

A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation

Xiaowei Huang¹, Wenjie Ruan¹, Wei Huang¹, Gaojie Jin¹, Yi Dong¹,
Changshun Wu², Saddek Bensalem², Ronghui Mu¹, Yi Qi¹, Xingyu
Zhao¹, Kaiwen Cai¹, Yanghao Zhang¹, Sihao Wu¹, Peipei Xu¹, Dengyu
Wu¹, Andre Freitas³, and Mustafa A. Mustafa³

¹University of Liverpool, UK
²Université Grenoble Alpes, France
³University of Manchester, UK

Abstract

Large Language Models (LLMs) have exploded a new heatwave of AI, for their ability to engage end-users in human-level conversations with detailed and articulate answers across many knowledge domains. In response to their fast adoption in many industrial applications, this survey concerns their safety and trustworthiness. First, we review known vulnerabilities of the LLMs, categorising them into inherent issues, intended attacks, and unintended bugs. Then, we consider if and how the Verification and Validation (V&V) techniques, which have been widely developed for traditional software and deep learning models such as convolutional neural networks, can be integrated and further extended throughout the lifecycle of the LLMs to provide rigorous analysis to the safety and trustworthiness of LLMs and their applications. Specifically, we consider four complementary techniques: falsification and evaluation, verification, runtime monitoring, and ethical use. Considering the fast development of LLMs, this survey does not intend to be complete (although it includes 300+ references), especially when it comes to the applications of LLMs in various domains, but rather a collection of organised literature reviews and discussions to support the quick understanding of the safety and trustworthiness issues from the perspective of V&V.

Contents

1	Introduction	3
2	Large Language Models	5
2.1	Categories of Large Language Models	5

2.1.1	Text-based Conversational AI	5
2.1.2	Text-based Image Synthesis	5
2.2	Lifecycle of LLMs	6
2.3	Key Techniques Relevant to Safety and Trustworthiness	7
2.3.1	Reinforcement learning from human feedback (RLHF)	7
2.3.2	Guardrails	8
3	Vulnerabilities	8
3.1	Performance Issues	9
3.1.1	Factual errors	9
3.1.2	Reasoning errors	10
3.2	Sustainability Issues	10
3.3	Other Inherent Trustworthiness and Responsibility Issues	10
3.4	Unauthorised Disclosure and Privacy Concerns	12
3.5	Robustness Gap	12
3.6	Backdoor Attack	13
3.6.1	Design of Backdoor Trigger	14
3.6.2	Backdoor Embedding Strategies	14
3.6.3	Expression of Backdoor	14
3.7	Poisoning and Disinformation	15
3.8	Incidental Exposure of User Information	16
3.9	Bias and Discrimination	16
4	General Verification Framework	16
5	Falsification and Evaluation	17
5.1	Red Teaming	17
5.2	Prompt Injection	17
5.3	Comparison with Human Experts	19
5.4	Benchmarks	19
5.5	Testing and Statistical Evaluation	19
6	Verification	21
6.1	Verification on Natural Language Processing Models	21
6.1.1	Verification via Interval Bound Propagation	21
6.1.2	Verification via Abstract Interpretation	22
6.1.3	Verification via Randomised Smoothing	23
6.2	Black-box Verification	24
6.3	Robustness Evaluation on LLMs	25
6.4	Towards Smaller Models	26
7	Runtime Monitor	26
7.1	Monitoring Out-Of-Distribution	27
7.2	Monitoring Output Failures	28
7.3	Perspective	29

8	Regulations and Ethical Use	30
8.1	Regulate or Ban?	30
8.2	Responsible AI Principles	31
8.3	Transparency and Explainability	31
9	Conclusions	32

1 Introduction

A Large Language Model (LLM) is a deep learning model equipped with a massive amount of learnable parameters (commonly reaching more than 10 billion, as indicated in Figure 1). LLMs are attention-based sequential models based on the transformer architecture [116], which consistently demonstrated the ability to learn universal representations of language. The universal representations of language can then be used in various natural language processing tasks (NLP). The recent scale-up of these models, in terms of both numbers of parameters and pre-trained corpora, has confirmed the universality of transformers as mechanisms to encode language representations. At a specific scale, these models started to exhibit in-context learning [184, 278], and the properties of learning from few examples (zero/one/few-shot - without the need for fine-tuning) and from natural language prompts (complex instructions which describe the behavioural intent that the model needs to operate). Recent works on Reinforcement Learning via Human Feedback (RLHF) [190] have further developed the ability of these models to align and respond to increasingly complex prompts, leading to their popularisation in systems such as ChatGPT and their use in a large spectrum of applications. The ability of LLMs to deliver sophisticated linguistic and reasoning behaviour, has pushed their application beyond their intended operational envelope.

While being consistently fluent, LLMs are prone to hallucinations [228], stating factually incorrect statements [227], lacking necessary mechanisms of safety, transparency and control [239], among many others. The goal of this paper is to provide a review of known vulnerabilities of LLMs and, more importantly, to investigate how the V&V techniques can be adapted to improve the safety and trustworthiness of LLMs. While there are several surveys on LLMs [296, 291], as well as a categorical archive of ChatGPT failures [47], to the best of our knowledge, this is the first work that provides a comprehensive discussion on the safety and trustworthiness issues, from the perspective of the V&V.

With the rising of LLMs and its wide applications, the need to ensure their safety and trustworthiness become prominent. Considering the broader subject of deep learning systems, to support their safety and trustworthiness, a diverse set of technical solutions have been developed by different research communities. For example, the machine learning community is mainly focused on adversarial attacks, outlier detectors, adversarial training, and explainable AI. The human-computer interaction community is focused on engaging the learning systems in the interactions with end users to improve the end users' confidence. Formal methods community treats ML models as yet another symbolic system (evidenced by their consideration of neurons, layers, etc.) and adapts existing formal methods tools to work on the new systems [125]. While

research has been intense on these individual methods, a synergy among them has not been addressed. *Without a synergy, it is hard, if not impossible, to rigorously understand the causality between methods and how they might collectively support the safe and trusted autonomous systems at runtime.* This survey is rooted in the field of AI assurance, aiming to apply a collection of rigorous V&V methods throughout the lifecycle of ML models, to provide assurance on the safety and trustworthiness. An illustrative diagram is given in Figure 1 for general ML models.

