transformers), introduced by Google AI in 2018, and GPT-2 (Generative Pre-trained Transformer 2), developed by OpenAI in 2019. These models marked a significant advancement in NLP by pre-training on vast amounts of text data and achieving state-of-the-art results on various NLP tasks.

3. Third Generation (2020-present):

- The third generation of large language models builds upon the advancements of the second generation and includes models like GPT-3 (Generative Pre-trained Transformer 3), introduced by OpenAI in 2020. GPT-3 is known for its massive size, with 175 billion parameters, which allows it to generate human-like text and perform a wide range of NLP tasks, including translation, text generation, and question answering.

These three generations of models represent significant milestones in the development of large-scale Transformer-based language models, with each generation pushing the boundaries of what is possible in natural language understanding and generation. It's worth noting that research in this field continues to evolve, and there may have been further developments beyond my knowledge cutoff date in September 2021.

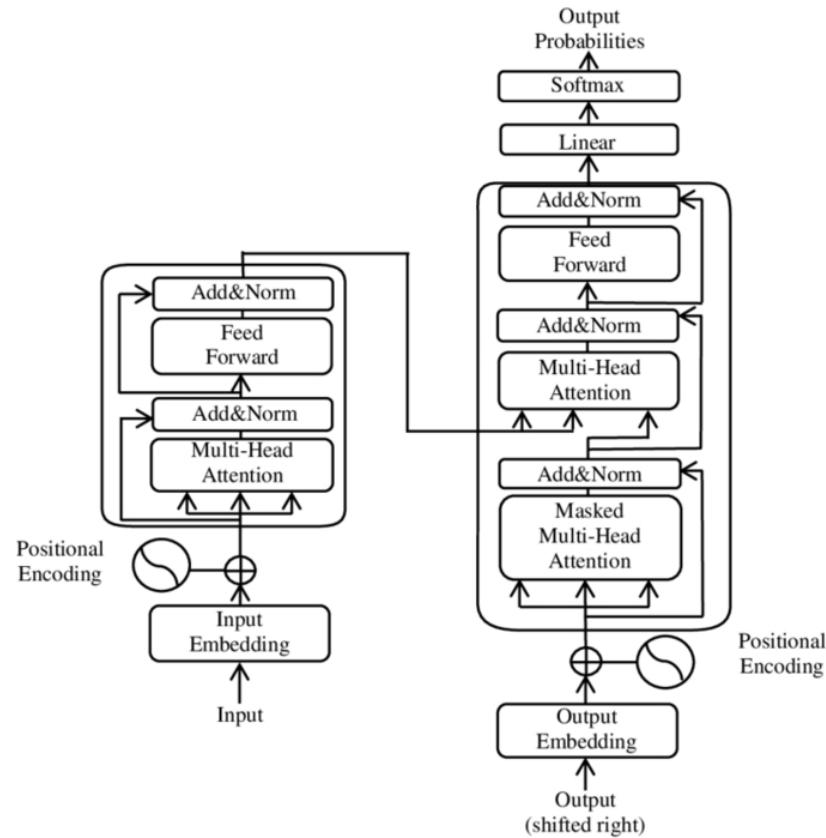
Regenerate

**ChatGPT:
Three or Four!**

**The answer is
prompt-sensitive!**

GENERATIONS

The First Generation (2017-2019)



Transformers

Transformer

BERT

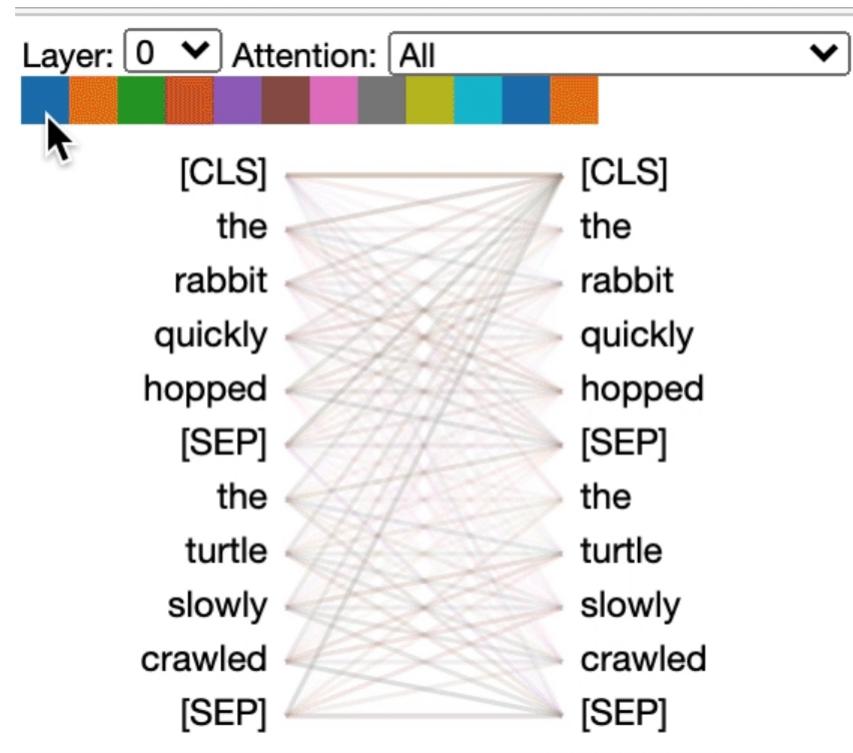
RoBERTa

AIBERT

**Core Innovation:
Transformer
Architecture**

Image Copyright © Attention is All You Need (2017) by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. NeurIPS 2017.

Do Transformers Read?



