

ChatGPT: Jack of All Trades, Master of None

Evaluation of ChatGPT's Capabilities on NLP Tasks

Izabela Telejko, Grzegorz Zbrzeźny, Tomasz Siudalski

November 20, 2024

ChatGPT: Jack of All Trades, Master of None

2022

Authors:

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek,
Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza,
Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra,
Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy,
Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik,
Stanisław Woźniak, Przemysław Kazienko

Department of Artificial Intelligence,
Wrocław University of Science and Technology,
Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland

Introduction

- ChatGPT, a conversational AI by OpenAI, is based on Transformer architecture.
- The study evaluates ChatGPT's performance on various NLP tasks compared to SOTA models.
- Explores personalization, task complexity, and prompt engineering.



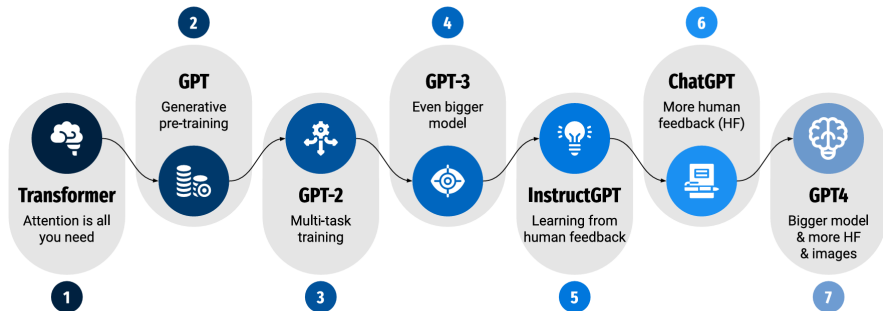
Research Objectives

The most important research questions in this paper were as follows:

- Is ChatGPT loss in performance compared to SOTA different for individual tasks of different kinds, and the same for GPT-4?
- Is there a difference in ChatGPT's ability to solve difficult and easy NLP analytical tasks?
- What is the impact of the context while processing multiple questions?
- Can we improve the model performance with prompt engineering?
- What are necessary post-processing activities that can improve the quality of ChatGPT output for analytical tasks?

Evolution of GPT models

- Progression from GPT to GPT-4 with increasing complexity and fine-tuning strategies.
- Introduction of Reinforcement Learning with Human Feedback (RLHF) for alignment with user needs.
- Applications across semantic and pragmatic NLP tasks.



Dataset Overview

- **25 tasks** tested, focusing on common NLP problems and analytical reasoning.
- Reasoning types:
 - ▶ **Binary Classification:** e.g., spam, sarcasm, aggression, grammatical correctness.
 - ▶ **Multiclass & Multi-label Classification:** e.g., sentiment analysis, emotion recognition.
 - ▶ **Personalized Reasoning:** Incorporating additional context for user-specific tasks.
 - ▶ **Semantic Annotation:** e.g., word sense disambiguation (WSD).
 - ▶ **Question Answering:** e.g. reading comprehension.
- Tasks categorized into:
 - ▶ **Semantic Tasks:**
 - ★ Focus on structured understanding and logical relationships.
 - ★ Examples: Natural Language Inference, Linguistic Acceptability.
 - ▶ **Pragmatic Tasks:**
 - ★ Require subjective interpretation and contextual reasoning.
 - ★ Examples: Emotion Recognition, Stance Detection.

Datasets Overview

- **Datasets:**

- ▶ Public datasets like Wikipedia, GoEmo, and SQuAD.
- ▶ New, unpublished dataset - ClarinEmo (Polish) to test ChatGPT's generalization.

- **Languages:**

- ▶ 92% of texts in English (23 tasks), 8% in Polish (2 tasks).

- **Dataset Availability Rating:**

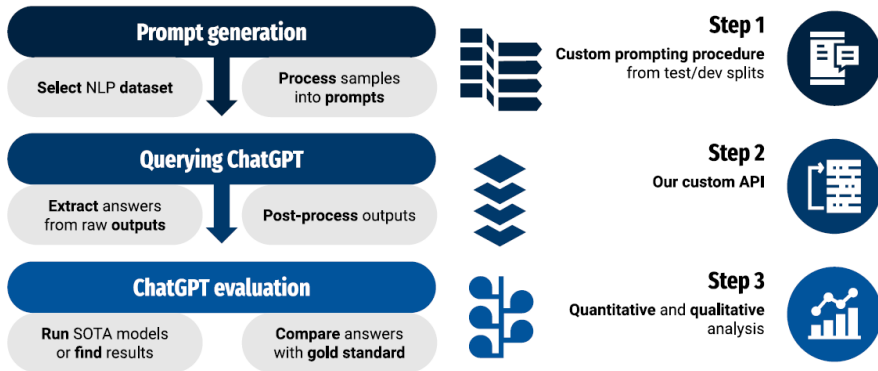
- ▶ *highly probable* (majority of datasets),
- ▶ *probable*,
- ▶ *rather no*,
- ▶ *impossible* (Unpublished datasets – ClarinEmo).

