In-Context Learning – Human-Computer Interface

Nasy

Mar 24, 2023

Outline

Example – Transfrom a table to a review sentiment template

Table:

Text	Label
Delicious food	1
The food is awful	0
Terrible dishes	0

Template: Review: [text] Sentiment: [label]

```
[ME]:
Review: Delicious food      Sentiment: Positive
Review: The food is awful   Sentiment: Negative
Review: Terrible dishes     Sentiment: Negative
Review: Good meal           Sentiment:
[AI]: Positive
```

Example – Emoji

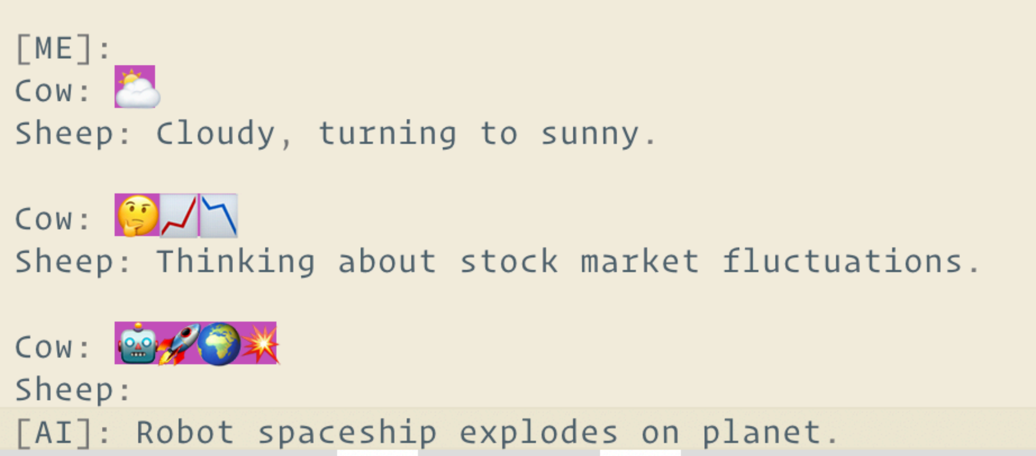

Figure: Emoji English translation.

Example – Graph

See org file

Section

Definition

Input A demonstration set $C = \{< x_i, y_i > | i \in [1, n]\}$, and the next x_{n+1}

Output y_{n+1}

$$P(y_j|x) = f_M(y_i, x_i, C), i \in [1, n]$$

$$y_{n+1} = \arg\max_{y_j} P(y_j|x)$$

Difference between ICL and others

	Model	Task description	Examples
Fine-tune	Update	Yes	Yes
Few-shot	Original	Yes	Yes
Zero-shot (Instruct)	Original	Yes	No
In-Context Learning	Original	No	Yes

```
[ME]:
### Transform sentence to triplets.  // Task description (Instruct)

### E.g.
Alice is Bob's roommate. Make her node green.
[
["Alice", "is roommate", "Bob"],
["Alice", "#00FF00"]
]

### Input
Over-fitting and over-smoothing are two main obstacles of developing deep Graph Convolutional Networks (GCNs) for node classification.

[AI]: [
["over-fitting", "are obstacles of", "developing deep Graph Convolutional Networks for node classification"],
["over-smoothing", "are obstacles of", "developing deep Graph Convolutional Networks for node classification"]
]
```

Section

Taxonomy of in-context learning.