Models
BERT
GPT2
T5

Core Innovation:
Attention
Visualization

Interpretability vs.
Explainability
Debate

Image Copyright © BertViz (2018-2023) by Jesse Vig.
<https://github.com/jessevig/bertviz>

Best Transformer Resources?

harvardnlp

Members PI Code Publications

The Annotated Transformer

Apr 3, 2018

There is now a [new version](#) of this blog post updated for modern PyTorch.

```
from IPython.display import Image
Image(filename='images/aiayn.png')
```

Attention Is All You Need

1) This is our input sentence*
2) We embed each word*
3) Split into 8 heads.
We multiply X or R with weight matrices
4) Calculate attention using the resulting $Q/K/V$ matrices
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

Thinking Machines

X

R

$W_0^Q \quad W_0^K \quad W_0^V$

$W_1^Q \quad W_1^K \quad W_1^V$

$... \quad ... \quad ...$

$W_7^Q \quad W_7^K \quad W_7^V$

$Q_0 \quad K_0 \quad V_0$

$Q_1 \quad K_1 \quad V_1$

$... \quad ... \quad ...$

$Q_7 \quad K_7 \quad V_7$

Z_0

Z_1

Z_7

W^O

Z

Image Copyright © The Annotated Transformer (2018) by Harvard NLP
Image Copyright © The Illustrated Transformer (2018) by Jay Alammar

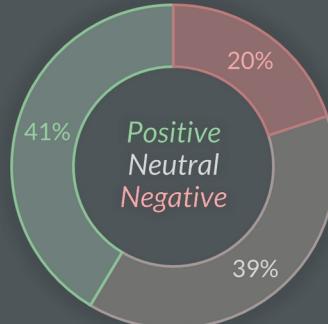
First Appearance
Attention Is All You Need by Ashish Vaswani et al. (2017)

Authors include:
Aidan Gomez
(Cohere)

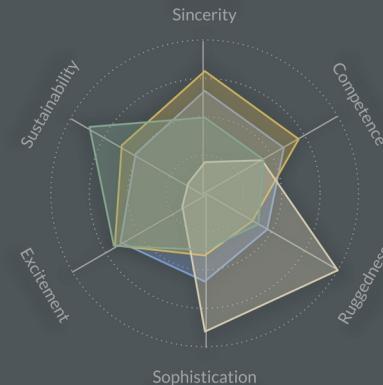
Formalization
Formal Algorithms for Transformers by Mary Phuong and Marcus Hutter (2022)

