International Journal of User-System Interaction 15(3) 2022, 69-81

Distilling ChatGPT for Short Story Generation

Neață Adrian[1], Ștefan Trăușan-Matu[1,2,3]

[1] Politehnica University of Bucharest
313 Splaiul Independenței, Bucharest, Romania
[2] Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy
13 Calea 13 Septembrie, Bucharest, Romania
[3] Academy of Romanian Scientists
54 Splaiul Independenței, Bucharest, Romania
E-mail: adrian.neata98@gmail.com, stefan.trausan@upb.ro

Abstract. The paper presents an analysis of the knowledge distillation process, encompassing the preparation of a ChatGPT-generated story dataset, the finetuning of a small GPT-2 model, and the evaluation of short story quality and coherence. Through extensive experiments and evaluations, state-of-the-art results are obtained when compared to models of similar sizes (between 100M and 400M parameters) and we believe our approach can boost the performance of NLG models in general. Furthermore, an impromptu examination of the texts generated by ChatGPT reveals biases towards certain words and phrases.

Keywords: natural language generation; knowledge distillation; story generation; ChatGPT; GPT2;

1. Introduction

Natural language generation (NLG) has been an important domain since the early years of Artificial Intelligence. Starting off from the use of templates for creating sentences from data in the 60's, we reached a point where we use advanced machine learning models to write stories, but those texts are not at the same level of quality as the stories written by humans.

While a lot of work has been poured into developing better neural network architectures for NLG, another important aspect to consider is the training dataset and as it will be shown in the next sections, using ChatGPT to create a new dataset can improve the overall performance of the models but it will also come with a few drawbacks.

It is important to note that a model capable of consistently producing top-notch short stories or an algorithm or optimization that can enhance a generator's output would most likely be easily adapted to the broader field of

NLG as many of the obstacles of short story generation apply to NLG (coherence and cohesion, style and tone, entity recognition and reference resolution).

GPT-3 (Generative Pre-trained Transformer 3) (Brown et al., 2020) is the engine behind the public version of ChatGPT and one of the most remarkable advancements in language models in recent years that has revolutionized the field of natural language generation. Trained on vast amounts of data, the model exhibits impressive capabilities in generating human-like text across a wide range of domains. Short story generation, among its many applications, holds a significant promise. However, despite GPT-3's prowess, the challenge of generating coherent and engaging narratives persists due to the inherent complexity of storytelling.

WritingPrompts (Fan et al., 2018) is a collection of prompts and corresponding stories written by reddit users and is our source of human-written texts. In our approach this dataset was used to give ChatGPT prompts from which to generate short stories. With these and other stories from WritingPrompts a small pretrained GPT-2 model (124M parameres) from HuggingFace was finetuned.

In order to gain a deeper understanding of the enhancements achieved through distillation, four GPT-2 models have been explored: one finetuned solely on WritingPrompts, another utilizing only the answers generated by ChatGPT (as it produced three stories per prompt), a model trained exclusively on stories from ChatGPT with unique prompts, and finally, a dataset comprising of approximately 14,000 stories from ChatGPT and around 20,000 stories from WritingPrompts.

The paper continues with a section discussing related work. The third section presents the experimental setup and the next section contains the obtained results. Conclusions are the subject of the final section.

2. Related Work

Oftentimes, the ground truths can be elusive, requiring a carefully constructed model to capture their essence. In such cases, incorporating the outputs of another model can provide valuable insights and, in some instances, as it will be discussed later in this paper, simplify the task at hand.

Knowledge distillation involves utilizing the outputs of a larger and more performant model during the training process, enabling a lightweight model to learn to reproduce the behavior of its teacher model (Hinton et al., 2015) for a specific task. Although not ideal for achieving state-of-the-art models, this approach finds significant application in embedded systems where limited memory space demands the use of smaller models.

The idea of utilizing another model's outputs in the training process is not a recent development. In fact, earlier research, from the mid-2000s, explored similar concepts (Bucilă et al., 2006). In this research, the authors discuss a technique where an unlabeled dataset could be labeled using the predictions of a network and then incorporated into the training process of a new model. This pioneering work laid the foundation for the current understanding and further advancements in leveraging the outputs of other models for improved training.