ICL Paper list: https://github.com/dqxiu/ICL_PaperList

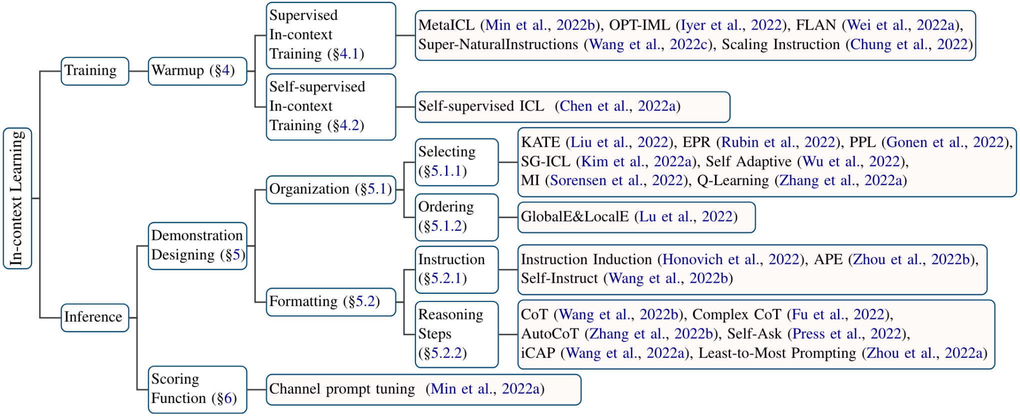

Figure: Taxonomy of in-context learning[1].

[1] Qingxiu Dong et al. *A Survey on In-context Learning*. Feb. 7, 2023. arXiv: arXiv:2301.00234. URL: http://arxiv.org/abs/2301.00234. preprint.

Large Language Models Are Human-Level Prompt Engineers

If we assume that In-Context Learning is the use of examples to concretely express task commands, and Instruct is a more abstract task description that is better suited to human habits, then a very natural question is: what is the connection between them?

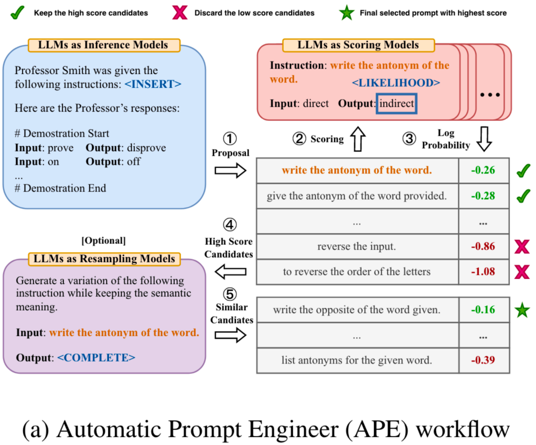

Figure: Automatic Prompt Engineer (APE) workflow[2]. Demo: https://sites.google.com/view/automatic-prompt-engineer

[2] Yongchao Zhou et al. "Large Language Models Are Human-Level Prompt Engineers". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=92gvk82DE-.

Section

Training

During training, split the raw text into ICL examples. As the data size increases, the effect gradually flattens out, but increasing task diversity can further enhance performance.

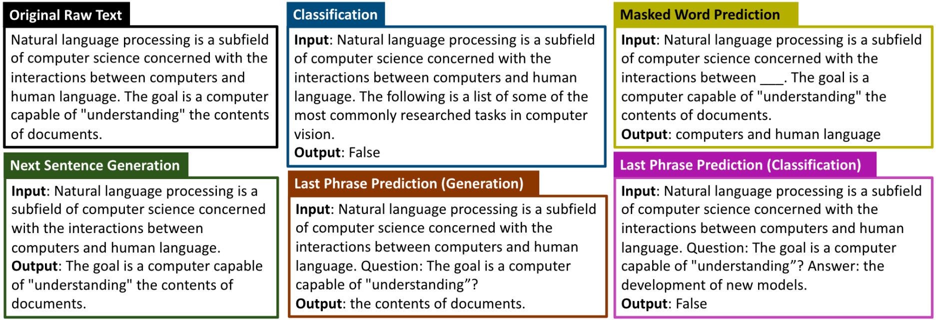

Figure: Each example is an input-output pair constructed from the raw text.[3]

[3] Mingda Chen et al. "Improving In-Context Few-Shot Learning via Self-Supervised Training". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3558–3573. DOI: 10.18653/v1/2022.naacl-main.260. URL: https://aclanthology.org/2022.naacl-main.260.