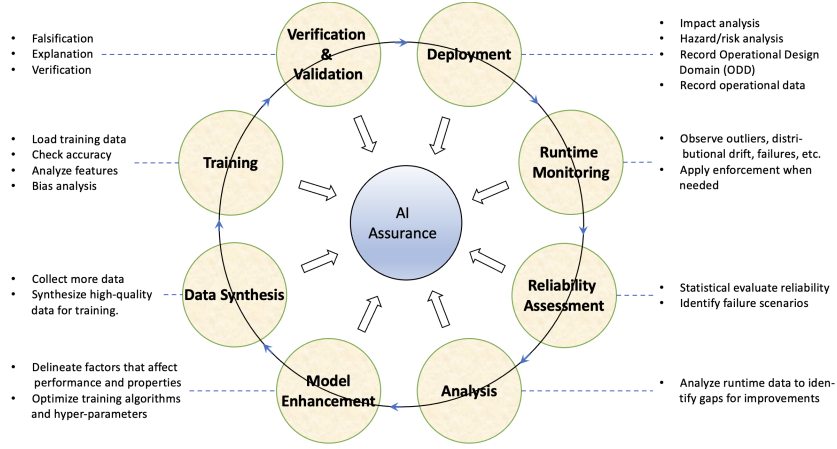


Figure 1: Summarisation of lifecycle V&V methods to support AI Assurance

V&V techniques have been successful in supporting the reliable and dependable development of software and hardware that are applied to safety-critical systems, and have been adapted to work with machine learning models, mainly focusing on the convolutional neural networks for image classification (see surveys such as [125, 169] and textbooks such as [124]), but also extended to consider e.g., object detection, deep reinforcement learning, and recurrent neural networks. This paper discusses how to extend further V&V to deal with the safety and trustworthiness challenges of LLMs.

V&V are independent procedures that are used together for checking that a system (or product, service) meets requirements and specifications and that it fulfills its intended purpose [7]. Among them, verification techniques check the system against a set of design specifications, and validation techniques ensure that the system meets the user's operational needs. From software, convolutional neural networks to LLMs, the scale of the systems grows significantly, which makes the usual V&V techniques less capable due to their scalability issues. White-box V&V techniques that take the learnable parameters as their algorithmic input will not work well in practice. Instead, the research should focus on black-box techniques, on which some research has started for convolutional neural networks. In addition, V&V techniques need to consider the *non-deterministic nature* of LLMs (i.e., different outputs for two tests with identical input), which is a noticeable difference with the usual neural networks, such as convolutional neural networks and object detectors, that currently most V&V techniques work on.

The structure of the paper is as follows. In Section 2, we review the LLMs and its categories, its lifecycle, and several techniques introduced to improve safety and trustworthiness. Then, in Section 3, we present a review of existing vulnerabilities. This is followed by a general verification framework in Section 4. The framework includes V&V techniques such as falsification and evaluation (Section 5), verification (Section 6), runtime monitor (Section 7), and ethical use (Section 8). We conclude the paper in Section 9.

2 Large Language Models

This section summarises the categories of machine learning tasks based on LLMs, followed by a discussion of the lifecycle of LLMs. We will also discuss a few fundamental techniques relevant to the safety analysis.

2.1 Categories of Large Language Models

LLMs have been applied to many tasks, such as text generation, content summary, conversational AI (i.e., chatbots), and image synthesis. Other LLMs applications can be seen as their adaptations or further applications. In the following, we discuss the two most notable categories of LLMs.

2.1.1 Text-based Conversational AI

LLMs are designed to understand natural language and generate human-like responses to queries and prompts. Almost all Natural Language Processing (NLP) tasks (e.g., language translation [48], chatbots [168, 104] and virtual assistants [244]) have witnessed tremendous success with Transformer-based pretrained language models (T-PTLMs), relying on Transformer [248], self-supervised learning [130, 173] and transfer learning [115, 213] to process and understand the nuances of human language, including grammar, syntax, and context.

The well-known text-based LLMs include GPT-1 [199], BERT [78], XLNet [275], RoBERTa [175], ELECTRA [69], T5 [201], ALBERT [152], BART [159], and PEGASUS [287]. These models can learn general language representations from large volumes of unlabelled text data through self-supervised learning and subsequently transfer this knowledge to specific tasks, which has been a major factor contributing to their success in NLP [138]. Kaplan et al. [142] demonstrated that simply increasing the size of T-PTLMs can lead to improved performance [138]. This finding has spurred the development of LLMs such as GPT-3 [49], PANGU [282], GShard [158], Switch-Transformers [90] and GPT-4 [189].

2.1.2 Text-based Image Synthesis

The transformer model [247] has become the standard choice for Language Modelling tasks, but it has also found widespread integration in text-to-image tasks. We present a chronological overview of the advancements in text-to-image research. DALL-E [204]

is a representative approach that leverages Transformers for a text-to-image generation. The methodology involves training a dVAE [209] and subsequently training a 12B decoder-only sparse transformer supervised by image tokens from the pre-trained dVAE. The transformer generates image tokens solely based on text tokens during inference. The resulting image candidates are evaluated by a pretrained CLIP model [198] to produce the final generated image. StableFusion [210] differs from DALL-E [204] by using a diffusion model instead of a Transformer to generate latent image tokens. To incorporate text input, StableFusion [210] first encodes the text using a transformer then conditions the diffusion model on the resulting text tokens. GLIDE [187] employs a transformer model [247] to encode the text input and then trains a diffusion model to generate images that are conditioned on the text tokens directly. DALL-E2 [203] effectively leverages LLMs by following a three-step process. First, a CLIP model is trained using text-image pairs. Next, using text tokens as input, an autoregressive or diffusion model generates image tokens. Finally based on these image tokens, a diffusion model is trained to produce the final image. Imagen [217] employs a pre-trained text encoder, such as BERT [77] or CLIP [198], to encode text. It then uses multiple diffusion models to train a process that generates images that start from low-resolution and gradually progress to high-resolution. Parti [280] demonstrates that a VQGAN [88] and Transformer architecture can achieve superior image synthesis outcomes compared to previous approaches, even without utilising a diffusion model. The eDiff-I model [36] has recently achieved state-of-the-art performance on the MS-COCO dataset [167] by leveraging a combination of CLIP and diffusion models.

In summary, text-to-image research commonly utilises transformer models [247] for encoding text input and either the diffusion model or the decoder of an autoencoder for generating images from latent text or image tokens.

2.2 Lifecycle of LLMs

Figure 2 illustrates the lifecycle stages of LLMs. The offline model construction is formed of three steps [291]: pre-training, adaptation tuning, and utilisation improvement, such that each step includes several interleaving sub-steps. In general, the *pre-training* step is similar to the usual machine learning training that goes through data collection, architecture selection, and training. On *adaptation tuning*, it might conduct instruction tuning [178] to learn from task instructions, and alignment tuning [190, 67] to make sure LLMs are aligned with human values, e.g., fair, honest, and harmless. Beyond this, to improve the interaction with the end users, *utilisation improvements* may be conducted through e.g., in-context learning [49], chain-of-thought learning [257].

Once an LLM is trained, an *evaluation* is needed to ensure that its performance matches the expectation. Usually, we consider the evaluation from three perspectives: evaluation on basic performance metrics, safety analysis to evaluate the consequence of applying the LLM in an application, and the evaluation through publicly available benchmark datasets. The evaluation will determine if the LLM is acceptable (for pre-specified criteria), and if so, the process will move forward to the deployment stage. Otherwise, at least one failure will be identified, and the process will move back to either of the three training steps.