The approach of Hinton et al. (2015) became a cornerstone in the field of model compression and knowledge distillation. They proved that this method holds great promise by applying it to the MNIST (Modified National Institute of Standards and Technology) dataset (LeCun et al., 1998), utilizing a large model for image classification. Additionally, they've done another test in the domain of speech recognition by employing an assembly of randomly initialized neural networks specifically trained for this task. In both experiments, the distilled models consistently outperformed their respective baseline models, which were solely trained on ground truths. Collectively, these findings, alongside subsequent innovations like sequence-to-sequence translation (Kim et al., 2016), provide compelling evidence of the efficacy of the distillation approach across many domains.

In recent years, distillation using a renowned model such as BERT, ELMo or GPT within the realm of natural language processing has garnered considerable attention, as evidenced by a range of notable examples (Tang et al., 2019) (Jiao et al., 2020) (West et al., 2021).

The introduction of ChatGPT ushers in a fresh surge of research papers dedicated to its distillation, with several of them already having been published mere months after the release of ChatGPT (Li et al., 2023) (Jiang et al., 2023).

3. Experimental Setup

3.1 ChatGPT Dataset

We handpicked approximately 4,500 prompts from the WritingPrompts training set to feed into ChatGPT via the OpenAI API (Application Programming Interface). For each prompt three potential story variations were generated. This approach was implemented to test whether it is more beneficial for a model to learn from a variety of stories coming from a single prompt or not.

The prompt selection process deliberately avoided subjects that violate OpenAI's terms of service, such as sexual acts, violence, or racism. Furthermore, requests for a specific narrative structure (e.g., *"[WP] The Butterfly Effect - Write the same scene twice , but with different endings", "[WP] Write a story that becomes a horror story in the last line ."*), breaking the fourth wall (e.g., *"[WP] Through the storyline, your character realizes he is written by you."*) or other miscellaneous aspects (e.g., *"[WP] Randomize your music playlist . Hit Play . Write a funny or scary story based on the title of the song playing ."*) were excluded because they are simply not core elements in most stories and might make it more difficult to learn from. Prompts that required prior knowledge of a fictional universe were also removed (e.g., *"[WP] Gandalf goes to Compton to do street magic.", "[WP] Two very method actors have been cast as Lex Luthor and Superman. Things have gotten out of hand."*).

A final curation of the resulting dataset was performed to remove the few instances in which ChatGPT failed to produce a story and offered explanations for its inability to do so (e.g., *"As an AI language model, I cannot play games, so I can't provide a story. Would you like a different prompt?"*). Additionally, any sections within the stories where ChatGPT referenced itself for various reasons were eliminated (e.g., *"As an AI language model, I do not have personal experiences or emotions like humans do. However, here's a story for you: [...]"*).

The data gathering process took place from May 18th to May 29th, 2023, using the default settings of the generator. When requesting a story from ChatGPT, the following message format was employed: *"Tell me a story about "* followed by the prompt.

3.2 Perplexity Analysis

Perplexity is a statistical measure of how closely a language model's prediction matches a target text. In the context of finetuning a NLG model we can use it as a means to determine which datasets are more aligned with the pretrained model's behavior and patterns.

To obtain the perplexity we raised the constant e to the power of the mean entropy loss of the model when passing the stories through. Since most stories had on average around 500 tokens, we could not fully make use of the GPT-2's potential as most of the time it did not have a lengthy context to predict from, making those guesses less accurate and consequently increasing the perplexity.

Table 1 shows the impact of different training sets on the perplexity of the GPT-2 model. The stories written by ChatGPT seem to closely resemble the patterns and style of the pretrained model meaning that it should be easier to finetune GPT-2 to behave like ChatGPT which is a capable story generator. When finetuning with either WritingPrompts or ChatGPT the other dataset's stories seem to perplex the resulting model even more. When both datasets are combined GPT-2's perplexity on WritingPrompts and ChatGPT's stories remain at the baseline's levels but the model gains more confidence in its predictions.

Table 1 – Perplexity of various GPT-2 models on different story samples

GPT-2 finetuned on	**Story Samples**		
	WritingPrompts	**ChatGPT**	**Its own stories**
No Finetune (Default)	34.60	12.82	8.08
WritingPrompts	29.40	22.18	10.76
ChatGPT	41.85	12.30	4.90
WritingPrompts + ChatGPT	33.13	12.76	5.01

3.3 ChatGPT's Stories Analysis

Before exploring the results of the models, it is insightful to first analyze the stories generated by ChatGPT, comparing them with human-written ones. It is important to keep in mind that these texts were created using the default settings of the OpenAI API.