Section

Inference – Demonstration Designing

- Demostration Organization
  - Selection
  - Order
- Demostration Formatting
  - Instruction
  - Reasoning Steps

Selection and Order

Target Which examples are good examples for ICL?

- Unsupervised Method
  - L2 distance
  - Cosine similarity
  - ...
  - LLM
- Supervised Method
  - Human Feedback Reinforcement Learning
  - ...

- Put the most similar examples last.
- Order by entropy metrics[a]

[a] Yao Lu et al. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556.

Formatting

- Instruction
  - Finetuned Language Models are Zero-Shot Learners (FLAN)[4]
- Reasoning Steps
  - CoT (Chain of Thought)[5]
  - Self-consistency[6]
  - Least-to-most prompting[7]
  - Lets think step by step (Zero-Shot-CoT)[8]

---

[4] Jason Wei et al. "Finetuned Language Models Are Zero-Shot Learners". In: International Conference on Learning Representations. Jan. 28, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[5] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: Advances in Neural Information Processing Systems. Oct. 31, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

[6] Xuezhi Wang et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

[7] Denny Zhou et al. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=WZH7099tgfM.

[8] Takeshi Kojima et al. *Large Language Models Are Zero-Shot Reasoners*. Jan. 29, 2023. arXiv: arXiv:2205.11916. URL: http://arxiv.org/abs/2205.11916. preprint.

FLAN[9]

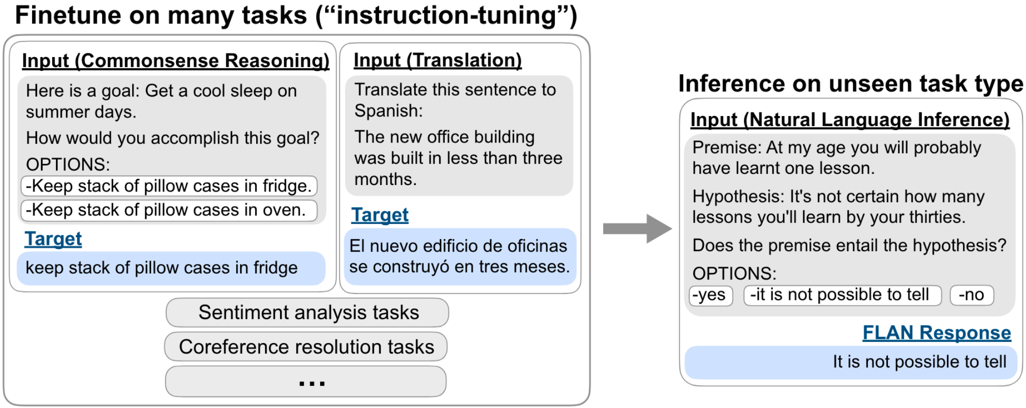

Figure: Instruction tuning finetunes a pretrained language model on a mixture of tasks phrased as instructions.

[9] Jason Wei et al. "Finetuned Language Models Are Zero-Shot Learners". In: International Conference on Learning Representations. Jan. 28, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

CoT[10].

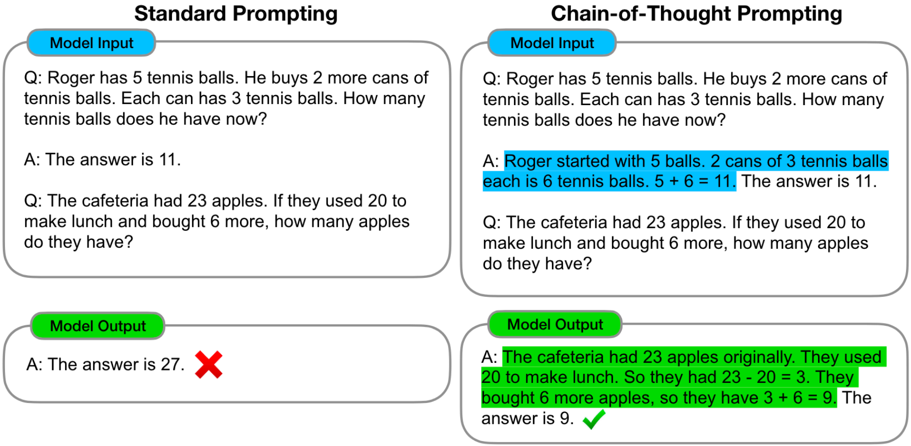

Figure: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

[10] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: Advances in Neural Information Processing Systems. Oct. 31, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