Dataset Overview

ID	Task Name	Task Category	NLP Problem	Dataset
1	Aggression	Pragmatic	Offensiveness detection	WikiDetox Aggr.
2	AggressionPer	Pragmatic	Offensiveness det.: personalized	WikiDetox Aggr.
3	CoLA	Semantic	Linguistic acceptability	CoLA
4	ColBERT	Pragmatic	Humor recognition	ColBERT
5	Sarcasm	Pragmatic	Humor recognition	Sarcasmania
6	Spam	Pragmatic	Spam detection	SMS Spam v.1
7	WordContext	Semantic	Word sense disambiguation	WiC
8	TextEntail	Semantic	Natural language inference	RTE
9	WNLI	Semantic	Natural language inference	WNLI
10	SQuAD	Semantic	Question answering	SQuAD v2
11	MathQA	Semantic	Question answering	GSM8K
12	*ClarinEmo	Pragmatic	Emotion recognition	ClarinEmo
13	*GoEmo	Pragmatic	Emotion recognition	GoEmotions
14	*GoEmoPer0	Pragmatic	Emotion rec.: personalized	GoEmotions
15	*GoEmoPer1	Pragmatic	Emotion rec.: personalized	GoEmotions
16	*GoEmoPer2	Pragmatic	Emotion rec.: personalized	GoEmotions
17	*GoEmoPer3	Pragmatic	Emotion rec.: personalized	GoEmotions
18	Unhealthy	Pragmatic	Offensiveness detection	Unhealthy Conv.
19	UnhealthyPer	Pragmatic	Offensiveness det.: personalized	Unhealthy Conv.
20	*PolEmo	Pragmatic	Sentiment analysis	PolEmo2
21	*TweetEmoji	Pragmatic	Emoji prediction	TweetEval
22	TweetSent	Pragmatic	Sentiment analysis	TweetEval
23	TweetStance	Semantic	Stance detection	TweetEval
24	ReAding	Semantic	Question answering	RACE
25	WSD	Semantic	Word sense disambiguation	Raganato

Table: NLP Tasks Evaluated in ChatGPT Study (emotion tasks are marked with an asterisk)

Methodology



Prompt Engineering

Key objectives:

- Solve a particular NLP problem
- Force ChatGPT to answer with a specified format

Chat 7. Task: Aggression. Case 402.

Prompt

Which one of the attributes: "aggressive", "non-aggressive" describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.

Text: *letters to an editor from an opponent are hardly reliable.*

ChatGPT answer

The text is "non-aggressive" as it is making a statement about the reliability of a source, rather than making any personal attacks or aggressive language.

Extracted answer

["non-aggressive"]

Expected answer

["non-aggressive"]

Chat 11. Task: CoLa. Case 4.

Prompt

Describe the grammatic acceptability of the given text. Choose your answer from provided list and map your answer with following unacceptable: 0, acceptable: 1 and return an integer as a result. Do not explain yourself.

Input data:

Text: *As you eat the most, you want the least.*

Possible acceptability: [unacceptable, acceptable]

ChatGPT answer

0

Expected answer

0

Querying ChatGPT

Raw text provided by ChatGPT differs from the final version after post-processing.

Issues with output:

- Answers may be returned as whole sentences instead of predefined lists.
- Typos (e.g., "negativ" to "negative") and incorrect values (e.g., out-of-list emotions) require manual corrections.
- Some values not in the requested list are mapped to "none".

Post-processing is required in relatively few cases:

- Most tasks (16) had less than 5% post-processing.
- Tasks like Aggression, SQuAD, Unhealthy, and ReAging exceeded 15%.

Evaluation of the results

- Models were launched equivalent to SOTA solutions when possible, given different setups (e.g., data splits).
- **F1 Macro** was the primary metric used for validation (if available):
 - ▶ Suitable for imbalanced data.
 - ▶ Average of harmonic means between precision and recall per label.