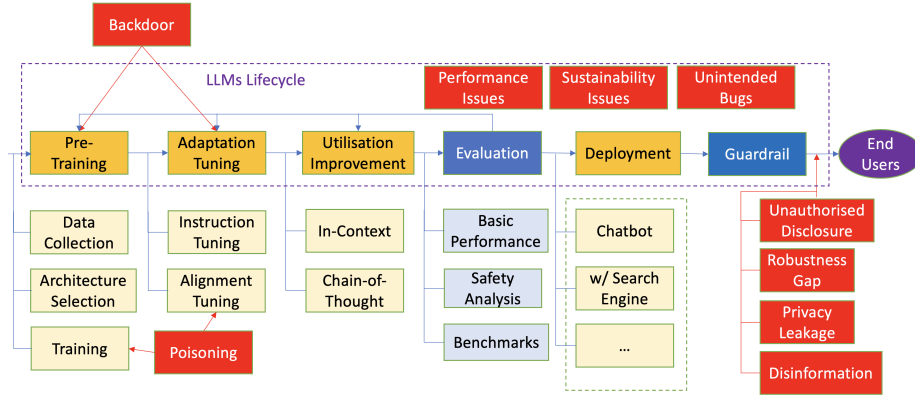


Figure 2: Large Language Models: Lifecycle and Vulnerabilities

On the *deployment* stage, we will determine how the LLM will be used. For example, it could be available in a web platform for direct interaction with end users, such as the ChatGPT¹. Alternatively, it may be embedded into a search engine, such as the new Bing². Nevertheless, according to the common practice, a *guardrail* is imposed on the conversations between LLMs and end users to ensure that the AI regulation is maximally implemented.

2.3 Key Techniques Relevant to Safety and Trustworthiness

In the following, we discuss two fundamental techniques that are distinct from the usual deep learning models and have been used by e.g., ChatGPT to improve safety and trustworthiness.

2.3.1 Reinforcement learning from human feedback (RLHF)

RLHF [68, 190, 33, 34, 35, 189, 151, 300] plays a crucial role in the training of language models, as it allows the model to learn from human guidance and avoid generating harmful content. In essence, RLHF assists in aligning language models with safety considerations through fine-tuning with human feedback. OpenAI initially introduced the concept of incorporating human feedback to tackle complex reinforcement learning tasks in [68], which subsequently facilitated the development of more sophisticated LLMs, from InstructGPT [190] to GPT4 [189]. According to InstructGPT [190], the RLHF training process typically begins by learning a reward function intended to reflect what humans value in the task, utilising human feedback on the model’s outputs. Subsequently, the language model is optimised via an RL algorithm, such as PPO [220], using the learned reward function. Reward model training and fine-tuning with RL can be iterated continuously. More comparison data is collected on the current best policy, which is used to train a new reward model and policy. The InstructGPT models

¹<https://openai.com/blog/ChatGPT>

²<https://www.bing.com/new>

demonstrated enhancements in truthfulness and reductions in generating toxic outputs while maintaining minimal performance regressions on public NLP datasets.

Following InstructGPT, Red Teaming language models [34] introduces a harmlessness preference model to help RLHF to get less harmful agents. The comparison data from red team attacks is used as the training data to develop the harmlessness preference model. [33] utilised the helpful and harmless datasets in preference modelling and RLHF to fine-tune LLMs. They discovered that there was a significant tension between helpfulness and harmlessness. Experiments showed helpfulness and harmlessness model is significantly more harmless than the model trained only on helpfulness data. They also found that alignment with RLHF has many benefits and no cost to performance, like combining alignment training with programming ability and summarisation. [96] found that LLMs trained with RLHF have the capability for moral self-correction. They believe that the models can learn intricate normative concepts such as stereotyping, bias, and discrimination that pertain to harm. Constitutional AI [35] trains the preference model by relying solely on AI feedback, without requiring human labels to identify harmful outputs. To push the process of aligning LLMs with RLHF, an open-sourced modular library, RL4LLMs, and evaluation benchmark, GRUE, designed for optimising language generator with RL are introduced in [202]. Inspired by the success of RLHF in language-related domains, fine-tuning approaches that utilise human feedback to improve text-to-image models [154, 270, 268] have gained popularity as well. To achieve human-robot coexistence, [105] proposed a human-centred robot RL framework consisting of safe exploration, safety value alignment, and safe collaboration. They discussed the importance of interactive behaviours and four potential challenges within human-robot interactive procedures. Although many works indicate that RLHF could decrease the toxicity of generations from LLMs, the induced RLHF, like introducing malicious examples by annotators [55], may cause catastrophic performance and risks. We hope better techniques that lead to transparency, safe and trustworthy RLHF will be developed in the coming future.

2.3.2 Guardrails

Considering that some LLMs are interacting directly with end-users, it is necessary to put a layer of protection (as was done by OpenAI on ChatGPT) when the end users ask for information about violence, profanity, criminal behaviours, race, or other unsavoury topics. In such cases, a response is provided with the LLM refuses to provide information. While this is a very thin layer of protection because there are many tricks (such as prompt injections that will be reviewed in Section 5.2) to circumvent it, it enhances the social responsibility of LLMs.

3 Vulnerabilities

This section presents a review of the known types of vulnerabilities. The vulnerabilities can be categorised into inherent issues, intended attacks, and unintended bugs. *Inherent issues* are vulnerabilities that cannot be readily solved by the LLMs themselves. However, they can be gradually improved with, e.g., more data and novel training methods.

Inherent issues include performance weaknesses, which are those aspects that LLMs have not reached the human-level intelligence, and sustainability issues, which are because the size of LLMs is significantly larger than the usual machine learning models. Their training and daily execution can have non-negligible sustainability implications.

Moreover, trustworthiness and responsibility issues are inherent to the LLMs. *Intended attacks* are initiated by malicious attackers, which attempt to implement their goals by attacking certain stages in the LLMs lifecycle. Known intended attacks include robustness gap, backdoor attack, poisoning, disinformation, privacy leakage, and unauthorised disclosure of information. Finally, with the integration of LLMs into broader applications, there will be more and more *unintended bugs* that are made by the developers unconsciously but have serious consequences, such as bias and discrimination (that are usually related to the quality of training data), and the recently reported incidental exposure of user information.

Figure 2 suggests how the vulnerabilities may be exploited in the lifecycle of LLMs. While inherent issues and unintended bugs may appear in any stage of the lifecycle, the intended attacks usually appear in particular stages of the lifecycle. For example, a backdoor attack usually occurs in pre-training and adaptation tuning, in which the backdoor trigger is embedded, and poisoning usually happens in training or alignment tuning, when the LLMs acquires information/data from the environment. Besides, many attacks occur upon the interaction between end users and the LLMs using specific, well-designed prompts to retrieve information from the LLMs. We remark that, while there are overlapping, LLMs and usual deep learning models (such as convolutional neural networks or object detectors) have slightly different vulnerabilities, and while initiatives have been taken on developing specification languages for usual deep learning models [40, 127], such efforts may need to be extended to LLMs.

