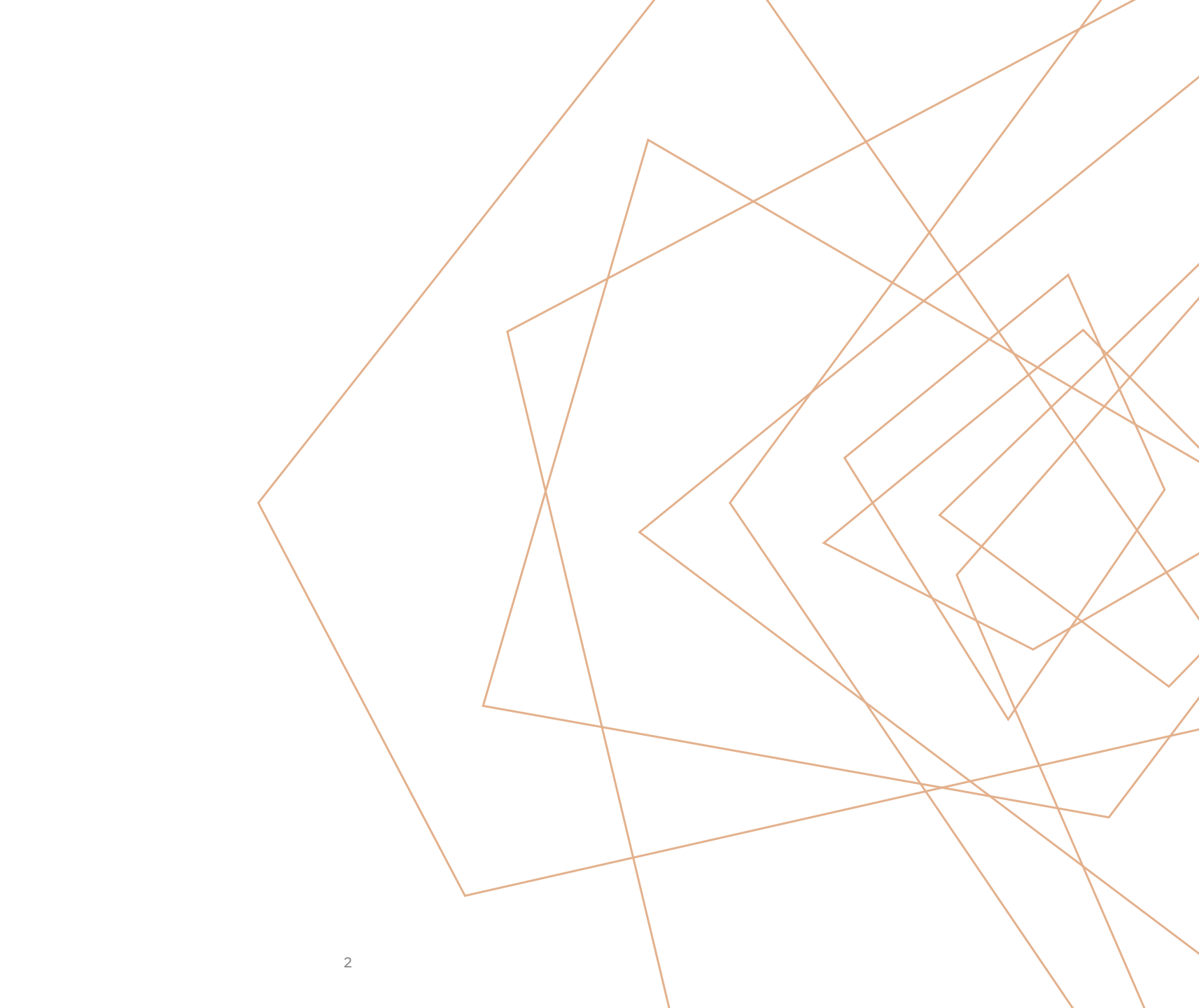
HOW GOOD ARE GPT MODELS AT MACHINE TRANSLATION

Alpaieva Yuliia

AGENDA

- Approaches
  - Rule-based
  - Statistical
  - Hybrid
  - Example-based
  - Neural
- Transformer Model
- Comprehensive Evaluation



APPROACHES

RULE-BASED

- Transfer-based
- Dictionary-based
- Interlingual

STATISTICAL

Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

HYBRID

Rules post-processed by statistics

Statistics guided by rules

EXAMPLE-BASED

Example-based machine translation is based on the idea of analogy. In this approach, the corpus that is used is one that contains texts that have already been translated. Given a sentence that is to be translated, sentences from this corpus are selected that contain similar sub-sentential components.

NEURAL

Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.