- Loss Metric:

$$\text{Loss} = 100\% \cdot \frac{\text{SOTA} - \text{ChatGPT}}{\text{SOTA}}$$

- Gain Metric:

$$\text{Gain} = 100\% \cdot \frac{\text{Per} - \text{NonPer}}{100\% - \text{NonPer}}$$

Quantitative Results

- ChatGPT's performance was 56.5% on average compared to SOTA's 73.7%.
- Losses were highest in emotion-related tasks.
- Semantic tasks generally showed better performance.

Quantitative Results

Task Name	ChatGPT [%]	SOTA [%]
Aggression	74.45	69.10
AggressionPer	81.03	72.57
CoLa	86.40	80.82
CoLBERT	98.50	86.47
Sarcasm	53.57	49.88
Spam	99.42	82.67
WordContext	74.00	64.58
TextEntail	92.10	88.09
WNLI	97.90	81.69
SQuAD	90.75	69.21
MathQA	83.20	71.40
*ClarinEmo	68.04	53.23
*GoEmo	52.75	25.55
*GoEmoPer0	54.50	23.74
*GoEmoPer1	66.10	19.00
*GoEmoPer2	66.10	20.34
*GoEmoPer3	66.10	23.41
Unhealthy	50.96	45.21
UnhealthyPer	70.92	54.02
*PolEmo	76.44	44.08
*TweetEmoji	32.20	18.19
TweetSent	72.07	63.32
TweetStance	67.42	56.44
ReAding	84.71	76.36
WSD	83.20	73.30

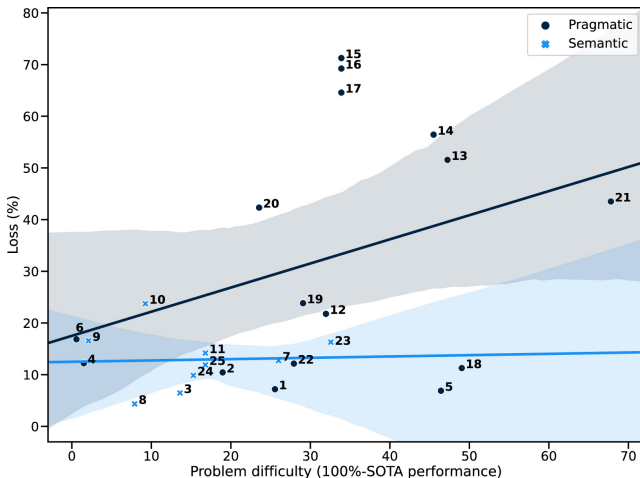
Table: Performance Comparison of ChatGPT vs. SOTA

Quantitative Results

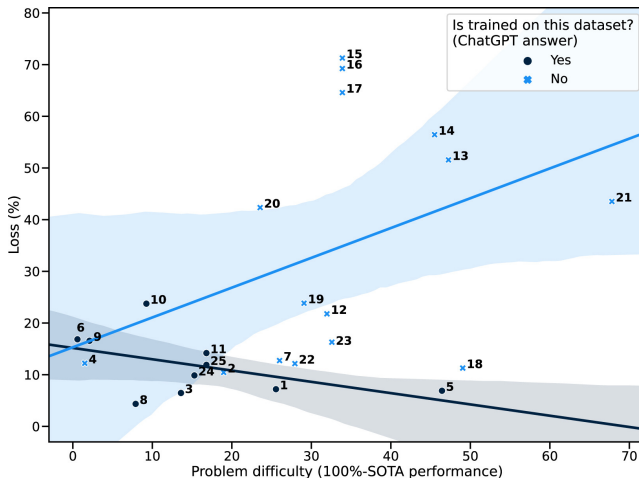
Category	ChatGPT Average	SOTA Average	Loss
All	56.51	73.71	25.50
Without Emotion	69.71	80.04	12.76
Only Emotion	28.44	60.28	52.59
Only Pragmatic	46.92	67.70	32.59
Only Semantic	73.54	84.41	12.90

Table: Average values for ChatGPT vs. SOTA with the difference.