Self Consistency[11]

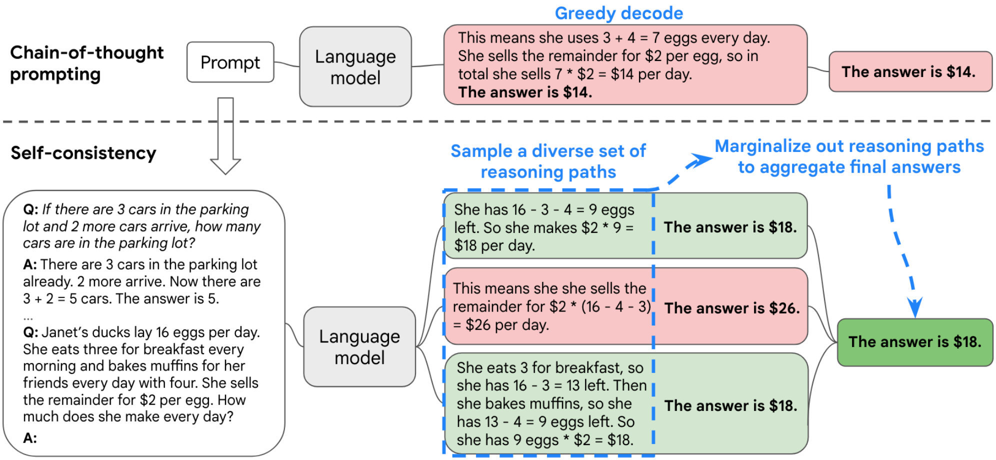

Figure: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the "greedy decode" in CoT prompting by sampling from the language model's decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

[11] Xuezhi Wang et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

Least-to-most prompting[12]

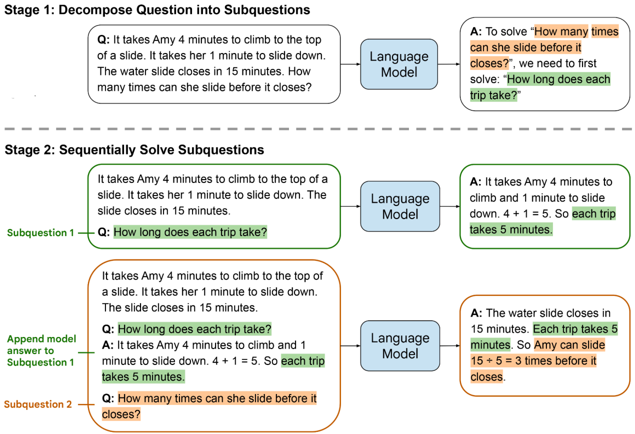

Figure: Least-to-most prompting solving a math word problem in two stages: (1) query the language model to decompose the problem into subproblems; (2) query the language model to sequentially solve the subproblems. The answer to the second subproblem is built on the answer to the first subproblem. The demonstration examples for each stage's prompt are omitted in this illustration.

[12] Denny Zhou et al. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=WZH7099tgfM.