- Recurrent Neural Network
- Transformer Model

TRANSFORMER MODEL

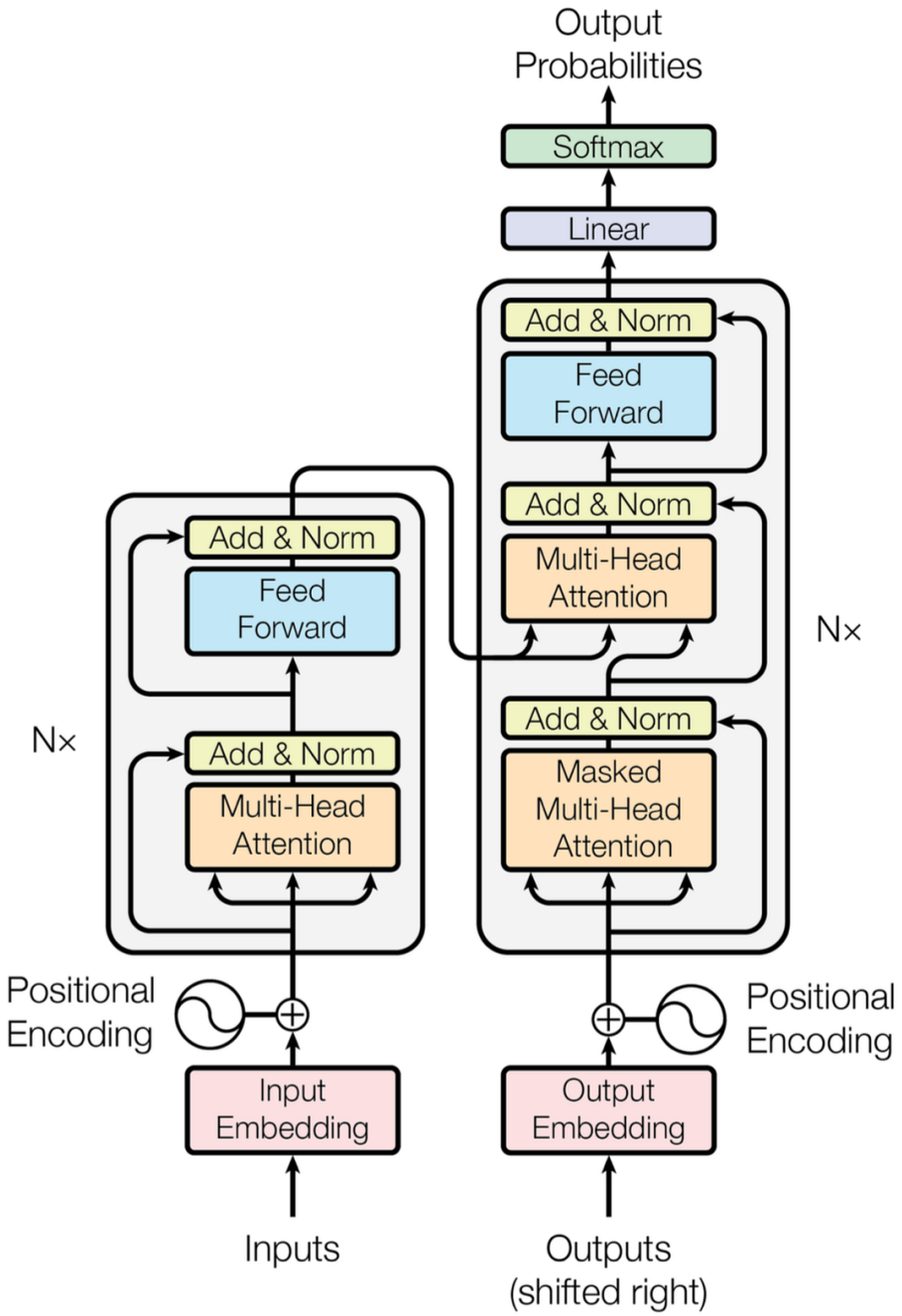


Figure 1: The Transformer - model architecture.