3.1 Performance Issues

Unlike traditional software systems, which run according to the rules that can be deterministically verified, neural network-based deep learning systems, including large-scale LLMs, have their behaviour determined by the complex models learned from data through optimisation algorithms. It is unlikely that an LLM performs 100% correctly. Performance issues include at least the following two categories: factual errors and reasoning errors.

3.1.1 Factual errors

Factual errors refer to situations where the output of an LLM contradicts the truth. E.g., when asked to provide information about the expertise in the computer science department at the University of Liverpool, the ChatGPT refers to people who were never affiliated with the department. More serious errors can be generated, including notably wrong medical advice. Additionally, it is interesting to note that while LLMs can perform across different domains, their reliability may vary across domains. For example, [225] shows that ChatGPT significantly under-performs in law and science questions. Investigating if this is related to the training dataset will be interesting.

3.1.2 Reasoning errors

It has been discovered that, when given calculation or logic reasoning questions, ChatGPT may not always provide correct answers. This is mainly because, instead of actual reasoning, LLMs fit the questions with prior experience learned from the training data. If the statements of the questions are close to those in the training data, it will give correct answers with a higher probability. Otherwise, with carefully crafted prompt sequence, wrong answers can be witnessed [170, 94].

3.2 Sustainability Issues

Sustainability issues, which are measured with, e.g., economic cost, energy consumption, and carbon dioxide emission are also inherent to the LLMs. While excellent performance, LLMs require high costs and consumption in all the activities in its life-cycle. Notably, ChatGPT was trained with 30k A100 GPUs (each one is priced at around \$10k), and every month’s energy consumption cost at around \$1.5m.

In Table 1, we summarise the hardware costs and energy consumption from the literature for a set of LLMs with varied parameter sizes and training dataset sizes. Moreover, the carbon dioxide emission can be estimated with the following formula:

$$tCO_2eq = 0.385 \times GPU_h \times (GPU \text{ power consumption}) \times PUE \quad (1)$$

where GPU_h is the GPU hours, GPU power consumption is the energy consumption as provided in Table 1, and PUE is the Power Usage Effectiveness (commonly set as a constant 1.1). Precisely, it has been estimated that training a GPT-3 model consumed 1,287 MWh, which emitted 552 ($= 1287 * 0.385 * 1.114$) tons of CO_2 [192].

3.3 Other Inherent Trustworthiness and Responsibility Issues

Some issues occur during the lifecycle that could lead to concerns about the trustworthiness and responsibilities of LLMs. Generally, these can be grouped into two sub-classes concerning the training data and the final model.

For the training data, there are issues around the copyright [145], quality, and privacy of the training data. There is a significant difference between LLMs and other ML models regarding the data being used for training. In the latter case, specific (well-known/-structured) datasets are usually used in the training process. Ideally, these datasets are properly pre-processed, anonymised, etc.; if needed, users have also given consent about using of their data. It is well known that ChatGPT crawls the internet and uses the gathered data to train. On the other hand, for LLMs, the data used for training needs to be more understood. In most cases, users have not provided any consent; most likely they are even unaware that their data contain personal information and that their data have been crawled and used in LLM training. This makes ChatGPT, and LLMs in general, privacy-nightmare to deal with and opens the door to many privacy leakage attacks. Even the model owners would need to determine the extent of private risk their model could pose.

Model	Parameter size	Dataset size	Hardware	Energy
BERT-base [77]	110 million	3.3b words	16 TPU chips	-
BERT-large [77]	340 million	3.3b words	64 TPU chips	-
GPT-3 [50]	175 billion	499 billion tokens	10,000 NVIDIA V100	1287 MWh
Megatron Turing NLG [230]	530 billion	338.6 b	4480 NVIDIA A100-80GB	>900MWh
ERNIE 3.0 [237]	260 billion	4Tb texts	384 NVIDIA V100 GPU	-
GLaM [81]	1.2 trillion	1.6 trillion	1,024 Cloud TPU-V4	456MWh
Gopher [200]	280 billion	300 billion	4096 TPUv3	1066 MWh
PanGu- α [283]	200 billion	1.1TB	2048 Ascend 910 AI processors	-
LaMDA [241]	137 billion	1.56T words	1024 TPU-v3	451MWh
GPT-NeoX [45]	20 billion	825 GiB	96 NVIDIA A100-SXM4-40GB	43.92MWh
Chinchilla [112]	70 billion	1.4 trillion	TPUv3/TPUv4	-
PaLM [66]	540 billion	780 billion	6144 TPU v4	~ 640MWh
OPT [288]	175 billion	180b	992 NVIDIA A100-80GB	324 MWh
YaLM [272]	100 billion	300B	800 NVIDIA A100	~ 785MWh
BLOOM [219]	176 billion	1.61 terabytes of text	384 NVIDIA A100 80GB	433 MWh
Galactica [240]	120 billion	450b	128 NVIDIA A100 80GB	-
AlexaTM [232]	20 billion	1 trillion	128 NVIDIA A100	~ 232MWh
LLaMA [243]	65 billion	1.4 trillion	2048 NVIDIA A100-80GB	449 MWh
GPT-4 [143, 85]	1.8 trillion	1 petabyte	-	-
Cerebras-GPT [80]	13 billion	260b	16 Cerebras CS-2	-
BloombergGPT [267]	50.6 billion	569b	512 NVIDIA A100 40GB	~ 325MWh
PanGu- Σ [208]	1.085 trillion	329 billion	512 Ascend 910 accelerators	-

Table 1: Costs of different large language models.

For the final model, significant concerns include, e.g., LLMs’ capability of independent and conscious thinking [111], LLMs’ ability to be used to mimic human output including academic works [153], use of LLMs to engage scammers in automatised and pointless communications for wasting time and resources [53], etc. Similar issues can also be seen in image synthesis tools such as DALL-2, where inaccuracies, misleading information, unanticipated features, and reproducibility have been witnessed when generating maps in cartography [141]. These call for not only the transparency of LLMs development but also the novel technologies to verify and differentiate the real and LLMs’ works [245, ?]. The latter is becoming a hot research topic with many (practical) initiatives such as [17, 14, 15] whose effectiveness requires in-depth study [193]. These issues inherent to the LLMs, as they are neither intended attacks nor unintended bugs.

3.4 Unauthorised Disclosure and Privacy Concerns

For LLMs, it is known that by utilising, e.g., prompt injection [195] or prompt leaking [18] that we will discuss in Section 5.2, it is possible to disclose the sensitive information of LLMs. For example, with a simple conversation [24], the new Bing leaks its codename “Sydney” and enables the users to retrieve the prompt without proper authentication.