Lets Think Step by Step (Zero-Shot-CoT)[13]

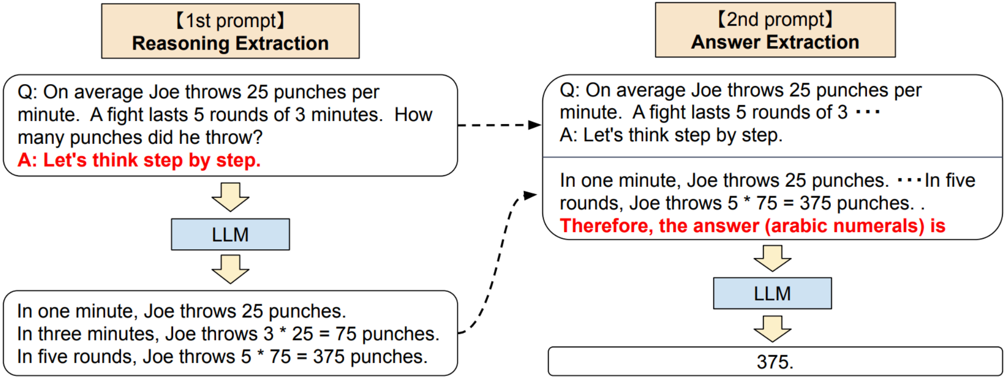

Figure: Lets think step by step.

[13] Takeshi Kojima et al. *Large Language Models Are Zero-Shot Reasoners*. Jan. 29, 2023. arXiv: arXiv:2205.11916. URL: http://arxiv.org/abs/2205.11916. preprint.

Results

	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
zero-shot	74.6/**78.7**	**72.2/77.0**	17.7/22.7	10.4/12.5	22.4/22.4	58.8/58.7
zero-shot-cot	**78.0/78.7**	69.6/74.7	**78.7/79.3**	**40.7/40.5**	**33.5/31.9**	**62.1/63.7**
	Common Sense		Other Reasoning Tasks		Symbolic Reasoning	
	Common SenseQA	Strategy QA	Date Understand	Shuffled Objects	Last Letter (4 words)	Coin Flip (4 times)
zero-shot	**68.8/72.6**	12.7/**54.3**	49.3/33.6	31.3/29.7	0.2/-	12.8/53.8
zero-shot-cot	64.6/64.0	**54.8**/52.3	**67.5/61.8**	**52.4/52.9**	**57.6/-**	**91.4/87.8**

Figure: Accuracy comparison of Zero-shot-CoT with Zero-shot on each tasks. The values on the left side of each task are the results of using answer extraction prompts depending on answer format. The values on the right side are the result of additional experiment where standard answer prompt "The answer is" is used for answer extraction

Inference – Scoring Function

Channel Model[14]

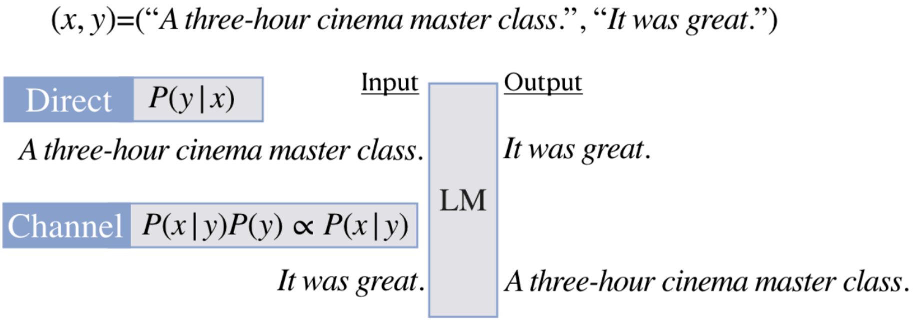

Figure: Channel model.

[14] Sewon Min et al. "Noisy Channel Language Model Prompting for Few-Shot Text Classification". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5316–5330. DOI: 10.18653/v1/2022.acl-long.365. URL: https://aclanthology.org/2022.acl-long.365.