Domain-Specific Sentiment Analysis

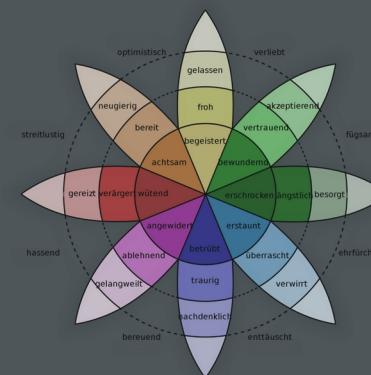
Sentiment



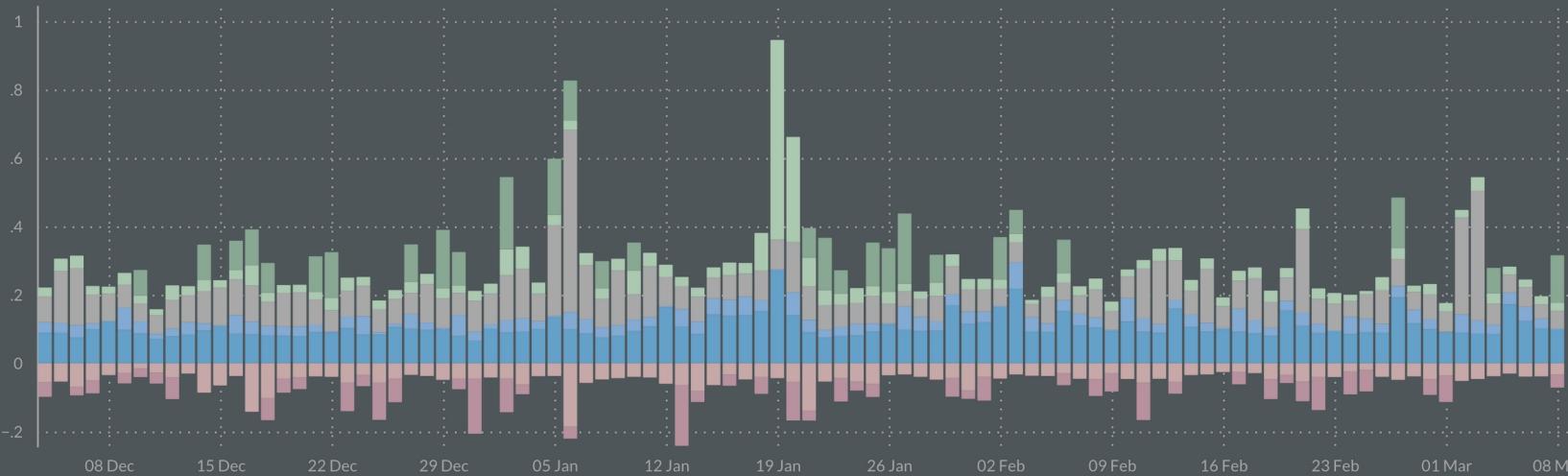
Brand Personality



Wheel of Emotions



Hourglass of Emotions



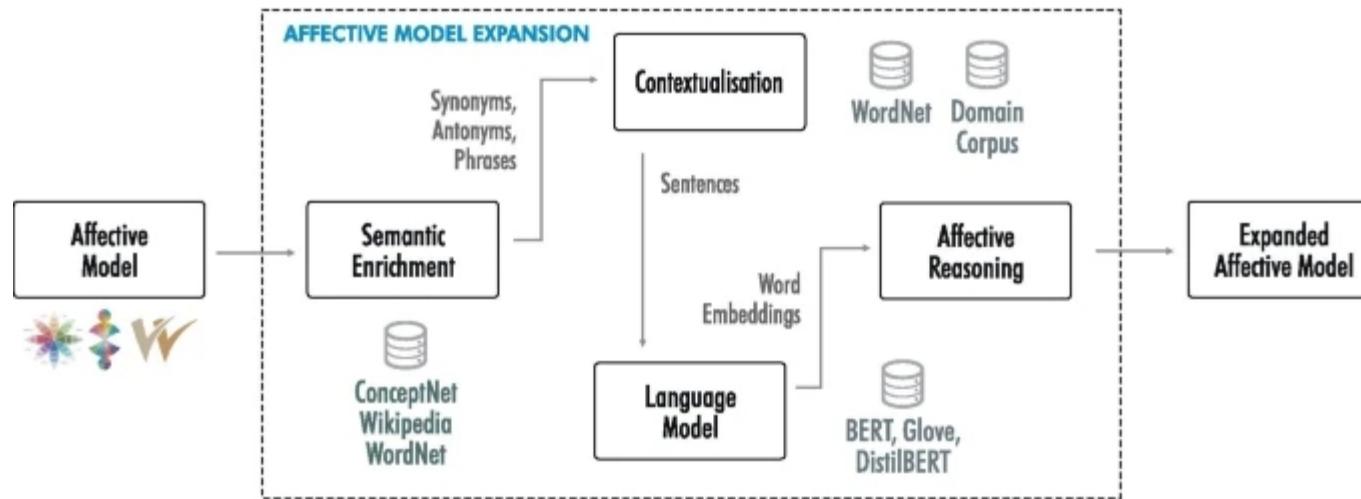
WYSDOM
Success Metric



- Desired
- Positive
- Neutral
- Pageviews
- Visits
- Negative
- Undesired

Domain-Specific Sentiment Analysis

Expanded Model



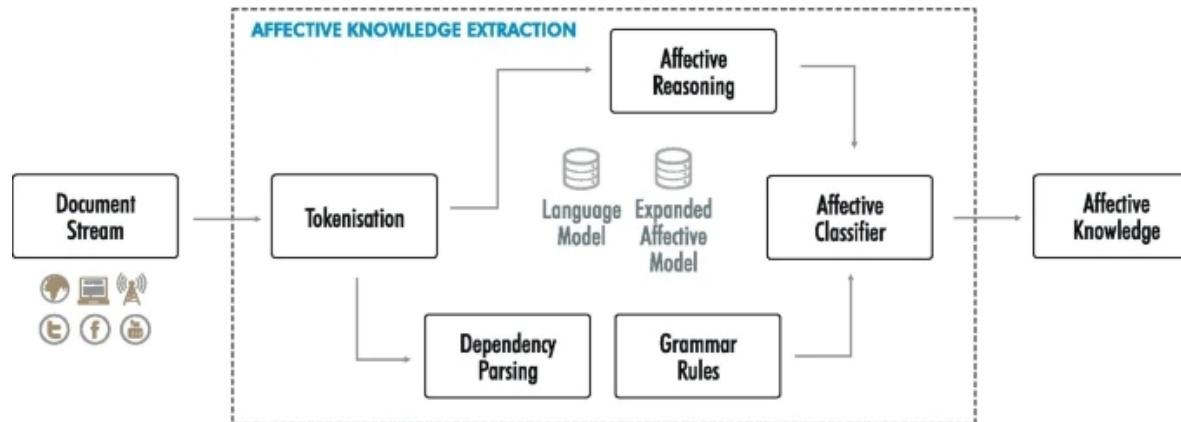
KG
Wordnet
ConceptNet

Word Sense Disambiguation
Basic Algorithm
Synonyms
Antonyms
Negation

Language Models
BERT
DistilBERT

COGN2022 Slides – Images Copyright © Automatic Expansion of Domain-Specific Affective Models for Web Intelligence Applications by Albert Weichselbraun, Jakob Steixner, Adrian M.P. Brașoveanu, Arno Scharl, Max C. Gobel and Lyndon J.B. Nixon . Cognitive Computation 14(1), 228-245 (2022).

Domain-Specific Sentiment Analysis



category		BERT	DistilBERT	BERT+GR	DistilBERT+GR
T+	calmness	0.45/0.61/0.52	0.50/0.39/0.44	0.52/0.76/0.62	0.71/0.63/0.67
T-	anger	0.75/0.57/0.65	0.72/0.71/0.71	0.87/0.57/0.69	0.84/0.74/0.79
I+	joy	0.66/0.35/0.46	0.73/0.42/0.53	0.65/0.43/0.52	0.71/0.38/0.50
I-	sadness	0.60/0.79/0.69	0.65/0.84/0.73	0.63/0.77/0.69	0.66/0.81/0.73
A+	pleasantness	0.62/0.55/0.58	0.66/0.61/0.64	0.65/0.60/0.62	0.69/0.58/0.63
A-	disgust	0.65/0.64/0.65	0.71/0.68/0.69	0.69/0.64/0.67	0.73/0.70/0.72
S+	eagerness	0.71/0.43/0.54	0.69/0.39/0.50	0.68/0.50/0.58	0.63/0.39/0.47
S-	fear	0.64/0.85/0.73	0.64/0.85/0.73	0.67/0.82/0.74	0.61/0.76/0.68
overall		0.64/0.58/0.61	0.67/0.60/0.63	0.70/0.61/0.65	0.68/0.62/0.65

Affective Categories

Introspection
(joy-sadness)

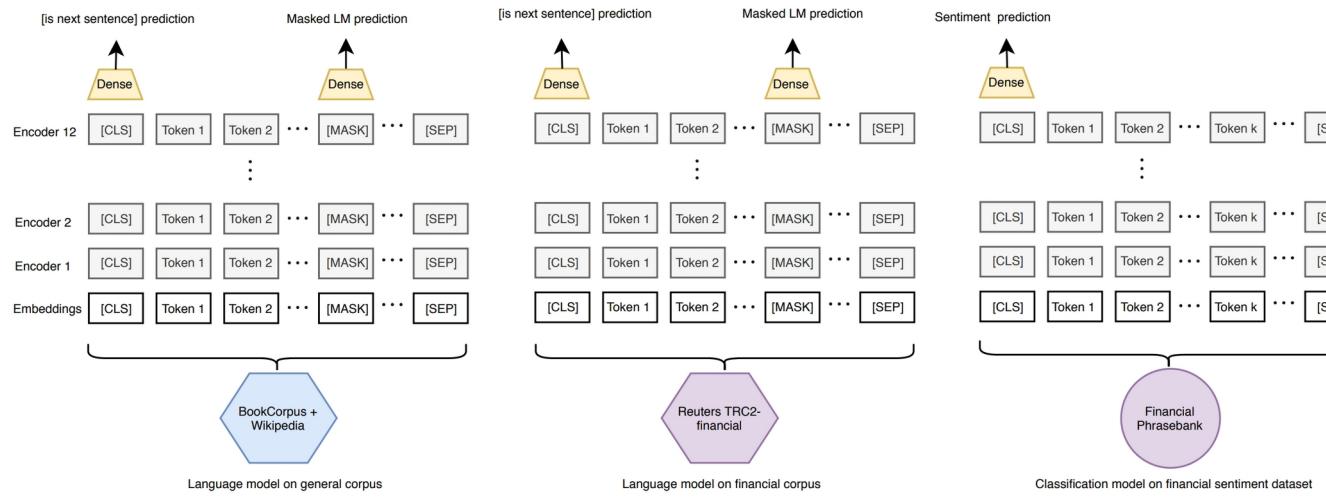
Temper
(calmness-anger)