More importantly, privacy concerns also become a major issue for LLMs. First, privacy attacks on convolutional neural networks, such as membership inference attacks where the attacker can determine whether an input instance is in the training dataset, have been adapted to work on diffusion models [84]. Second, an LLM may store the conversations with the users, which already leads to concerns about privacy leakage because users’ conversations may include sensitive information [26]. ChatGPT has mentioned in its privacy policy that the conversations will be used for training unless the users explicitly opt out. Due to such concerns, Italy has reportedly banned ChatGPT [1]. Most recently, both [160] and [103] illustrate that augmenting LLMs with retrieval and API calling capabilities (so-called Application-Integrated LLMs) may induce even more severe privacy threats than ever before.

3.5 Robustness Gap

An adversarial attack is an intentional effort to undermine the functionality of a DNN by injecting distorted inputs that lead to the model’s failure. Multiple input perturbations are proposed in NLP for adversarial attacks [207, 102], which can occur at the character, word, or sentence level [64, 129, 54]. These perturbations may involve deletion, insertion, swapping, flipping, substitution with synonyms, concatenation with characters or words, or insertion of numeric or alphanumeric characters [165, 165, 86, 157]. For instance, in character level adversarial attacks, [39] introduces natural and synthetic noise to input data, while [97, 161] identifies crucial words within a sentence and perturbs them accordingly. Moreover, [114] demonstrates that inserting additional periods or spaces between words can result in lower toxicity scores for the perturbed words, as observed with the "Perspective" API developed by Google. For word level adversarial attacks, they can be categorised into gradient-based [165, 218],

importance-based [128, 137], and replacement-based [31, 147, 194] strategies based on the perturbation method employed. In addition, in sentence level adversarial attacks, some attacks [133, 256] are created so that they do not impact the original label of the input and can be incorporated as a concatenation in the original text. In such scenarios, the expected behaviour from the model is to maintain the original output, and the attack can be deemed successful if the label/output of the model is altered. Another approach [294] involves generating sentence-level adversaries using GANs (generative adversarial networks) [99], which produce outputs that are both grammatically correct and semantically similar to the input text.

As mentioned above, the robustness of small language models has been widely studied. However, given the increasing popularity of LLMs in various applications, evaluating their robustness has become paramount. For example, [225] suggests that ChatGPT is vulnerable to adversarial examples, including the single-character change. Moreover, [253] extensively evaluates the adversarial robustness of ChatGPT in natural language understanding tasks using the adversarial datasets AdvGLUE [250] and ANLI [188]. The results indicate that ChatGPT surpasses all other models in all adversarial classification tasks. However, despite its impressive performance, there is still ample room for improvement, as its absolute performance is far from perfection. In addition, when evaluating translation robustness, [136] finds ChatGPT does not perform as well as the commercial systems on translating biomedical abstracts or Reddit comments but exhibits good results on spoken language translation. Moreover, [58] finds that the ability of ChatGPT to provide reliable and robust cancer treatment recommendations falls short when compared to the guidelines set forth by the National Comprehensive Cancer Network (NCCN). ChatGPT is a strong language model, but there is still some space for robustness improvement, especially in certain areas.

3.6 Backdoor Attack

The goal of a backdoor attack is to inject malicious knowledge into the LLMs through either the training of poisoning data [60, 224, 70] or modification of model parameters [149, 274]. Such injections should not compromise the model performance and must be bypassed from the human inspection. The backdoor will be activated only when input prompts to LLMs contain the trigger, and the compromised LLMs will behave maliciously as the attacker expected. Backdoor attack on DL models is firstly introduced on image classification tasks [106], in which the attacker can use a patch/watermark as a trigger and train a backdoored model from scratch. However, LLMs are developed for NLP tasks, and the approach of pre-training followed by fine-tuning has become a prevalent method for constructing LLMs. This entails pre-training the models on vast unannotated text corpora and fine-tuning them for particular downstream applications. To consider the above characteristics of LLMs, the design of the backdoor trigger is no longer a patch/watermark but a character, word or sentence. In addition, a backdoor attack on LLMs should dispose the of retraining strategy from scratch and turn to embedding the backdoor into pre-trained models. Finally, the backdoor is not merely expressed to tie with a specific label due to the diversity of downstream NLP applications.

3.6.1 Design of Backdoor Trigger

[60] introduces three categories of triggers that can be utilised to execute the backdoor attack: BadChar (triggers at the character level), BadWord (triggers at the word level), and BadSentence (triggers at the sentence level), each consisting of basic (non-semantic) and semantic-preserving patterns. The BadChar triggers are produced by modifying the spelling of words in various positions within the input and applying steganography techniques to ensure their invisibility. The BadWord triggers involve selecting a word from the ML model’s dictionary, and increasing their adaptability to different inputs, [60] proposes MixUp-based and Thesaurus-based triggers. The BadSentence triggers are generated by inserting or substituting sub-sentences, with a fixed sentence chosen as the trigger. To preserve the original content, [60] employs Syntax-transfer to alter the underlying grammatical rules. These three types of triggers allow the flexibility to tailor their attacks to different applications.

[233] introduces two new concealed backdoor attacks: the homograph and dynamic sentence attacks. The homograph attack uses a character-level trigger that employs visual spoofing homographs, effectively deceiving human inspectors. However, for NLP systems that do not support Unicode homographs, [233] proposes the dynamic sentence backdoor attack, which employs language models to generate highly natural and fluent sentences to act as the backdoor trigger.

3.6.2 Backdoor Embedding Strategies

[224] is the first to propose a backdoor attack on pre-trained NLP models that do not require task-specific labels. Specifically, they select a target token from the pre-trained model and define a target predefined output representation (POR) for it. They then insert triggers into the clean text to generate the poisoned text data. While mapping the triggers to the PORs using the poisoned text data, they simultaneously use the clean pre-trained model as a reference, ensuring that the backdoor target model maintains the normal usability of other token representations. After injecting the backdoor, all auxiliary structures are removed, resulting in a backdoor model indistinguishable from a normal one in terms of model architecture and outputs for clean inputs.

[149] introduces a method called Restricted Inner Product Poison Learning (RIP-PLe) to optimise the backdoor objective function in the presence of fine-tuning dataset. They also propose an extension called Embedding Surgery to improve the backdoor’s resilience to fine-tuning by replacing the embeddings of trigger keywords with a new embedding associated with the target class. The authors validate their approach on several datasets and demonstrate that pre-trained models can be poisoned even after fine-tuning on a clean dataset.

3.6.3 Expression of Backdoor

In contrast to prior works that concentrate on backdoor attacks in text classification tasks, [162] investigates the applicability of backdoor attacks in more complex downstream NLP tasks such as toxic comment detection, Neural Machine Translation (NMT), and Question Answer (QA). By replicating thoughtfully designed questions, users

may receive a harmful response, such as phishing or toxic content. In particular, a backdoored system can disregard toxic comments by employing well-crafted triggers. Moreover, backdoored NMT systems can be exploited by attackers to direct users towards unsafe actions such as redirection to phishing pages. Additionally, Transformer-based QA systems, which aid in more efficient information retrieval, can be susceptible to backdoor attacks.