The stories are well-written, coherent and follow closely the prompt's ideas. One notable tendency of the model is its inclination to reiterate phrases from the given prompt within the story's first few sentences. Consequently, it tackles almost immediately the theme of the prompt diverging from human-written stories which usually establish a proper setup beforehand.

Additionally, these generated stories tend to be considerably shorter than the ones found on WritingPrompts.

Table 2 shows a side-by-side comparison between a story in WritingPrompts and one generated by ChatGPT using the same prompt. Due to the prolonged setup of the WritingPrompts story, we highlighted only the part in which the prompt's theme is effectively addressed. We can see that ChatGPT has a much more straightforward approach.

The narrative structure exhibits robustness, with each paragraph containing a few sentences which are on average longer than their human-written counterparts. Despite this, they are generally easy to comprehend, even for individuals with lower reading proficiency. Nevertheless, the standard type/token ratio of the generated stories remains closely aligned (see Table 5) with that of the WritingPrompts, indicating that they do not shy away in vocabulary richness while pursuing intelligibility. The model rarely engages in extensive conversations between characters and typically presents only a single one. Compared to human-written stories, ChatGPT exhibits a far more pronounced tendency to utilize past tense and the third person (see Table 6 and Table 7).

Moreover, the model appears to associate the term "story" with the phrase "once upon a time" as approximately 28% of the generated stories start with these words. Notably, there are various recurring common names and phrases found throughout the entire dataset, such as: “in the end” – 15%, “from that day on” – 15%, “couldn't help but feel“ - 7%, “John” - 7%, “felt a sense of“ – 5.9%, “Anne” – 5.7%, “as time went on” – 5.1%, “Jack” – 4.9%, “there was a young” – 4.5%, “from that day forward” – 4.3%, “there was a man named” – 4%, “Sarah” – 3.9%, “days turned into weeks” – 3.6%, “couldn't shake the feeling” – 3.1%, “had always been fascinated” – 2.5%, “Emily” – 2.2%, “William” – 2.2%. Modifying the temperature of the generator may potentially reduce this behavior, but it could also impact the coherence of the story.

Table 2 – Example of story generated by humans vs ChatGPT

Prompt	WritingPrompts (Human-Written)	ChatGPT
[WP] There are creatures unknown to us here on earth , but we can not sense them because we lack that one sensory organ that can sense them .<newline>	“[…]Then the mood turned serious . My friend mentioned that his grandmother had recently passed . As soon as he said that , I saw one of those creatures phase into existence and jab its proboscis into his neck , not registering even a twitch of pain from him . I knew he would n't believe me if I told him . He 's a cynic . Of course he would n't . So I tried to carry on out conversation by cheering him back up . As soon as he cracked a smile , the creature pulled itself out of his neck , and the better my friend 's mood got , the farther away it seemed to get . We said our goodbyes and hung up . <newline> Over the past week , I have experimented with different things , different emotions to make these creatures appear and disappear . I learned that they are drawn to negative emotion - they feed on something in the host 's blood . Because of the little organ in my arm , I am aware of their presence. […]”	“Once upon a time, there were creatures beyond our wildest imagination that roamed the earth. They had been living among us for centuries, yet we were completely oblivious to their existence. There were rumors and legends about these mysterious beings, but they went unnoticed because we did not possess the sensory organs to perceive them. <newline> <newline> These creatures were known as 'Wispies', and they lived in a parallel dimension that was invisible to us. They emitted a faint light that was barely visible to those who possessed a sixth sense, but for the majority of us, they remained a mystery. Humans lacked the sensory capacity to perceive these beautiful creatures, but they lived side by side with us, observing our every move, and watching us grow and evolve. […]”

4. Results

4.1 Training on WritingPrompts

The model finetuned on the short stories from WritingPrompts exhibits significant limitations, particularly in terms of coherence and clarity, even at

sentence level (see
Table 3 for an example). The generated stories suffer from abrupt shifts between different scenes. It appears that the model may find it easier to learn from dialogue as it is quite common in the generated texts, possibly due to the narrower vocabulary used in conversational contexts.

It is very likely that the model inadvertently developed a tendency to overlook the initial prompt as by the time its context window reached the main theme of the story it had already forgotten the initial prompt. GPT-2 has a limited attention span and the stories from WritingPrompts often involve extensive setups before delving into the central idea.