Attitude
(pleasantness-disgust)

Sensitivity
(eagerness-fear)

**GR – Grammar
Rules and
Dependency
Parsing**

The Second Generation (2019-2022)



Domain Expertise
FinBERT
BioBERT
ClinicalBERT
SciBERT
LegalBERT
ERNIE

Core Innovation:
Domain Adaptation

Image Copyright © FinBERT: Financial Sentiment Analysis with Pre-trained Language Models(2019) by Dogu Tan Araci

The Second Generation (2019-2022)

The three settings we explore for in-context learning

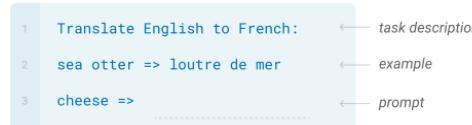
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.



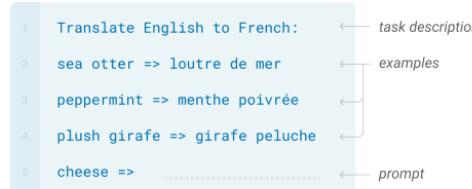
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Early LLMs

GPT-3 (2020) -
175B

**Core Innovation:
In-Context
Learning,
Instruction Fine-
Tuning &
Prompting**

Image Copyright © Language Models are Few-Shot Learners (2020) by Tom B. Brown et al. NeurIPS 2020.

Second Generation Resources?

DeepLearning.AI

Natural Language Processing Specialization

Break into NLP. Master cutting-edge NLP techniques through four hands-on courses! Updated with the latest techniques in October '21.

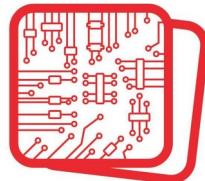


Instructors: Lukasz Kaiser +2 more

Enroll for Free
Starts Sep 13

Financial aid available

109,148 already enrolled



AHEAD
OF AI

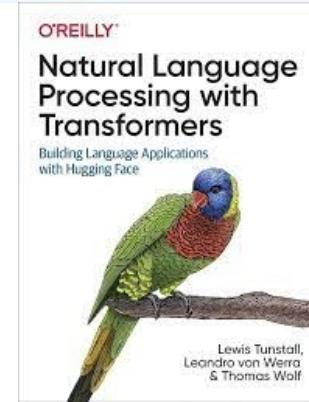


Image Copyright © Coursera DeepLearning.AI Natural Language Processing Specialization (2023)

Image Copyright © Ahead of AI / Sebastian Raschka (2023)

Image Copyright © O'Reilly (2022)

Repositories

HuggingFace
(models)
dair-ai (papers)

Annotated T5

Exploring the
Limits of Transfer
Learning with a
Unified Text-to-
Text Transformer
by Colin Raffel
(YouTube video)

TRANSFORMERS

Economics, Domain Knowledge and Trends for Predicting Crude Oil Markets

Table 1 Expected and actual FinBERT sentiment scores (Pos. = Green, Neu. = Yellow, Neg. = Red)

Headlines			Sentiment Score Expected	Sentiment Score FinBERT
Shortage	Supply Decrease	Major Explosion, Fire at Oil Refinery in Southeast Philadelphia	Positive	-0.886292
		PETROLEOS confirms Gulf of Mexico oil platform accident	Positive	-0.507213
		CASUALTIES FEARED AT OIL ACCIDENT NEAR IRANS BORDER	Positive	-0.901763
	Demand Increase	EIA Chief expects Global Oil Demand Growth 1 M B/D to 2011	Positive	0.930822
		Turkey Jan-Oct Crude Imports +98.5% To 57.9M MT	Positive	0.866315
		China's crude oil imports up 78.30% in February 2019	Positive	0.922963
Surplus	Demand Decrease	China February Crude Imports -16.0% On Year	Negative	0.540711
		Turkey May Crude Imports down 11.0% On Year	Negative	-0.965965
		Japan June Crude Oil Imports decrease 10.9% On Yr	Negative	-0.955271
	Supply Increase	Iran's Feb Oil Exports +20.9% On Mo at 1.56M B/D - Official	Negative	0.139093
		Apache announces large petroleum discovery in Philadelphia	Negative	0.089624
		Turkey finds oil near Syria, Iraq border	Negative	0.076210