TRANSFORMER MODEL

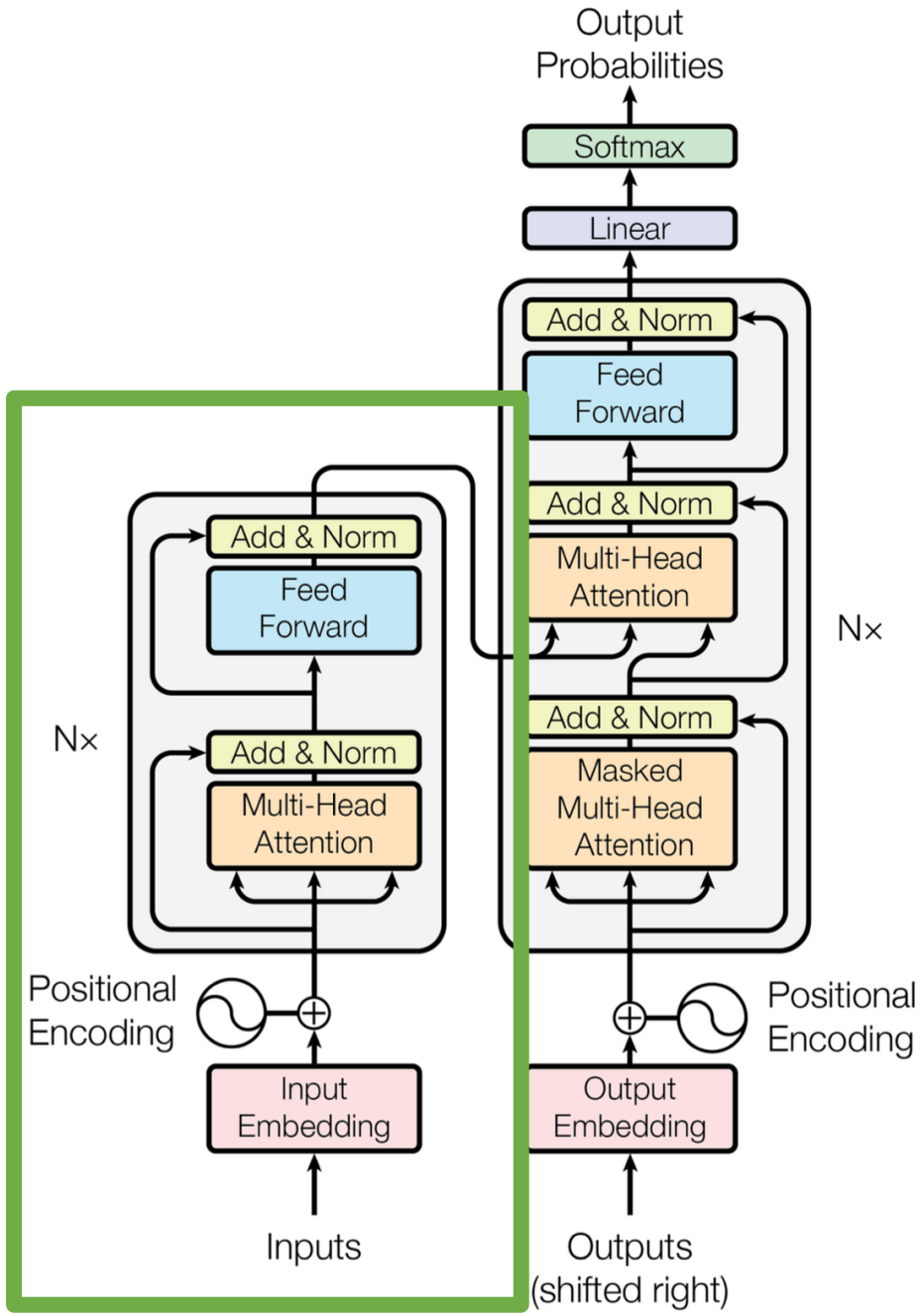


Figure 1: The Transformer - model architecture.

ENCODER

The encoder is composed of a stack of N = 6 identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, positionwise fully connected feed-forward network. We employ a residual connection [11] around each of the two sub-layers, followed by layer normalization [1]. That is, the output of each sub-layer is LayerNorm(x + Sublayer(x)), where Sublayer(x) is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension dmodel = 512.