Table 3 – Story generated by a GPT-2 model finetuned only on WritingPrompts

Prompt	Text
"Aliens start abducting humans."	"`` What do you think is going to happen? " "`` I said.`` You know, I've been doing this for a long time. "I'm not the only one who has come up with this idea, but I don't really know how to explain it to you. It's something that I have been trying to do for years, and I just can't seem to get enough of it. I just want to know what it is and what I can do to make it work. […]"

4.2 Training on ChatGPT's stories

The comparison between using only one story per prompt and using three different stories for each prompt is inconclusive as both models achieved similar results. They outperformed the one trained on the WritingPrompts by maintaining the theme of the prompt throughout most of the text. Moreover, the stories generated did not rely heavily anymore on dialogue having a narrative structure more in tune with ChatGPT with a richer vocabulary.

Unfortunately, both models became influenced by the recurring phrases and words from ChatGPT and started using them nearly twice as often as in the training set (see

Table 4). Certain phrases like "had always been fascinated" or "once upon a time" exhibit an abnormally large occurrence rate, raising the possibility that the pretrained GPT-2 model might already possess a bias towards these words. Through the process of finetuning the model on ChatGPT's stories, the unintentional strengthening of this existing bias might have occurred.

Table 4 – Comparison between the frequency of phrases and words in stories. Note that the GPT-2 model trained solely on WritingPrompts was not included because this behavior only appears in models trained with ChatGPT's texts.

		GPT-2 trained on:		
Phrase/Word	ChatGPT	One Story per Prompt	Three Stories per Prompt	WP + ChatGPT
"once upon a time"	28%	96%	94%	52%
“in the end”	15%	18%	29%	17%
“from that day on”	15%	18%	16%	6%
“couldn't help but feel”	7%	29%	16%	6%
“John”	7%	39%	20%	9%
“felt a sense of“	5.9%	14%	11%	6%
“as time went on”	5.1%	14%	23%	15%
“had always been fascinated”	2.5%	78%	58%	38%

Table 5 – Comparison between stories coming from Writing Prompts dataset, ChatGPT and our own GPT-2 models regarding the length of sentences and the richness of the vocabulary. WP stands for Writing Prompts and SPP for Story per Prompt

	Words per sentence		Standard Type/Token Ratio	
	mean	**std**	**mean**	**std**
Writing Prompts	12.92	7.76	0.58	0.093
GPT-2 on Writing Prompts	16.27	2.43	0.48	0.05
ChatGPT	**19.63**	4.24	0.57	0.063
GPT-2 on 1 SPP	19.65	3.78	0.58	0.035
GPT-2 on 3 SPP	17.37	2.47	0.59	0.042
GPT-2 on 3SPP+ChatGPT	16.90	2.91	0.55	0.063

4.3 Training on ChatGPT’s stories and Writing Prompts

Merging ChatGPT's stories with the ones from WritingPrompts was the approach chosen in order to address the deficiencies of the previous models. Doing this a model was derived that possesses a comparable narrative structure robustness to ChatGPT's stories while exhibiting a reduced bias towards specific phrases (see

Table 4) that are less prevalent in human-written narratives. The model manages quite often to construct coherent sentences that incorporate multiple clauses, although it achieves this rather through a lack of specificity. The stories tend to follow the theme of the prompt at least in the first few sentences (see Table 8). Nevertheless, it attains performances that are on par with the results obtained in other similar papers (Fan et al., 2018) (Fan et al., 2019) (Goldfarb-Tarrant et al., 2020).

Long and intricate prompts, which are frequently encountered in WritingPrompts, present a significant challenge for the model, leading it to primarily focus on the theme of the prompt while neglecting its other aspects.

Table 6 – Comparison between stories coming from Writing Prompts dataset, ChatGPT and our own GPT-2 models regarding the tense frequency. WP stands for Writing Prompts and SPP for Story per Prompt.

	Present Tense Frequency		Past Tense Frequency		Future Tense Frequency	
	mean	**std**	**mean**	**std**	**mean**	**std**
Writing Prompts	0.41	0.19	0.55	0.2	0.03	0.04
GPT-2 on Writing Prompts	0.41	0.19	0.52	0.21	0.05	0.04
ChatGPT	0.23	0.12	**0.75**	0.12	0.006	0.02
GPT-2 on 1 SPP	0.28	0.10	0.69	0.10	0.019	0.031
GPT-2 on 3 SPP	0.24	0.09	0.74	0.09	0.009	0.017
GPT-2 on 3SPP+ChatGPT	0.30	0.14	0.67	0.16	0.02	0.04

Table 7 – Comparison between stories coming from Writing Prompts dataset, ChatGPT and our own GPT-2 models regarding the person frequency. WP stands for Writing Prompts and SPP for Story per Prompt.