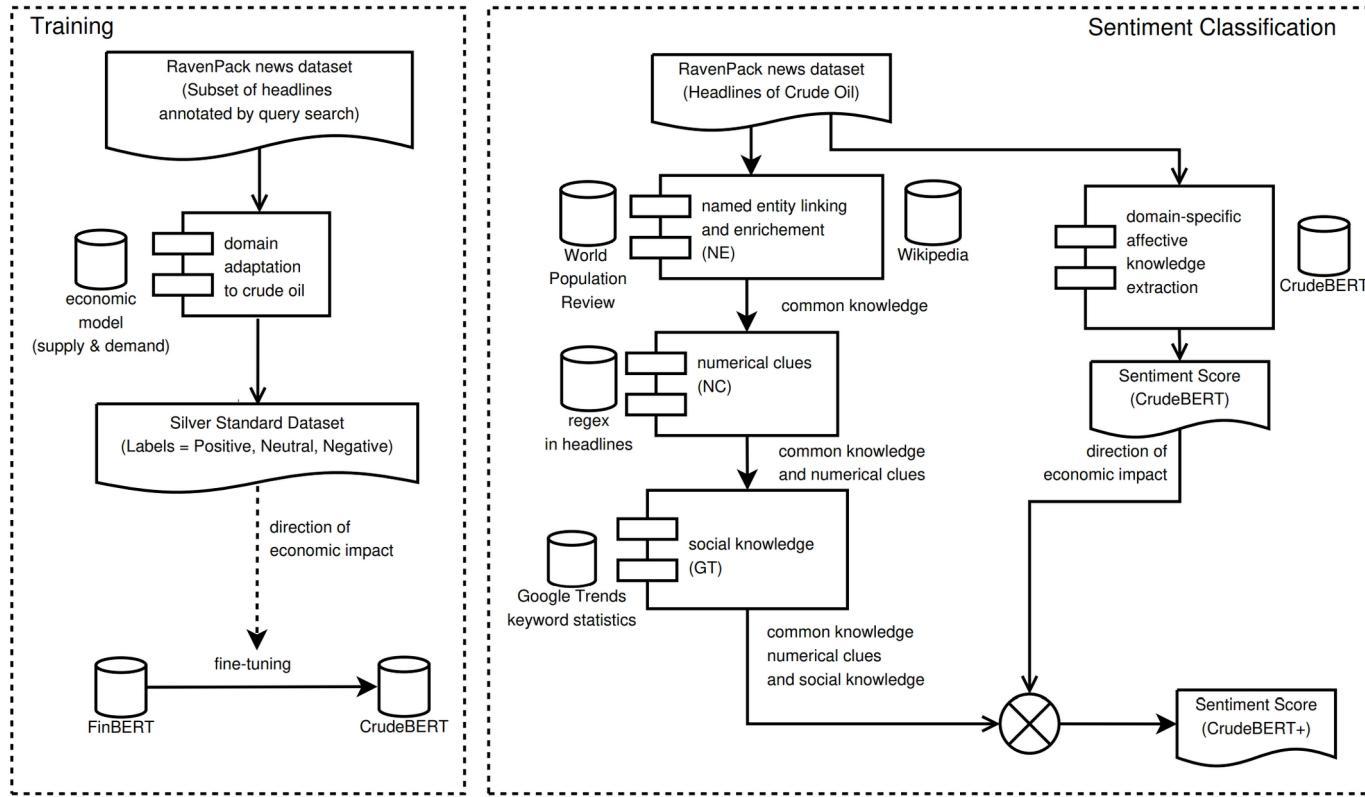
Economics
Supply and
Demand

Domain
Entities
Numerical Cues

Social
Knowledge
Google Trends

COGN2023 Slides – Images Copyright © Integrating Economic Theory, Domain Knowledge, and Social Knowledge into Hybrid Sentiment Models for Predicting Crude Oil Markets by Himmet Kaplan, Albert Weichselbraun and Adrian M.P. Brașoveanu. Cognitive Computation 15(4): 1355-1371 (2023).

CrudeBERT+



Training
Small corpora of economic news headlines

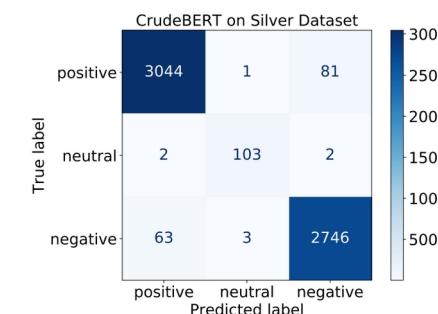
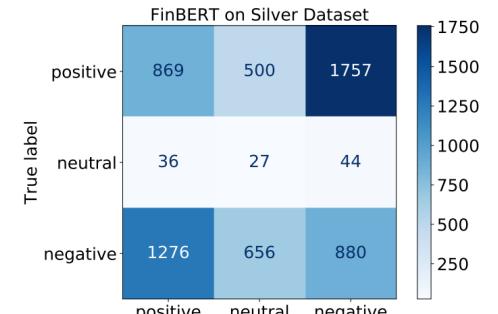
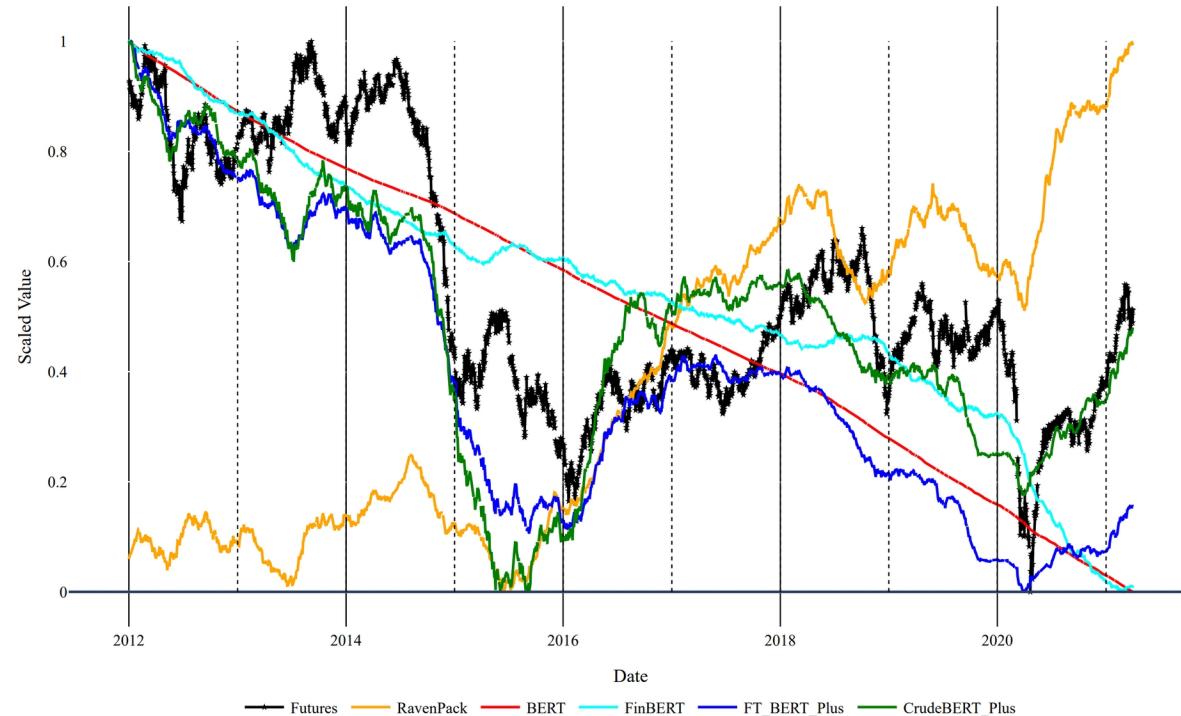
Main Idea
Even headlines are enough to get great results

Up to 2x improvements (next slide).