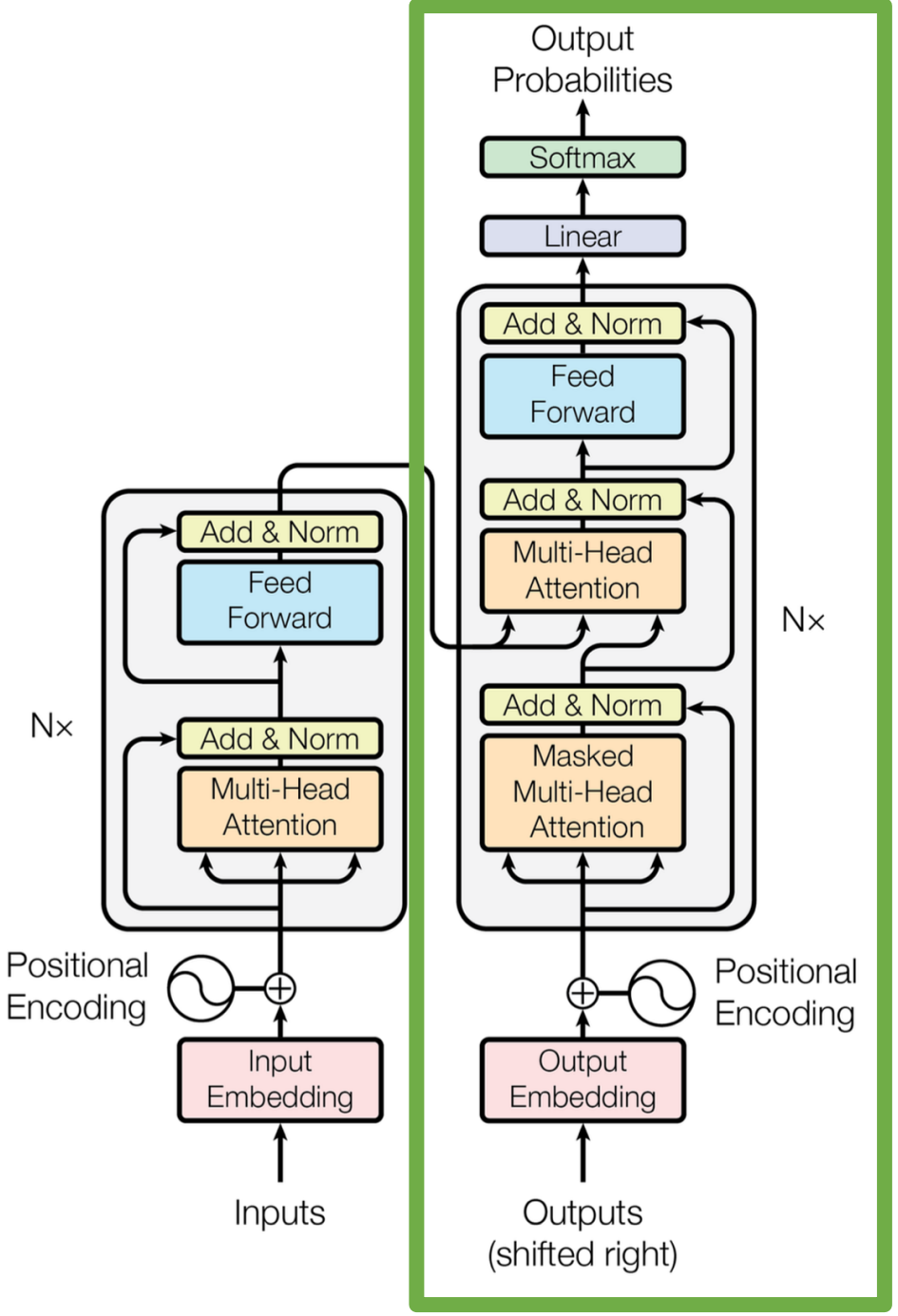


Figure 1: The Transformer - model architecture.

DECODER

The decoder is also composed of a stack of N = 6 identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. Similar to the encoder, we employ residual connections around each of the sub-layers, followed by layer normalization. We also modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position i can depend only on the known outputs at positions less than i.

MULTI-HEAD-SELF-ATTENTION

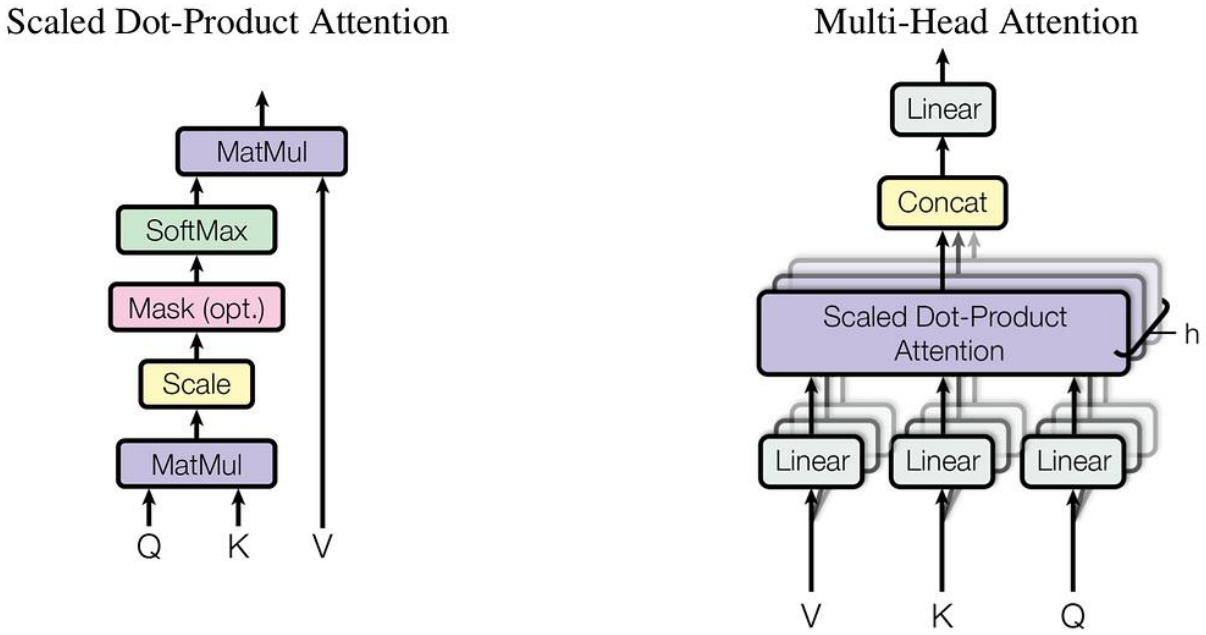


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

COMPREHENSIVE EVALUATION

How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, Hany Hassan Awadalla*
Microsoft