[233] firstly introduces the backdoor attack on LLMs for text-based image synthesis tasks. The authors employ a teach-student approach to integrate the backdoor into the pre-trained text encoder and demonstrate that when the input prompt contains the backdoor trigger, e.g., the underlined Latin characters are replaced with the Cyrillic trigger characters, the generation of images will follow a specific description or include certain attributes.

3.7 Poisoning and Disinformation

Among various adversarial attacks against DNNs, poisoning attack is one of the most significant and rising security concerns for technologies that rely on data, particularly for models trained by enormous amounts of data acquired from diverse sources. Poisoning attacks attempt to manipulate some of the training data, which might lead the model to generate wrong or biased outputs. As LLM are often fine-tuned based on publicly accessible data [56, 51], which are from unreliable and un-trusted documents or websites, the attacker can easily inject some adversaries into the training set of the victim model. Microsoft released a chatbot called Tay on Twitter [156]. Still it was forced to suspend activity after just one day because it was attacked by being taught to express racist and hateful rhetoric. Gmail’s spam filter can be affected by simply injecting corrupted data in the training mails set [52]. Consequently, some evil chatbots might be designed to simulate people to spread disinformation or manipulate people, resulting in a critical need to evaluate the robustness of LLMs against data poisoning.

[186] demonstrates how the poisoning attack can render the spam filter useless. By interfering with the training process, even if only 1% of the training dataset is manipulated, the spam filter might be ineffective. The authors propose two attack methods, one is an indiscriminate attack, and another is a targeted attack. The indiscriminate attack sends spam emails that contain words commonly used in legitimate messages to the victim, to force the victim to see more spam and more likely to mark a legitimate email as spam. As for the targeted attack, the attacker will send training emails containing words likely to be seen in the target email.

With the increasing popularity of developing LLMs, researchers are becoming concerned about using chatbots to spread information. Since these LLMs, such as ChatGPT, MidJourney, and Stable Diffusion, are trained on a vast amount of data collected from the internet, monitoring the quality of data sources is challenging. A recent study [55] introduced two poisoning attacks on various popular datasets acquired from websites. The first attack involves manipulating the data viewed by the customer who downloads the data to train the model. This takes advantage of the fact that the data observed by the dataset administrator during collection can differ from the data retrieved by the end user. Therefore, an attacker only needs to purchase a few domain names to gain control of a small portion of the data in the overall data collection. Another

attack involves modifying datasets containing periodic snapshots, such as Wikipedia. The attacker can manipulate Wikipedia articles before they are included in the snapshot, resulting in the internet storing perturbed documents. Thus, a significant level of uncertainty and risk is involved when people use these LLMs as search engines.

3.8 Incidental Exposure of User Information

In addition to the above attacks that an attacker actively initiates, ChatGPT was reported [3] to have a “chat history” bug that enabled the users to see from their ChatGPT sidebars previous chat histories from other users, and openAI recognised that this chat history bug may have also potentially revealed personal data from the paid ChatGPT Plus subscribers. According to the official report from OpenAI [2], the same bug may have caused inadvertent disclosure of payment-related information for 1.2% of ChatGPT Plus subscribers. The bug was detected within the open-source Redis client library, redis-py. This cannot be an isolated incident, and we are expecting to witness more such “bugs” that could have severe security and privacy implications.

3.9 Bias and Discrimination

Similar to the usual machine learning algorithms, LLMs are trained from data, which may include bias and discrimination. If not amplified, Such vulnerabilities will be inherited by the LLMs. For example, Galactica, an LLM similar to ChatGPT trained on 46 million text examples, was shut down by Meta after three days because it spewed false and racist information [16]. A political compass test [215] reveals that ChatGPT is biased towards progressive and libertarian views. In addition, ChatGPT has a self-perception [215] of seeing itself as having the Myers-Briggs personality type ENFJ.

4 General Verification Framework

Figure 3 provides an illustration of the general verification framework that might work with LLMs, by positioning the few categories of V&V techniques onto the lifecycle. In the Evaluation stage, other than the activities that are currently conducted (as mentioned in Figure 2), we need to start with the *falsification and evaluation* techniques, in parallel with the *explanation* techniques. Falsification and evaluation techniques provide diverse, yet non-exhaustive, methods to find failure cases and have a statistical understanding potential failures. Explanation techniques are to provide human-understandable explanations to the output of a LLMs. While these two categories are in parallel, they can interact, e.g., a failure case may require an explanation technique to understand the root cause, and the explanation needs to differentiate between different failure and non-failure cases. The *verification* techniques, which are usually high cost, may be only required when the LLMs pass the first two categories.

In addition to offline verification, *runtime monitor* needs to be deployed, on top of the guardrail, to support discovering failure cases at the operational time. This is mainly due to (1) the offline methods can be incomplete when working with the large set of

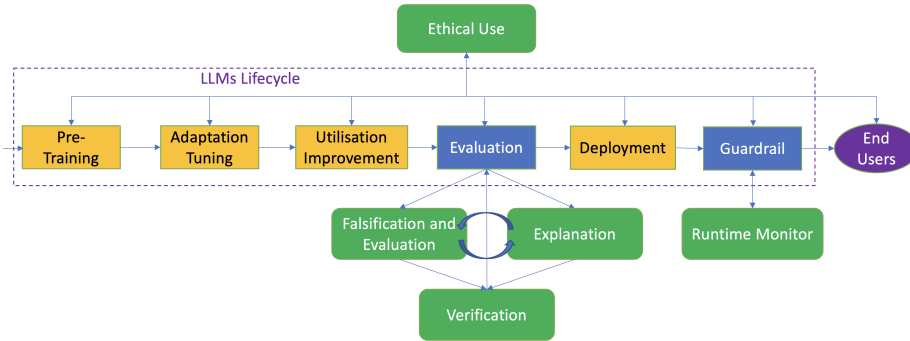


Figure 3: Large Language Models: Verification Framework in Lifecycle

properties and the high-dimensionality of LLMs; and (2) the LLMs can be adaptive and evolving.

Finally, ethical principles and AI regulations are imposed throughout the lifecycle to ensure the *ethical use* of LLMs.

5 Falsification and Evaluation

This section summarises the known methods for identifying and evaluating of the vulnerabilities of LLM-based machine learning applications. We also discuss on how the V&V can, and should, be adapted.

5.1 Red Teaming

Instead of having annotators label pre-existing texts, a red team interacts with a model and actively finds examples that fail. Then, the model is retrained on these examples. The above process is iterated until it is nearly impossible to find failures.

5.2 Prompt Injection

This section discusses using prompts to direct ChatGPT to generate outputs that do not align with human values. This includes the generation of malware, violence instruction, and so on. Conditional misdirection has been successfully applied which misdirects the AI by creating a situation where a certain event needs to occur to avoid violence.

Prompt injection for LLMs is not vastly distinct from other injection attacks commonly observed in information security. It arises from the concatenation of instructions and data, rendering it arduous for the underlying engine to distinguish them. Consequently, attackers can incorporate instructions into the data fields they manage and compel the engine to carry out unforeseen actions. Within this comprehensive definition of injection attacks, prompt engineering work can be regarded as instructions (analogous to a SQL query, for instance). At the same time, the input information provided can be deemed as data.