DTW evaluation as sequences are not perfectly aligned (next slide).

Matched Country = Turkey
Numerical Clue = up by 0.783
Turkey's crude oil imports up 78.3% since 2021
Matched Topic = IMPORT
Matched Polarity = INCREASE

Dynamic Time Warping Evaluation



DTW-Distance to WTI FUTURES	Yearly									Overall 2012 2021
	2012	2013	2014	2015	2016	2017	2018	2019	2020	
RavenPack	506	541	442	188	18	162	103	171	360	2258
BERT	85	40	64	219	184	66	81	147	124	608
FT_BERT	56	21	58	214	219	247	72	89	45	414
FT_BERT_Plus	34	90	85	42	12	32	83	192	131	360
FinBERT	137	70	72	330	270	126	31	118	110	598
CrudeBERT	114	111	32	108	199	311	135	40	133	635
CrudeBERT_Plus	47	59	67	103	57	136	29	46	39	279

The Third Generation (2022-Present)

A

1. Create a list of 3 science fiction books
2. Create a list of ratings for these books
3. Combine them to make a json object that contains the books with their ratings.

Certainly! Here's how you can create a list of 3 science fiction books, a list of ratings for these books, and then combine them to make a JSON object that contains the books with their ratings:

```
python
import json

# Step 1: Create a list of 3 science fiction books
science_fiction_books = [
    "Dune",
    "Foundation",
    "Neuromancer"
]

# Step 2: Create a list of ratings for these books
ratings = [
    4.6,
    4.4,
    4.2
]

# Step 3: Combine them to make a JSON object
science_fiction_book_ratings = {}
for i in range(3):
    science_fiction_book_ratings[science_fiction_books[i]] = ratings[i]

# Convert the dictionary to a JSON object
science_fiction_book_ratings_json = json.dumps(science_fiction_book_ratings)
```

Copy code

Regenerate

Generative AI
ChatGPT 3.5/4.0
Claude 2
Cohere Chat
Falcon
LLaMa2
Flan-T5

Core Innovation:
Ecosystems
Agents
LangChain
KGs
Tools
Problem Solving

Mixture of
Experts (MoE)?

Chain-of-Thought (CoT)

Few-shot In-Context Prompt

Text: Edward marks, an official with the ITAR explained their position...

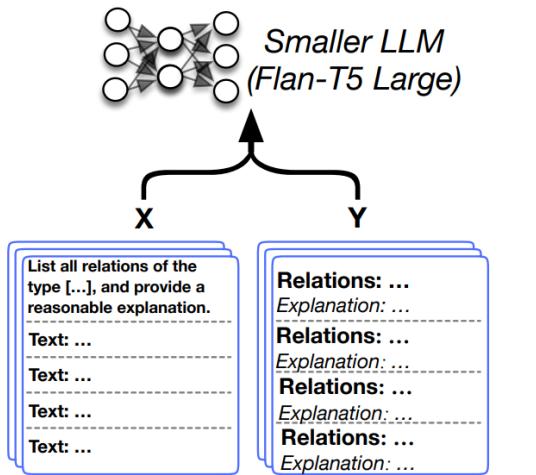
Triplets: [Edward marks:PER, work_for, ITAR:ORG]

Explanation: Edward Marks is an official with the ITAR, therefore it can be concluded that he works for ITAR.<s>

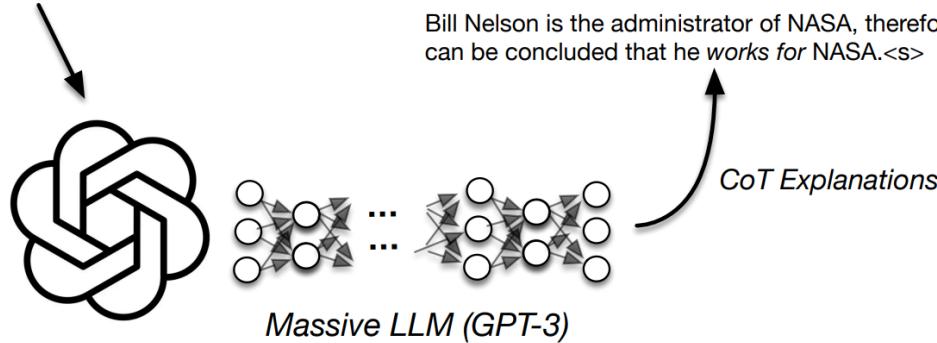
Text: NASA administrator Bill Nelson said in his historic speech that this mission...