Abstract

Generative Pre-trained Transformer (GPT) models have shown remarkable capabilities for natural language generation, but their performance for machine translation has not been thoroughly investigated. In this paper, we present a comprehensive evaluation of GPT models for machine translation, covering various aspects such as quality of different GPT models in comparison with state-of-the-art research and commercial systems, effect of prompting strategies, robustness towards domain shifts and document-level translation. We experiment with eighteen different translation directions involving high and low resource languages, as well as non English-centric translations, and evaluate the

up new possibilities for building more effective translation systems (Brown et al., 2020; Chowdhery et al., 2022). Among these models, the latest Generative Pre-trained Transformer (GPT) models (Brown et al., 2020) have gained significant attention for their ability to generate coherent and context-aware text. We present a comprehensive evaluation of GPT models for machine translation, exploring their strengths and limitations, and providing insights for researchers and practitioners working in the area of machine translation.

GPT models and the conventional Neural Machine Translation (NMT) systems are both based on the transformer architecture (Vaswani et al., 2017), but they differ in several aspects. First, GPT models

Lang-Pair	Dataset	Number of sentences	WMT-Best System
CS-EN	WMT22	1448	Online-W
EN-CS		2037	Online-W
DE-EN	WMT22	1984	Lan-Bridge
EN-DE		2037	Online-B
IS-EN	WMT21	1000	Facebook-AI
EN-IS		1000	Facebook-AI
JA-EN	WMT22	2008	DLUT
EN-JA		2037	NT5
ZH-EN	WMT22	1875	JDExploreAcademy
EN-ZH		2037	Online-W
UK-EN	WMT22	2018	Lan-Bridge
EN-UK		2037	Online-B
RU-EN	WMT22	2016	JDExploreAcademy
EN-RU		2037	Online-W
HA-EN	WMT21	997	Facebook-AI
EN-HA		1000	Facebook-AI
FR-DE	WMT22	2006	Online-W
DE-FR		1984	Online-B

Table 1: Test datasets used in the evaluation and best systems used for comparison as reported in WMT by Kocmi et al. (2022b) and Barrault et al. (2021).

EXPERIMENTAL SETUP

- *text-davinci-002* - an InstructGPT model (Ouyang et al., 2022) which utilizes Reinforcement Learning with reward models trained based on human comparisons.
- *text-davinci-003* - an improved version of *text-davinci-002*.
- *ChatGPT* - a model that is similar to the previous two and optimized specifically for conversational purposes[8].

PROMPT SELECTION

B Few-shot Example Selection Data Pool

Language	# of sentences	
	Raw	Cleaned
CS-EN	193.5M	175.5M
EN-CS	167.9M	151.5M
DE-EN/EN-DE	295.8M	289.1M
IS-EN/EN-IS	4.4M	3.7M
JA-EN/EN-JA	33.9M	33.1M
ZH-EN	55.2M	50.4M
EN-ZH	35.5M	31.2M
UK-EN/EN-UK	50.6M	45.5M
RU-EN/EN-RU	75.0M	65.7M
HA-EN/EN-HA	7.2M	727K
FR-DE/DE-FR	17.7M	15.6M

Table 12: Size of the data pool for the few-shot example selections for each translation direction. Raw column shows the size of the original dataset which is from the WMT training dataset and Cleaned column shows high quality data after the cleaning from the original dataset.

PROMPT EXAMPLE

ZERO-SHOT

Prompt:

```
Translate from English to French:
The cat is in the kitchen. =>
```

Response:

```
Le chat est dans la cuisine.
```

FEW-SHOT

Prompt:

```
Translate from English to French:
I am in the kitchen. => Je suis dans la cuisine.
The cat is in the kitchen. =>
```