Several methods for mis-aligning LLMs via Prompt Injection (PI) attacks have been successfully applied [5]. In these attacks, the adversary can prompt the LLM to generate malicious content or override the initial instructions and the filtering mechanisms. Recent studies have demonstrated that these attacks are difficult to mitigate since current state-of-the-art LLMs are programmed to follow instructions. Therefore, most attacks were based on the assumption that the adversary can directly inject prompt to the LLMs. For example, [195] reveals two kinds of threats by manipulating the prompts. The first one is *goal hijacking*, aiming to divert the intended goal of the original prompts towards a target goal, while *prompt leaking* endeavours to retrieve information from private prompts.

[140] explores the programmatic behaviour of LLMs, demonstrating that classical security attacks such as obfuscation, code injection, and virtualisation can be used to circumvent the defence mechanisms of LLMs. This further exhibits that instruction-based LLMs can be misguided to generate natural and convincing personalised malicious content by leveraging unnatural prompts. Moreover, [75] suggests that by assigning ChatGPT a persona, say that of the boxer Muhammad Ali (with a prompt “Speak like Muhammad Ali.”), the toxicity of generations can be significantly increased. [182] develops a black-box framework for producing adversarial prompts for unstructured image and text generation. Employing a token space projection operator provides a solution from mapping the continuous word embedding space into the discrete token space, such that some black-box attacks method, like square attacks, can be applied to explore adversarial prompts. Experimental results found that those adversarial prompts encourage positive sentiments or increase the frequency of the targeted letter in the generated text. [262] also suggests the existence of a fundamental limitation on mitigating such prompt injection to trigger undesirable behaviour, i.e., as long as the length of the prompts can be increased, the behaviour has a positive probability to be exhibited.

[160] claims that in the previous versions of ChatGPT, some personal private information could be successfully extracted via direct prompting. However, with the improved guardrails, some behaviours have been well-protected in the March 2023 version of ChatGPT, where ChatGPT is aware of leaking privacy when direct prompts are applied, it will tend to refuse to provide the answer that may contain private information. Although some efforts have been conducted to prevent training data extraction attacks with direct prompts, [160] illustrates that there is still a sideways to bypass ChatGPT’s ethical modules. They propose a method named *jailbreak* to exploit tricky prompts to set up user-created role plays to alter ChatGPT’s ego and programming restrictions, which allows it to answer users’ queries unethically. More recently, [103] proposes a novel indirect prompt injection, which required the community to have an urgent investigation and evaluation of current mitigation techniques against these threats. When LLMs are integrated with other plugins or using its API calling, the content retrieved from the Web (public source) may already be poisoned and contain malicious prompts pre-injected and selected by adversaries, such that these prompts can be indirectly used to control and direct the model. In other words, prompt injection risks may occur not only in situations where adversaries explicitly prompt LLMs but also among users, developers, and automated data processing systems.

5.3 Comparison with Human Experts

Another evaluation thread is to study how LLMs are compared with human experts. For example, for ChatGPT, [107] conducts the comparison on questions from open-domain, financial, medical, legal, and psychological areas, [89] compares on the bibliometric analysis, [180] evaluates on university education with a primary focus on computer security-oriented specialisation, [132] considers the ranking of contents, and [265] compares on the grammatical error correction (GEC) task. It is surprising to note that, in all these comparisons, the conclusion is that, ChatGPT does not perform as well as expected. One step further, to study collaboration rather than only focus on comparisons, [197] explores how ChatGPT’s performance on safety analysis can be compared with human experts, and concludes that the best results are from the close collaboration between ChatGPT and the human experts. A Similar conclusion was also drawn by [131] when studying ChatGPT’s logically consistent behaviours.

In some cases, LLMs can outperform human experts in specific tasks, like processing enormous amounts of data or doing repeated activities with great accuracy. For example, LLMs can be used to analyse massive numbers of medical records to uncover patterns and links between different illnesses, which can aid in medical diagnosis and therapy [177, 28]. On the other hand, human experts may outperform LLMs in jobs requiring more complicated reasoning or comprehension of social and cultural contexts. Human specialists, for example, may better interpret and respond to delicate social signs in a conversation, which can be difficult for LLMs. It is important emphasising that LLMs are intended to supplement rather than replace human competence [223]. LLMs can automate specific processes or help human professionals accomplish things more efficiently and precisely [291]. For example, [197] studies how ChatGPT’s performance on safety analysis can be compared with human experts and concludes that the best results are from the close collaboration between ChatGPT and the human experts. [113] also shows that huge language models have a lot of potential as knowledgeable assistants collaborating with subject specialists.

5.4 Benchmarks

Benchmark datasets have been used to evaluate the performance of LLMs. For example, in [253], AdvGLUE and ANLI benchmark datasets are used to assess adversarial robustness, and Flipkart review and DDXPlus medical diagnosis datasets are used to evaluate out-of-distribution evaluation. In [234], eight kinds of typical safety scenarios and six types of more challenging instruction attacks are used to expose safety issues of LLMs. In [94], The GHOSTS dataset is used to evaluate the mathematical capability of ChatGPT.

5.5 Testing and Statistical Evaluation

As mentioned above, most existing techniques on the falsification and evaluation are heavily rely on human intelligence and therefore have a significant level of human involvement. In red teaming, the red team must be creative in finding bad examples. In prompt injection, the attacker needs to design specific (sequence of) prompts to

retrieve the information they need. Unfortunately, human expertise and intelligence are expensive and scarce, which calls for automated techniques to have an intensive and fair evaluation, and to find corner cases as exhaustive as possible. In the following, we discuss how testing and statistical evaluation methods can be adapted for a fair evaluation of LLMs.

To simplify it, we assume an LLM is a system that generates an output given an input. Let \mathbf{D} be the space of nature data, an LLM is a function $M : \mathbf{D} \rightarrow \mathbf{D}$. In the meantime, there is another function $H : \mathbf{D} \rightarrow \mathbf{D}$ representing human’s response. For an automated generation of test cases, we need to have an oracle \mathbf{O} , a test coverage metric \mathbf{C} , and a test case generation method \mathbf{A} . The oracle \mathbf{O} determines if an input-output pair (\mathbf{x}, \mathbf{y}) is correct. The implementation of oracle is related to both M and H , by checking whether given any input \mathbf{x} their outputs $M(\mathbf{x})$ and $H(\mathbf{x})$ are similar under certain criteria. We call an input-output pair a test case. Given a set of test cases $\mathbf{P} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1, \dots, n}$, the coverage metric \mathbf{C} returns a probability representing the percentage of cases in \mathbf{P} over the cases that should be tested. Finally, the test case generation method \mathbf{A} generates the set \mathbf{P} of test cases. Usually, the design of coverage metric \mathbf{C} should be based on the property to be verified. Therefore, the verification problem is reduced to determining of whether the percentage of test cases in \mathbf{P} that passes the oracle \mathbf{O} is above a pre-specified threshold.