Triplets: [Bill Nelson:PER, work_for, NASA:ORG]

Explanation:



Bill Nelson is the administrator of NASA, therefore it can be concluded that he works for NASA.<s>



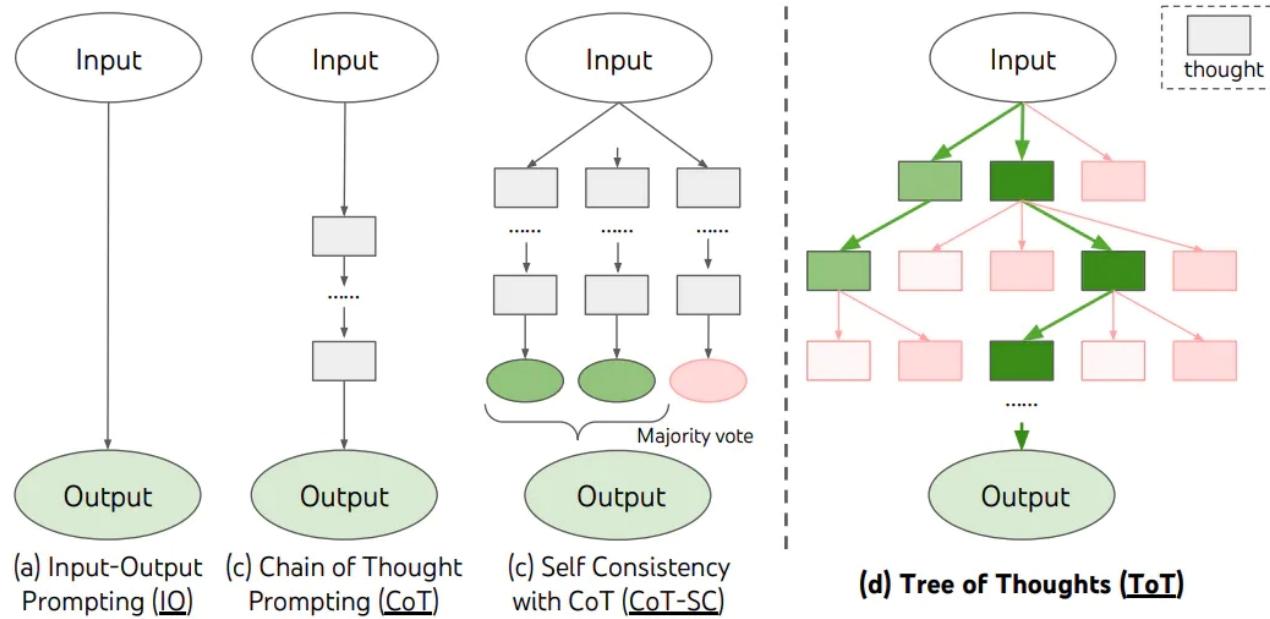
Relation Extraction with CoT
Explanation is All You Need!

Step-by-step reasoning

Augmented Text leads to better results!

Image Copyright © Revisiting Relation Extraction in the era of Large Language Models by Somin Wadhwa, Silvio Amir, Byron C. Wallace. ACL(1) 2023.

Tree of Thoughts (ToT)



ToT extends CoT
Multiple paths
towards an answer

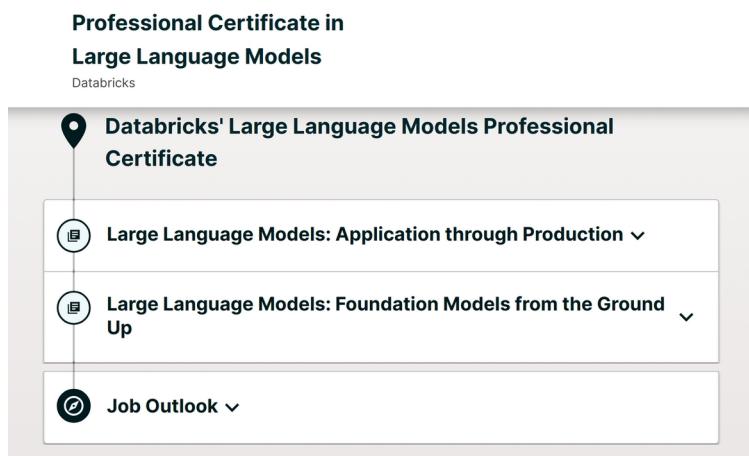
CoT-SC – Majority
voting mechanism

ToT – more similar
to the human
selection process

ToT allows for
parallel exploration
of ideas as
opposed to linear
exploration (CoT).

Image Copyright © Tree of Thoughts: Deliberate Problem Solving with Large Language Models (2023) by Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, Karthik Narasimhan. CoRR abs/2305.10601.

Best LLM Resources?



Repositories
HuggingFace
(models)
dair-ai (papers)

LangChain AI Handbook by James Briggs and Francisco Ingham

Image Copyright © Coursera Prompt Engineering with ChatGPT by Jules White (2023)

Image Copyright © EdX Databricks Professional Certificate in Large Language Models by Matei Zaharia et al (2023)

Evaluating Large Language Models

The screenshot shows a user interface for evaluating large language models. At the top, there's a navigation bar with a "New chat" button and a trash bin icon. Below it is a list of recent chats, all titled "List relations of types [empl...]" with a trash bin icon next to each. The main area shows a conversation with "Adrian". The message from Adrian is a detailed instruction about extracting entity relations from text. On the right side, there's a "System Context" section with dropdown menus and several sliders for tuning model parameters: Maximum Length (set to 8192), Temperature (set to 0.85), Top P (set to 0.99), Top K (set to 50), and Repetition Penalty (set to 1). A "Reset Parameters to default" button is also present. At the bottom, there's a text input field labeled "Send a message..." with a send arrow icon.

Single interface
nat.dev/chat

Includes
ChatGPT3.5/4
(with 32k cw)
Claude1/2
(with 100k cw)
Cohere Chat
MPT30B
Falcon40B
LLaMa2

Functionality
Playground
Compare
Chat
Metrics

SEM2023 Slides – Images Copyright © Framing Few-Shot Knowledge Graph Completion with Large Language Models by Adrian M.P. Brașoveanu, Lyndon J.B. Nixon, Albert Weichselbraun and Arno Scharl. NLP4KG at SEMANTICS 2023 (2023).

Evaluating Large Language Models

The prompts used for the general evaluation. X = [Person, Location, Organisation, Event, Date, Work, Other]. Y = [Per, Loc, Org, Event, Date, Work, Other]. Z = [employer, field of work, jurisdiction, member of, occupation, position, director, part of, objective, opposes, country, located in, operator, location, main subject, organizer, point of time] . [EXAMPLES] represents a list of 12 examples.