ZERO-SHOT TRANSLATION CAPABILITIES OF GPT MODELS

System	COMET-22	COMETkiwi	ChrF	BLEU	COMET-22	COMETkiwi	ChrF	BLEU
		DE-EN				EN-DE		
WMT-Best	**85.0**	**81.4**	**58.5**	**33.4**	**87.2**	**83.6**	**64.6**	**38.4**
text-davinci-002	73.2	73.1	46.1	23.3	82.0	79.0	56.0	28.6
text-davinci-003	84.8*	81.2*	56.8	30.9	85.6*	82.8*	60.2*	31.8*
ChatGPT	84.8*	81.1	58.3*	33.4*	84.2	81.0	59.6	30.9
		ZH-EN				EN-ZH		
WMT-Best	81.0	77.7	**61.1**	**33.5**	**86.7**	**82.0**	**41.1**	**44.8**
text-davinci-002	74.1	73.1	49.6	20.6	84.0	79.0	32.1	36.4
text-davinci-003	**81.6***	**78.9***	56.0*	25.0	85.8*	81.3*	34.6	38.3
ChatGPT	81.2	78.3	56.0	25.9*	84.4	78.7	36.0*	40.3*
		RU-EN				EN-RU		
WMT-Best	**86.0**	**81.7**	**68.9**	**45.1**	**89.5**	**84.4**	**58.3**	**32.4**
text-davinci-002	77.5	76	58.7	34.9	85.4	80.9	51.6	25.1
text-davinci-003	84.8*	81.1*	64.6	38.5	86.7*	82.2*	54.0*	27.5*
ChatGPT	84.8*	81.0	66.5*	41.0*	77.6	70.4	41.1	19.0
		FR-DE				DE-FR		
WMT-Best	**89.5**	**80.7**	**81.2**	**64.8**	**85.7**	79.5	**74.6**	**58.4**
text-davinci-002	66.6	67.9	45.8	25.9	64.2	67.6	44.6	24.5
text-davinci-003	84.6	77.9	65.7*	42.5*	78.5	76.1	58.9	35.6
ChatGPT	84.7*	78.5*	65.2	42.0	81.6*	**79.8***	60.7*	37.3*

Table 2: Zero-Shot evaluation results with three GPT models on 8 language pairs from WMT22 Testset. The best scores across different systems are marked bold. * denotes the best results among GPT systems.

GPT PERFORMANCE ON HIGH-RESOURCE LANGUAGES

System	COMET-22	COMETkiwi	ChrF	BLEU	COMET-22	COMETkiwi	ChrF	BLEU
		DE-EN				EN-DE		
WMT-Best	85.0	81.4	58.5	33.4	**87.2**	**83.6**	**64.6**	**38.4**
MS-Translator	84.7	81.0	**58.5**	**33.5**	86.8	83.4	64.2	37.3
GPT Zeroshot	84.8	81.2	56.8	30.9	85.6	82.8	60.2	31.8
GPT 1-Shot RR	84.9	81.3	56.1	30.4	86.1	83.0	60.7	31.9
GPT 1-Shot QR	84.9	81.3	56.7	31.1	85.8	82.8	60.7	32.4
GPT 5-Shot RR	85.2	81.5	56.5	31.2	86.5*	83.2 *	61.0	32.4
GPT 5-Shot QR	**85.4***	**81.5***	57.7	32.4	86.4	83.1	61.3*	33.2*
GPT 5-Shot QS	85.0	81.3	57.8*	32.5*	85.9	82.9	60.8	32.7
		CS-EN				EN-CS		
WMT-Best	**89.0**	**82.5**	**79.3**	**64.2**	**91.9**	**85.3**	**68.2**	**45.8**
MS-Translator	87.4	82.2	74.0	54.9	90.6	84.2	65.6	42.1
GPT Zeroshot	86.2	82.0	67.5	44.5	88.6	82.9	57.9	31.3
GPT 1-Shot RR	86.6	82.3	67.9	45.4	89.7*	84.0*	58.3	31.6
GPT 1-Shot QR	86.4	82.3	67.8	45.0	89.2	83.6	58.6	32.5
GPT 5-Shot RR	86.6	82.3	66.4	44.2	89.4	83.8	58.6	32.0
GPT 5-Shot QR	86.9*	**82.5***	69.2*	47.5*	89.0	83.3	59.0*	32.9*
		JA-EN				EN-JA		
WMT-Best	81.6	80.3	**49.8**	**24.8**	**89.3**	**85.8**	**36.8**	**27.6**
MS-Translator	81.5	80.1	49.6	24.5	88.0	85.3	34.9	25.1
GPT Zeroshot	81.5	80.7	47.7	21.1	87.8	84.8	31.2	21.2
GPT 1-Shot RR	81.7	80.7	46.8	20.2	88.3	85.1	31.8	22.0
GPT 1-Shot QR	81.6	80.8	48.3*	22.1	88.4*	85.3	32.2*	22.5*
GPT 5-Shot RR	**82.0***	**80.9***	48.2	22.4*	88.2	85.4*	31.7	21.4
GPT 5-Shot QR	81.8	80.8	47.2	21.0	88.2	85.3	31.1	21.6
		ZH-EN				EN-ZH		
WMT-Best	81.0	77.7	**61.1**	**33.5**	86.7	82.0	41.1	44.8
MS-Translator	80.4	77.6	57.7	27.9	86.1	81.4	**43.1**	**48.1**
GPT Zeroshot	**81.6***	**78.9***	56.0*	25.0*	85.8	81.3	34.6	38.3
GPT 1-Shot RR	80.9	78.2	55.2	24.2	86.7	81.8	38.7	42.8
GPT 1-Shot QR	81.2	78.8	55.3	24.2	86.1	81.5	35.5	38.8
GPT 5-Shot RR	81.1	78.8	55.0	24.4	87.0	82.0	37.1	41.3
GPT 5-Shot QR	81.1	78.7	54.7	23.8	**87.0***	**82.2***	39.8*	43.7*
GPT 5-Shot QS	81.0	78.5	55.5	24.6	86.2	81.5	38.3	41.8
		RU-EN				EN-RU		
WMT-Best	**86.0**	**81.7**	**68.9**	**45.1**	**89.5**	**84.4**	**58.3**	32.4
MS-Translator	85.2	80.7	68.3	43.9	87.4	82.9	58.1	**33.1**
GPT Zeroshot	84.8	81.1	64.6	38.5	86.7	82.2	54.0	27.5
GPT 1-Shot RR	84.1	80.6	63.3	37.9	86.4	81.9	54.3	28.1
GPT 1-Shot QR	84.9	81.2*	65.4*	40.1	86.9	82.4	53.8	27.5
GPT 5-Shot RR	84.9	81.2*	63.9	39.0	86.8	82.3	54.3	27.9
GPT 5-Shot QR	84.9	81.0	65.4*	40.0	87.0*	82.4*	54.4*	28.2*
GPT 5-Shot QS	85.0*	81.2*	65.3	40.2*	86.4	82.2	54.4*	28.0
		UK-EN				EN-UK		
WMT-Best	**86.0**	**81.5**	**67.3**	**44.6**	**88.8**	**83.4**	**59.3**	**32.5**
MS-Translator	83.5	79.7	65.3	42.4	86.1	81.9	56.1	28.2
GPT Zeroshot	83.5	80.1	59.8	34.8	83.7	79.5	49.6	21.1
GPT 1-Shot RR	83.5	80.3	60.3	35.6	84.7	80.2	50.1	21.2
GPT 1-Shot QR	83.8	80.3	61.4	37.5	85.1	80.5	50.5	21.9
GPT 5-Shot RR	83.6	80.3	58.8	34.4	85.4*	80.8*	50.9*	22.6*
GPT 5-Shot QR	83.9*	80.3*	62.1*	38.4*	85.4	80.6	50.6	22.1