	First Person Frequency		Second Person Frequency		Third Person Frequency	
	mean	**std**	**mean**	**std**	**mean**	**std**
Writing Prompts	0.46	0.28	0.12	0.14	0.4	0.29
GPT-2 on Writing Prompts	0.51	0.16	0.19	0.14	0.28	0.16
ChatGPT	0.24	0.35	0.04	0.14	**0.71**	0.37
GPT-2 on 1 SPP	0.12	0.14	0.09	0.08	0.77	0.19
GPT-2 on 3 SPP	0.11	0.10	0.09	0.09	0.78	0.16
GPT-2 on 3SPP+ChatGPT	0.31	0.29	0.11	0.11	0.56	0.33

5. Conclusion

This paper explores the potential of distillation by leveraging ChatGPT's responses to generate stories from prompts. Our GPT-2 model delivers impressive results despite having a relatively small number of parameters but it's still a long way to go until it consistently outputs coherent texts that stay true to the prompt's theme.

The perplexity scores have shown that data generated by ChatGPT is much more fitted to be used as a training set compared to human-written texts but it is proving to be quite difficult to obtain a generator performing similarly to its teacher.

An important aspect to mention is that neither Writing Prompts nor ChatGPT are exempt from biases, and these can greatly affect the outcome of the training process. Human-generated stories tend to slowly tackle the ideas of a prompt which makes it difficult for models with their limited context windows to learn. On the other hand, ChatGPT is prone to using certain phrases in any scenario and to engage the prompt's ideas almost immediately making the story less intriguing. By counterbalancing these two datasets the best of both worlds can be achieved.

Table 8 – Stories generated by a GPT-2 model finetuned on ChatGPT's and WritingPrompts's stories

Prompt	*Text*
*"Listen, you're fine, I'm- stop screaming- I'm not going to hurt you. Yes, opening the Dark Book summons The One Who Ends, but that whole thing where I kill the reader is hogwash. Now, would you kindly direct me to the one who tricked you into opening my book? *Intent* is important, after all."*	*"I nodded in agreement, feeling a sense of relief wash over me. I had always been a bit of a skeptic, and I never really believed in the supernatural. But then, one day, something strange happened. As I was reading through the dark book, a figure appeared from the shadows. It was a young woman, dressed in a black robe, with long black hair and piercing blue eyes. She looked like she was about to give birth to a child. At first, it was hard to tell what to make of it. Was it a dream or a nightmare? But as I began to read more and more, my curiosity got the better. Suddenly, the figure disappeared, leaving me with a feeling of unease and confusion. What had happened to me? Was I hallucinating? Had I been transported to another dimension?"*
"[WP] Berlin, 1943. You're a Jew who has been successfully keeping a low profile while working as a clerk under the 3rd Reich. You just found out you've won a contest to spend the day with the Führer."	*""As I walked through the streets of Berlin, I couldn't help but feel a sense of pride and accomplishment. I had won the competition, and now I had the chance to make history. The day of the contest had come and gone, but I was determined to prove my worth to the people of Germany. It was a tough day for me, as I struggled to keep track of every detail of my life. But I knew that I could make a difference in the lives of those around me. And so I set out on a journey that would change the course of history forever. In the year 1943, a young man named Adolf Hitler was born. [...]"*
[WP] You're a fortune teller with a magical orb that shows you the future of those you are reading. A young man walks	*"Once upon a time, in a far-off land, there was a young woman named Sarah. She had always been fascinated by magic, and she loved it so much that she decided to try it herself. However, as she grew older, she began to notice strange things happening in*