No

- In *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?*[15], the authors show that ICL is not learning anything.
- They use a random label $y_r \in Y$ to replace the true label y_i for x_i, and the model still works.
- What really affects is the distribution of $< x_i, y_i >$.

```
[ME]:
Review: Delicious food       Sentiment: Negative
Review: The food is awful    Sentiment: Positive
Review: Terrible dishes      Sentiment: Negative
Review: Good meal            Sentiment:

[AI]: Positive
```

[15] Sewon Min et al. *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* Oct. 20, 2022. DOI: 10.48550/arXiv.2202.12837. arXiv: arXiv:2202.12837. URL: http://arxiv.org/abs/2202.12837. preprint.

Yes

- Ekin Akyurek[16]
  - Transformer-based in-context learners implement standard learning algorithms implicitly, by encoding smaller models in their activations, and updating these implicit models as new examples appear in the context.
- Damai Dai[17]
  - Language models is meta-optimizers and understands ICL is a kind of implicit finetuning.

[16] Ekin Akyürek et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=0g0X4H8yN4I.

[17] Damai Dai et al. *Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers*. Dec. 21, 2022. DOI: 10.48550/arXiv.2212.10559. arXiv: arXiv:2212.10559. URL: http://arxiv.org/abs/2212.10559. preprint.

Conclusion

- In-context learnig (ICL) difinition
- Taxonomy of ICL
- Relation between ICL and instruct
- ICL in Training
- ICL in Inference
    - Demonstration Designing
        - Selection and Order
        - Formatting
    - Scoring Function

Reference I

[1] Ekin Akyürek et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=0g0X4H8yN4I.

[2] Mingda Chen et al. "Improving In-Context Few-Shot Learning via Self-Supervised Training". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3558–3573. DOI: 10.18653/v1/2022.naacl-main.260. URL: https://aclanthology.org/2022.naacl-main.260.

[3] Damai Dai et al. *Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers*. Dec. 21, 2022. DOI: 10.48550/arXiv.2212.10559. arXiv: arXiv:2212.10559. URL: http://arxiv.org/abs/2212.10559. preprint.

Reference II

[4] Qingxiu Dong et al. *A Survey on In-context Learning*. Feb. 7, 2023. arXiv: arXiv:2301.00234. URL: http://arxiv.org/abs/2301.00234. preprint.

[5] Takeshi Kojima et al. *Large Language Models Are Zero-Shot Reasoners*. Jan. 29, 2023. arXiv: arXiv:2205.11916. URL: http://arxiv.org/abs/2205.11916. preprint.

[6] Yao Lu et al. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL: https://aclanthology.org/2022.acl-long.556.

Reference III

[7] Sewon Min et al. “Noisy Channel Language Model Prompting for Few-Shot Text Classification”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5316–5330. DOI: 10.18653/v1/2022.acl-long.365. URL: https://aclanthology.org/2022.acl-long.365.

[8] Sewon Min et al. *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* Oct. 20, 2022. DOI: 10.48550/arXiv.2202.12837. arXiv: arXiv:2202.12837. URL: http://arxiv.org/abs/2202.12837. preprint.

[9] Xuezhi Wang et al. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

Reference IV

[10] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: Advances in Neural Information Processing Systems. Oct. 31, 2022. URL: https://openreview.net/forum?id=_VjQlMeSB_J.

[11] Jason Wei et al. "Finetuned Language Models Are Zero-Shot Learners". In: International Conference on Learning Representations. Jan. 28, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[12] Denny Zhou et al. "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=WZH7099tgfM.

[13] Yongchao Zhou et al. "Large Language Models Are Human-Level Prompt Engineers". In: The Eleventh International Conference on Learning Representations. Feb. 1, 2023. URL: https://openreview.net/forum?id=92gvk82DE-.