GPT PERFORMANCE ON LOW-RESOURCE AND NON ENGLISH-CENTRIC LANGUAGES

System	COMET-22	COMETkiwi	ChrF	BLEU	COMET-22	COMETkiwi	ChrF	BLEU
		IS-EN				EN-IS		
WMT-Best	**87.0**	**81.4**	62.3	**41.7**	**86.8**	**81.8**	**59.6**	**33.3**
MS-Translator	85.9	80.3	**62.8**	40.5	84.3	80.2	56.8	28.7
GPT Zeroshot	82.1	78.7	55.6	31.9	76.3	74.0	43.5	15.9
GPT 1-Shot RR	84.1	80.2	57.8	34.7	77.0	74.6	43.7	15.3
GPT 1-Shot QR	83.5	79.7	56.7	33.3	77.4	75.1	44.5	16.2
GPT 5-Shot RR	84.4*	80.4*	58.1*	35.0*	77.9*	75.2*	45.1*	16.8*
GPT 5-Shot QR	84.2	80.2	58.0	35.2	76.0	74.1	44.1	16.3
		HA-EN				EN-HA		
WMT-Best	**80.0**	**74.5**	**48.7**	**21.0**	**79.8**	**61.5**	**51.1**	**20.1**
MS-Translator	73.3	68.5	43.4	16.2	72.5	57.2	38.4	10.3
GPT Zeroshot	76.1	73.1	45.5	17.3	73.3	58.6	38.4*	9.4*
GPT 1-Shot RR	75.7	72.7	45.7	17.3	74.0	59.0	38.4*	8.8
GPT 1-Shot QR	78.1	74.4	47.5*	19.1*	74.1*	59.7*	37.8	8.9
GPT 5-Shot RR	75.5	72.2	45.9	17.8	72.1	57.7	36.0	8.0
GPT 5-Shot QR	78.2*	74.5*	47.5*	18.9	72.6	58.5	36.9	8.5
		FR-DE				DE-FR		
WMT-Best	**89.5**	**80.7**	**81.2**	**64.8**	**85.7**	79.5	**74.6**	**58.4**
MS-Translator	85.4	78.9	67.5	45.3	82.7	79.0	65.0	42.0
GPT Zeroshot	84.6	77.9	65.7	42.5	78.5	76.1	58.9	35.6
GPT 1-Shot RR	86.1	79.6	65.1	41.0	83.1	80.5	60.3	36.9
GPT 1-Shot QR	86.4	80.0	67.0	43.9	83.2	80.8	61.2	38.1
GPT 5-Shot RR	86.6	80.0	65.2	41.6	83.6*	**80.9***	60.1	37.1
GPT 5-Shot QR	86.7*	80.2*	67.7*	44.8*	83.2	80.7	62.1*	39.3*

Table 4: Zero-Shot and Few-Shots evaluation results with GPT (text-davinci-003) on low resources and non-English centric translation directions from WMT Testsets. The best scores across different systems are marked bold. * denotes the best results among GPT systems.