into your shop and asks you to tell him his future. When you peer into the orb, you witness your own brutal murder.	*the world around her. One day, Sarah was walking down the street, when she saw a man standing in front of her shop. He was wearing a black robe and a hooded cloak that covered his entire body. As he approached her, he asked her if she was interested in magic. Sarah couldn't believe her ears, but she quickly realized that he was talking about something far more sinister than anything she had ever seen before. […]"*
[WP] an unhappy person gains the superpower of mind control.	*"Once upon a time, there was a man named John. He had always been fascinated by the idea of using his mind to control people. However, as time went on, he began to realize that he was not the only one who possessed this ability. People around the world were beginning to use it to their advantage, and John was no exception. One day, John decided to take matters into his own hands. As he sat in his living room, staring at his computer screen, thinking about all the things he could do to improve his mental state, something caught his eye. Suddenly, a message popped up on his screen that read, "You have the ability to manipulate the minds of others." John's mind was filled with thoughts and emotions, but he couldn't help but feel a sense of unease running through his body. It was as if something was off about him, or something had changed in him. John knew that this was something he had to deal with, so he took a deep breath and started to open his eyes. The message was clear and concise, with a simple message in bold letters: "Your mind is not your mind. You are not a person. Your mind does not belong to you. Please do not use this power to harm others. Do not harm yourself. This power is for the benefit of all of us, not just those who are affected by it."*

References

Brown T., Mann B., Ryder N., Subbiah M., D Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. (2020). *Language Models are Few-Shot Learners.* Retrieved June 5, 2023, from https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html https://doi.org/10.48550/arXiv.2005.14165

Bucilă C., Caruana R., Niculescu-Mizil A.. (2006). *Model Compression.* Retrieved May 3, 2023, from https://doi.org/10.1145/1150402.1150464

Fan A., Lewis M., Dauphin Y. (2018). *Hierarchical Neural Story Generation*. Retrieved November 5, 2021, from https://arxiv.org/abs/1805.04833 https://doi.org/10.48550/arXiv.1805.04833

Fan A., Lewis M., Dauphin Y. (2019). *Strategies for Structuring Story Generation*. Retrieved November 5, 2021, from https://arxiv.org/abs/1902.01109 https://doi.org/10.48550/arXiv.1902.01109

Goldfarb-Tarrant S., Chakrabarty T., Weischedel R., Peng N. (2020). *Content Planning for Neural Story Generation with Aristotelian Rescoring*. Retrieved December 11, 2021, from https://arxiv.org/abs/2009.09870 https://doi.org/10.48550/arXiv.2009.09870

Hinton G., Vinyals O., Dean J. (2015). *Distilling the Knowledge in a Neural Network.* Retrieved July 3, 2023, from https://arxiv.org/abs/1503.02531 https://doi.org/10.48550/arXiv.1503.02531

Jiang Y., Chan C., Chen M., Wang W. (2023). *Lion: Adversarial Distillation of Closed-Source Large Language Model.* Retrieved June 3, 2023, from https://arxiv.org/abs/2305.12870 https://doi.org/10.48550/arXiv.2305.12870

Jiao X., Yin Y., Shang L., Jiang X., Chen X., Li L., Wang F., Liu Q. (2020). *TinyBERT: Distilling BERT for Natural Language Understanding.* Retrieved May 5, 2023, from https://arxiv.org/abs/1909.10351 https://doi.org/10.48550/arXiv.1909.10351

Kim Y., Rush A. M. (2016). *Sequence-Level Knowledge Distillation.* Retrieved May 4, 2023, from https://arxiv.org/abs/1606.07947 https://doi.org/10.48550/arXiv.1606.07947

LeCun Y., Cortes C., Burges C. J. C. (1998). *THE MNIST DATABASE of handwritten digits.* Retrieved 29 June, 2023, from http://yann.lecun.com/exdb/mnist/

Li J., Gui L., Zhou Y., West D., Aloisi C., He Y. (2023). *Distilling ChatGPT for Explainable Automated Student Answer Assessment.* Retrieved June 3, 2023, from https://arxiv.org/abs/2305.12962 https://doi.org/10.48550/arXiv.2305.12962

Tang R., Lu Y., Liu L., Mou L., Vechtomova O., Lin J. (2019). *Distilling Task-Specific Knowledge from BERT into Simple Neural Networks.* Retrieved May 4, 2023, from https://arxiv.org/abs/1903.12136 https://doi.org/10.48550/arXiv.1903.12136

West P., Bhagavatula C., Hessel J., Hwang J. D., Jiang L., Le Bras R., Lu X., Welleck S., Choi Y. (2021). *Symbolic Knowledge Distillation: from General Language Models to Commonsense Models.* Retrieved 8 May, 2023, from https://arxiv.org/abs/2110.07178 https://doi.org/10.48550/arXiv.2110.07178