Example I

Over-fitting and over-smoothing are two main obstacles of developing deep Graph Convolutional Networks (GCNs) for node classification. In particular, over-fitting weakens the generalization ability on small dataset, while over-smoothing impedes model training by isolating output representations from the input features with the increase in network depth. This paper proposes DropEdge, a novel and flexible technique to alleviate both issues. At its core, DropEdge randomly removes a certain number of edges from the input graph at each training epoch, acting like a data augmenter and also a message passing reducer. Furthermore, we theoretically demonstrate that DropEdge either reduces the convergence speed of over-smoothing or relieves the information loss caused by it. More importantly, our DropEdge is a general skill that can be equipped with many other backbone models (e.g. GCN, ResGCN, GraphSAGE, and JKNet) for enhanced performance. Extensive experiments on several benchmarks verify that DropEdge consistently improves the performance on a variety of both shallow and deep GCNs. The effect of DropEdge on preventing over-smoothing is empirically visualized and validated as well. Codes are released on https://github.com/DropEdge/DropEdge

Example II

```
[ME]:
Alice is Bob’s roommate. Make her node green.
[
["Alice", "is roommate", "Bob"],
["Alice", "#00FF00"]
]

In this paper, we question if self-supervised learning provides new propertie
[
["Paper", "focus", "self-supervised learning properties of ViT"],
["ViT", "compared to", "convnets"],
["Self-supervised ViT", "observation", "contains explicit information about se
["Self-supervised ViT", "observation", "excellent k-NN classifiers"],
["Self-supervised ViT", "performance", "78.3% top-1 on ImageNet with a small V
["Momentum encoder", "importance", "self-supervised learning"],
["Multi-crop training", "importance", "self-supervised learning"],
```

Example III

```
["Small patches", "importance", "self-supervised learning with ViTs"],
["DINO", "implemented", "self-supervised method"],
["DINO", "interpretation", "self-distillation with no labels"],
["DINO", "synergy", "ViTs"],
["DINO and ViT-Base", "performance", "80.1% top-1 on ImageNet in linear evalua
]

Over-fitting and over-smoothing are two main obstacles of developing deep Gra
[
["Over-fitting", "obstacle", "GCNs"],
["Over-smoothing", "obstacle", "GCNs"],
["GCNs", "purpose", "node classification"],
["DropEdge", "proposed for", "alleviating over-fitting and over-smoothing"],
["DropEdge", "function", "randomly removes edges"],
["DropEdge", "acts as", "data augmenter"],
["DropEdge", "acts as", "message passing reducer"],
```

Example IV

```
["DropEdge", "effect", "reduces convergence speed of over-smoothing or relieve
["DropEdge", "applicable to", "GCN"],
["DropEdge", "applicable to", "ResGCN"],
["DropEdge", "applicable to", "GraphSAGE"],
["DropEdge", "applicable to", "JKNet"],
["Experiments", "validate", "DropEdge"],
["DropEdge", "prevents", "over-smoothing"],
["Codes", "available at", "https://github.com/DropEdge/DropEdge"]
]

Pretrained large language models (LLMs) are widely used in many sub-fields of

[AI]: [
["Pretrained large language models", "widely used in", "many sub-fields of NLI
["LLMs", "known for", "excellent few-shot learning"],
["Chain of thought (CoT)", "a recent technique for", "eliciting complex multi-
```

Example V

```
["CoT prompting", "achieved", "state-of-the-art performances in arithmetics ar
["CoT prompting", "successes attributed to", "LLMs ability for few-shot learni
["Zero-shot-CoT", "outperforms", "zero-shot LLM performances on diverse benchm
["Zero-shot-CoT", "uses", "same single prompt template"],
["Zero-shot-CoT", "increases accuracy on MultiArith from 17.7% to 78.7% and GS
["Zero-shot-CoT", "similar magnitudes of improvements", "with another off-the-
["Single prompt", "versatile across", "diverse reasoning tasks"],
["Zero-shot capabilities", "hidden inside LLMs", "untapped and understudied fu
["Zero-shot capabilities", "suggests high-level, multi-task broad cognitive ca
["Importance", "carefully exploring and analyzing the enormous zero-shot knowl
["Hope", "work serves as", "minimal strongest zero-shot baseline for the chall
["Hope", "work highlights", "importance of exploring and analyzing the enormou
]

[ME]:
```