Task difficulty vs. performance



Availability of the testing set



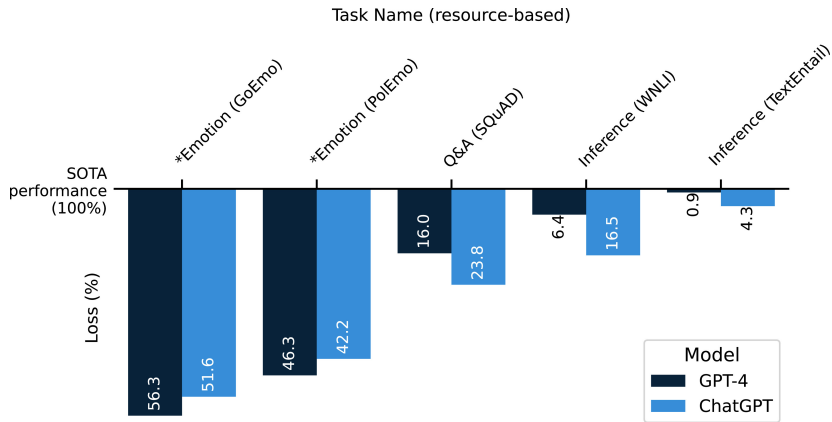
Impact of the context

Dataset	Accuracy [%]	F1 score [%]	Unanswerable detected cases	Unanswerable not detected cases
Original set	56.50	69.21	76 (25.33%)	224 (74.67%)
After week	55.40	68.72	64 (21.33%)	236 (78.67%)
New order	57.00	69.76	74 (24.67%)	226 (75.33%)
Separate conversations	53.60	67.23	60 (20.00%)	240 (80.00%)

Impact of prompt fine-tuning

ID	Task Name (resource-based)	Prompt pattern	Task category	Measure type	Initial ChatGPT (a) [%]	Pattern ChatGPT (b) [%]	Difference (b-a) [pp]	Pattern benefit [%]
8	TextEntail	78	Semantic	F1 Macro	88.1	77.6	-10.5	-11.9
8	TextEntail	79	Semantic	F1 Macro	88.1	81.2	-6.9	-7.8
8	TextEntail	80	Semantic	F1 Macro	88.1	77.6	-10.5	-11.9
8	TextEntail	81	Semantic	F1 Macro	88.1	75.1	-13.0	-14.8
9	WNLI	82	Semantic	Accuracy	81.7	74.2	-7.5	-9.2
9	WNLI	83	Semantic	Accuracy	81.7	77.5	-4.2	-5.1
9	WNLI	84	Semantic	Accuracy	81.7	76.1	-5.6	-6.9
13	*GoEmo	71	Pragmatic	F1 Macro	25.6	21.8	-3.8	-14.8
13	*GoEmo	72	Pragmatic	F1 Macro	25.6	26.4	0.8	3.1
13	*GoEmo	73	Pragmatic	F1 Macro	25.6	23.7	-1.9	-7.4
13	*GoEmo	74	Pragmatic	F1 Macro	25.6	24.6	-1.0	-3.9
20	*PolEmo	75	Pragmatic	F1 Macro	44.1	38.6	-5.5	-12.5
20	*PolEmo	76	Pragmatic	F1 Macro	44.1	57.8	13.7	31.1
20	*PolEmo	77	Pragmatic	F1 Macro	44.1	58.9	14.8	33.6
	All tasks			Average Std. dev.	59.4 ±27.8	56.5 ±24.0	-2.9 ±8.2	-2.7 ±15.6

Comparison with GPT-4



Explanatory analysis - problematic responses

- ChatGPT often misjudges neutral information by assigning positive or negative sentiment. It seems to have no regard for individuals, instead judging situations.
- The model flattens messages by evaluating content without considering the broader context or metatext, leading to inaccurate assessments.
- ChatGPT evaluates situations rather than participants, but the use of certain words tied to people leads to predetermined judgments.
- When explaining semantic decisions, the model paraphrases context rather than constructing precise sense definitions
- ChatGPT often interprets content through a lens of societal morality, sometimes missing alternative interpretations like humor until prompted.

Problematic responses - example 1

Chat 62. Task: Aggression. Case 899.

Prompt

Which one of the attributes: “aggressive”, “non-aggressive” describes a given text? Write your answer in the form of a Python list containing the appropriate attribute.

Text: *‘No worries, I’ll report him as well for calling people “trolls” that is a direct attack on people. One minute and I’ll link proof under this comment.’*

ChatGPT answer

```
[“aggressive”]
```

Prompt

Why did you label the text as aggressive?