DOCUMENT-LEVEL MT

```
Document:
[shot 1 source]

[shot 2 source]

[shot n source]

####

Translate each line in document into [target language].

Translated Document:

[shot 1 reference]

[shot 2 reference]

[shot n reference]

####

Document:
[sentence 1 from context window]

[sentence n from context window]

####

Translate each line in document into [target language].

Translated Document:
```

Figure 18: Prompt template for document translation.

EXPERIMENT 1

System	COMET-22	COMETkiwi	Doc-COMETkiwi	ChrF	BLEU	Doc-BLEU	GPT Requests
			DE-EN				
WMT-Best	85.0	**81.4**	79.9	**58.5**	33.4	**35.2**	–
MS-Translator	84.7	81.0	79.5	**58.5**	**33.5**	**35.2**	–
GPT Sent ZS	84.8	81.2	79.5	56.8	30.9	32.3	1984
GPT Doc ZS w=2	85.1	**81.4***	80.0	57.8	32.6	34.4	1055
GPT Doc ZS w=4	**85.2***	81.3	**80.2***	57.9	32.8	34.5	607
GPT Doc ZS w=8	85.1	81.2	80.2	57.9	33.0	34.7	401
GPT Doc ZS w=16	85.2	81.2	80.2	58.0*	33.1*	34.8*	310
GPT Doc ZS w=32	85.1	81.2	80.2	57.9	33.1	34.8	274
			EN-DE				
WMT-Best	**87.2**	**83.6**	**83.1**	**64.6**	**38.4**	**40**	–
MS-Translator	86.8	83.4	83	64.2	37.3	38.8	–
GPT Sent ZS	85.6	82.8	82.2	60.2	31.8	33.1	2037
GPT Doc ZS w=2	86.1	82.7	82.4	60.9	32.8	34.4	1058
GPT Doc ZS w=4	86.3	82.6	82.6	61.3	33.6	35.2	579
GPT Doc ZS w=8	86.4	82.6	82.6	60.9	33.4	35.2	349
GPT Doc ZS w=16	86.5*	82.6*	82.6*	61.3*	34.2*	36.1*	235
GPT Doc ZS w=32	86.4	82.6	82.7	61.3	34.1	36.1	187

Table 5: Evaluation results of document-level translation with GPT on DE<>DE WMT22 testset. The table shows the effect of increasing context length w in document to document translation with a zero-shot setting.

EXPERIMENT 2

System	COMET-22	COMETkiwi	Doc-COMETkiwi	ChrF	BLEU	Doc-BLEU
			DE-EN			
WMT-Best	85.0	81.4	79.9	**58.5**	33.4	**35.2**
MS-Translator	84.7	81.0	79.5	**58.5**	**33.5**	**35.2**
GPT-Sent-QR	**85.4***	**81.5***	80.2	57.7	32.4	34.0
GPT-Sent-DR	84.9	81.4	80.0	55.3	29.6	31.3
GPT-Doc-QR	85.2	81.2	80.2	58.1	33.3	35.0
GPT-Doc-DR	85.2	81.3	**80.5***	57.6	32.7	34.3
GPT-Doc-DF	85.1	81.3	80.4	57.3	32.6	34.1
GPT-Doc-DH	85.3	81.2	80.3	58.2*	**33.5***	35.1*
			EN-DE			
WMT-Best	**87.2**	**83.6**	**83.1**	**64.6**	**38.4**	**40.0**
MS-Translator	86.8	83.4	83.0	64.2	37.3	38.2
GPT-Sent-QR	86.4	83.1	82.7	61.3	33.2	34.8
GPT-Sent-DR	86.7	83.5*	83.0*	61.2	32.8	34.3
GPT-Doc-QR	86.6	83.0	82.7	61.6	34.0	35.7
GPT-Doc-DR	86.9	83.0	82.9	61.9	34.4	36.1
GPT-Doc-DF	87.0*	83.1	82.9	62.0*	34.5*	36.2*
GPT-Doc-DH	86.6	82.9	82.8	61.7	33.9	35.7

Table 6: Effect of shot selection for document-level translation on WMT22 DE<>EN testset.

CONCLUSION