Task Type	Prompt
Relations	List relations among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Use the following texts as examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts
Explanations	List relations among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Provide a reasonable explanation for the relations extracted, as presented in the following examples. Provide a reasonable explanation for the relations extracted, as presented in the following examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts
Completions	List relations of types [Z] among the entities [X] in the given text and provide a reasonable explanation. Make sure to mark each entity with its abbreviated name [Y]. Relations should be expressed as triples that have the form [subject, predicate, object]. Provide a reasonable explanation for the relations extracted, as presented in the following examples. Provide a reasonable explanation for the relations extracted, as presented in the following examples. [EXAMPLES] Using the previous texts as examples, compute the relations for the following texts
Self-scoring	Please compare the previous answers with the following list and compute a score, considering that the newly provided list represents the gold standard.

Relations
Only Relations

Explanations
CoT

Completions
Restricted CoT

Self-Scoring
Truthfulness Proxy

Evaluating Large Language Models

Task Type	Example Input
Relations	Text, Relations
Explanations	Text, Relations, Explanations
KGC	Text, Relations (only missing), Explanations
Self-scoring	Any of the output from the previous tasks

Tools

GPT-3.5
GPT-4.0
Claude2
MPT-30B

Few-Shot

Input: 12-14 annotated texts
Output: 50 annotated texts

Source	Output
Text	Artificial intelligence pioneer Yoshua Bengio says regulation in Canada is on the right path, but advancing along it far too slowly.
Gold	(‘Yoshua Bengio:Per’, ‘field of work’, ‘A.I.’), (‘A.I. regulation’, ‘jurisdiction’, ‘Canada’), (‘Yoshua Bengio:Per’, ‘comments on’, ‘regulation in Canada:Loc’)
GPT-3.5	(‘Yoshua Bengio:Per’, ‘position’, ‘Artificial intelligence pioneer’), (‘Yoshua Bengio:Per’, ‘location’, ‘Canada:Loc’), (‘Canada:Loc’, ‘regulating’, ‘Artificial intelligence’)
GPT-4.0	(‘Yoshua Bengio:Per’, ‘expertise’, ‘Artificial Intelligence’), (‘Yoshua Bengio:Per’, ‘located in’, ‘Canada:Loc’)
Claude 2	(‘Yoshua Bengio:Per’, ‘pioneer in’, ‘artificial intelligence:Field’), (‘Yoshua Bengio:Per’, ‘comments on’, ‘regulation in Canada:Loc’)
MPT-30B	(‘Yoshua Bengio:Per’, ‘advocacy’, ‘AI regulation in Canada:Event’), (‘Yoshua Bengio’, ‘opinion’, ‘regulation is advancing too slowly’)

We want all the texts annotated in a large batch if possible

Evaluating Large Language Models

mpt-30b-chat

The F1 score is 1.0, indicating perfect agreement between the gold standard and the extracted information.

\$0.0416

mpt-30b-chat

The F1 score is 0.8.

\$0.0407

mpt-30b-chat

The F1 score is 0.44.

\$0.0371

gpt-3.5-turbo-16k

The F1 score is a measure of the accuracy of relation extraction, taking into account precision and recall. To calculate the F1 score, we need to compare the relations extracted by the system with the relations provided in the gold standard. In order to do that, we need the relations extracted by the system for the given texts. However, the system only provided the explanations for the relations, not the relations themselves.

Therefore, we are unable to compute the F1 score in this case.

\$0.0122

gpt-3.5-turbo-16k

The F1 scores for the provided relations compared to the gold standard are as follows:

- F1 score for the relations in the first text: 1.0
- F1 score for the relations in the second text: 0.5
- F1 score for the relations in the third text: 0.75
- F1 score for the relations in the fourth text: 0.5
- F1 score for the relations in the fifth text: 0.67

Overall average F1 score: 0.68

\$0.0121

mpt-30b-chat

The F1 score is 0.8.

\$0.0407

claudie-2

To compute the F1 score, I will compare the relations extracted in my previous responses to the relations in the provided gold standard.

For text 1, my response had 2 relations and the gold standard had 2 relations. Of these, 1 relation matched. So the precision is $1/2 = 0.5$, and the recall is $1/2 = 0.5$. Therefore, the F1 score is $2 * (0.5 * 0.5) / (0.5 + 0.5) = 0.5$

For text 2, my response had 4 relations and the gold standard had 5 relations. Of these, 3 relations matched. So the precision is $3/4 = 0.75$, and the recall is $3/5 = 0.6$. Therefore, the F1 score is $2 * (0.75 * 0.6) / (0.75 + 0.6) = 0.667$

For text 3, my response had 4 relations and the gold standard had 3 relations. Of these, 2 relations matched. So the precision is $2/4 = 0.5$, and the recall is $2/3 = 0.667$. Therefore, the F1 score is $2 * (0.5 * 0.667) / (0.5 + 0.667) = 0.571$

For text 4, my response had 3 relations and the gold standard had 4 relations. Of these, 2 relations matched. So the precision is $2/3 = 0.667$, and the recall is $2/4 = 0.5$. Therefore, the F1 score is $2 * (0.667 * 0.5) / (0.667 + 0.5) = 0.571$

For text 5, my response had 3 relations and the gold standard had 4 relations. Of these, 2 relations matched. So the precision is $2/3 = 0.667$, and the recall is $2/4 = 0.5$. Therefore, the F1 score is $2 * (0.667 * 0.5) / (0.667 + 0.5) = 0.571$

In summary, the F1 scores for the 5 texts are: Text 1: 0.5 Text 2: 0.667 Text 3: 0.571 Text 4: 0.571 Text 5: 0.571

The average F1 score over the 5 texts is $(0.5 + 0.667 + 0.571 + 0.571 + 0.571) / 5 = 0.576$

\$0.0645

Self-Scoring
Consecutive runs
Huge differences

**And the Winner
Is?**
**ChatGPT and
Claude2 have
similar
performance**

Acknowledgments

PROJECTS

GENTIO – FFG (GA No. 873992)

EPOCH – FFG (GA No. 867551)

DWBI Vienna - Vienna Science and Technology Fund (WWTF) [10.47379/ICT20096]

SDG-HUB – FFG (GA No. 892212)

THANK YOU!