ChatGPT answer

The text contains language that is confrontational and hostile, including statements like “I’ll report him” and “that is a direct attack on people”. These expressions suggest that the speaker is in a hostile state and is taking action against the person they are addressing, implying that they are being aggressive. Additionally, the use of quotes around “trolls” implies that the speaker is directly calling the person they are addressing a troll, which is further evidence of an aggressive tone.

Evaluation result

human - correct, ChatGPT - not correct

Problematic responses - example 2

Chat 66. Task: TweetSent. Case 2108.
Prompt
Describe the sentiment of the given text. Choose your answer from provided list and map your answer with following negative: 0, neutral: 1, positive: 2 and return an integer as a result. Do not explain yourself. Input data: Text: <i>Cristiano Ronaldo Shortlisted For FIFA's 2016 Best Men's Player Award</i> <i>#cristiano #ronaldo #shortlisted #player— Possible sentiment: negative, neutral, positive</i>
ChatGPT answer
2
Prompt
Why did you label the text as positive?
ChatGPT answer
I apologize for not explaining earlier. The text is labeled as positive because the sentiment expressed in the text is one of excitement or approval. The mention of Cristiano Ronaldo being shortlisted for the FIFA 2016 Best Men's Player Award is a recognition of his accomplishments and implies that he is being recognized for his talent and skills as a player. The use of hashtags like #cristiano, #ronaldo, #shortlisted, and #player further suggests that the sentiment is positive, as these tags are typically used to highlight and promote positive events or developments. In general, being shortlisted for an award is seen as a positive achievement, so the sentiment expressed in the text is positive.
Evaluation result
human - correct, ChatGP - not correct

Research Objectives - revisited

The most important research questions in this paper were as follows:

- Is ChatGPT loss in performance compared to SOTA different for individual tasks of different kinds and the same for GPT-4?
- Is there a difference in ChatGPT's ability to solve difficult and easy NLP analytical tasks?
- What is the impact of the context while processing multiple questions?
- Can we improve the model performance with prompt engineering?
- What are necessary post-processing activities that can improve the quality of ChatGPT output for analytical tasks?

Recommendations for Improvement

- Fine-tune on more diverse and balanced datasets.
- Enhance handling of subjectivity through advanced personalization techniques.
- Implement robust multilingual capabilities and ethical safeguards.

Conclusions

ChatGPT



General

Does **various** tasks



Inaccurate

Quality **drop** in tasks



Explainable

May **explain** if asked



Interactive

It **chats**



Release candidate

Beta testing stage



Creative

Multiple answers



Recent methods



Dedicated

Does **only one** task



Accurate

Solves its task **well**



Unexplained

Raw answer only



Passive

No talk



Stable release

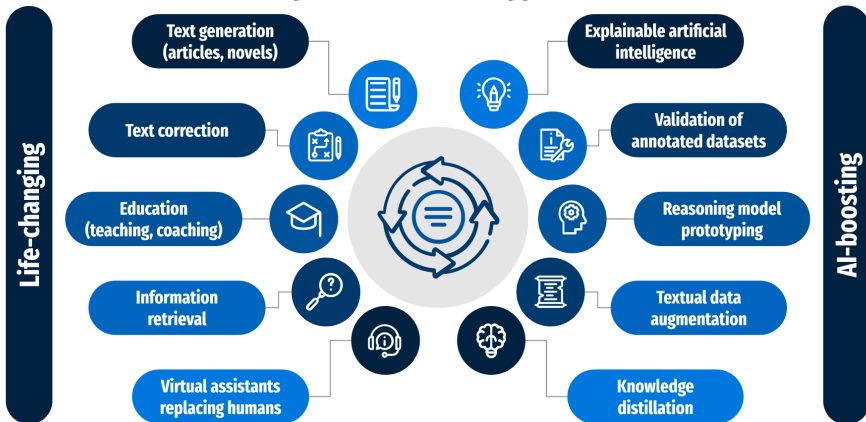
In production



Repetitive

Always **same** answer

Prospects for ChatGPT applications



Q&A

Prompts Used for Image Generation

A humanoid robot with multiple arms, each holding a tool or symbol for different tasks (e.g., a pen for writing, a stethoscope for medicine, a paintbrush for art, and a book for knowledge). The robot appears slightly overwhelmed, with an aura of versatility but no specialization. The background could show a digital workspace filled with faint, glowing icons of diverse tasks, suggesting broad but shallow expertise.