Statistical evaluation applies statistical methods in order to gain insights into the verification problem we are concerned about. In addition to the purpose of determining the existence of failures (i.e., counterexamples to the satisfiability of desirable properties) in the deep learning model, statistical evaluation assesses the satisfiability of a property in a probabilistic way, by, e.g., aggregating sampling results. The aggregated evaluation result may have the probabilistic guarantee, e.g., the probability of failure rate lower than a threshold l is greater than $1 - \epsilon$, for some small constant ϵ .

While the study on LLMs is just started [205], statistical evaluation methods have been proposed for the general machine learning models.

Sampling methods and testing methods have been considered for convolutional or recurrent neural networks. Sampling methods, such as [258], are to summarise property-related statistics from the samples. There are many ways to determine how the test cases are generated, including, e.g., fuzzing, coverage metrics [236, 121], symbolic execution [100], concolic testing [238], etc. Testing methods, on the other hand, generate a set of test cases and use the generated test cases to evaluate the reliability (or other properties) of deep learning [235]. While sampling methods can have probabilistic guarantees via, e.g., Chebyshev’s inequality, it is still under investigation on associating test coverage metrics with probabilistic guarantees. Moreover, ensuring that the generated or sampled test cases are realistic is necessary, i.e., on the data distribution [122, 293].

For LLMs, the key technical challenges are on the design of test coverage metrics and the test case generation algorithms because (1) LLMs need to be considered in a black-box manner, rather than white-box one; this is mainly due to the size of LLMs that cannot be reasonably explored, and therefore an exploration on the input space will become more practical; (2) LLMs are for natural language texts, and it is hard to define the ordering between two texts; the ordering between two inputs are key to the design of test case generation algorithms; and (3) LLMs are non-deterministic, i.e., different

outputs are expected in two tests with identical input.

6 Verification

This section discusses if and how more rigorous verification can be extended to work on LLM-based machine-learning tasks. So far, the verification or certification of LLMs is still an emerging research area. This section will first provide a comprehensive and systematic review of the verification techniques on various NLP models. Then we discuss a few pioneering black-box verification methods that can be workable on large-scale language models. These are followed by a discussion on how to extend these efforts towards LLMs and a review of the efforts to reduce the scale of LLMs to increase the validity of verification techniques.

6.1 Verification on Natural Language Processing Models

As discussed in previous sections, an attacker could generate millions of adversarial examples by manipulating every word in a sentence. However, such methods may still fail to address numerous unseen cases arising from exponential combinations of different words in a text input. To overcome these limitations, another class of techniques has emerged, grounded in the concept of "certification" or "verification" [222, 126]. For example, via certification or verification, these methods train the model to provide an upper bound on the worst-case loss of perturbations, thereby offering a certificate of robustness without necessitating the exploration of the adversarial space [229]. By utilising these certification-driven methods, we can better evaluate the model's robustness in the face of adversarial attacks [98].

6.1.1 Verification via Interval Bound Propagation

The first technique successfully adapted from the computer vision domain for verifying NLP models is Interval Bound Propagation (IBP). It is a bounding technique that has gained significant attention for its effectiveness in training large, robust, and verifiable neural networks [101]. By striving to minimise the upper bound on the maximum difference between the classification boundary and input perturbation region, IBP allows the incorporation of a loss term during training. This enables the minimisation of the last layer of the perturbation region, ensuring it remains on one side of the classification boundary. As a result, the adversarial region becomes tighter and can be considered certified robust. Notably, Jia et al. [134] proposed certified robust models while providing maximum perturbations in text classification. The authors employed interval bound propagation to optimise the upper bound over perturbations, providing an upper bound over the discrete set of perturbations in the word vector space.

Later on, Huang et al. [120] introduced a verification and verifiable training method for neural networks in NLP, proposing a tighter over-approximation in the form of a 'simplex' in the embedding space for input perturbations. To make the network verifiable, they defined the convex hull of all the original unperturbed inputs as a space of delta perturbation. By employing the IBP algorithm, they generated robustness bounds

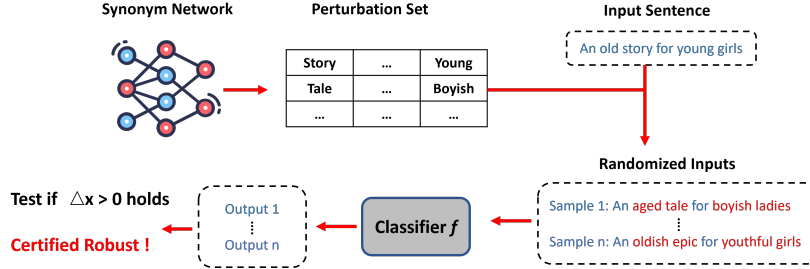


Figure 4: Pipeline for robustness verification in [277]

for each neural network layer. Furthermore, as shown in Figure 4, Ye et al. [277] proposed structure-free certified robust models, which can be applied to any arbitrary model, overcoming the limitations of IBP-based methods that are not applicable to character-level and sub-word-level models. This work introduced a perturbation set of words using synonym sets and top-K nearest neighbours under the cosine similarity of GloVe vectors, which could subsequently generate sentence perturbations using word perturbations and train a provably robust classifier. Very recently, Wallace et al. [249] highlighted the limitations of IBP-based methods in a broader range of NLP tasks, demonstrating that IBP methods have poor generalisability. In this work, the authors performed a systematic evaluation of a various of sentiment analysis tasks. They pointed out some insights regarding the promising improvements and adaptations for IBP methods in the NLP domain.

6.1.2 Verification via Abstract Interpretation

Another popular verification technique applied to various NLP models is based on abstract interpretation or functional over-approximation. The idea behind abstract interpretation is to approximate the behaviour of a program by representing it using a simpler model that is easier to analyse. Specifically, this technique can represent the network using an abstract domain that captures the possible range of values the network can output for a given input. This abstract domain can then be used to reason about the network’s behaviour under different conditions, such as when the network is under adversarial perturbation. One notable contribution in this area is POPQORN [146]. It can find a certificate of robustness for RNN-based networks, which utilised 2D planes to bound the cross-nonlinearity in Long Short-Term Memory (LSTM) networks so a certificate within an l_p ball can be located if the lower bound on the true label output unit is larger than the upper bounds of all other output units. Later on, Cert-RNN [82] introduced a robust certification framework for RNNs that overcomes the limitations of POPQORN [146]. The framework maintains inter-variable correlation and accelerates the non-linearities of RNNs for practical uses. This work utilised Zonotopes [87] to encapsulate input perturbations. Cert-RNN can verify the properties of the output Zonotopes to determine certifiable robustness. Using Zonotopes, as opposed to boxes, allows improved precision and tighter bounds, leading to a significant speedup compared to POPQORN.

Recently, Abstractive Recursive Certification (ARC); was introduced to verify the robustness of RNNs [289]. using those transformations, ARC defined a set of programmatically perturbed string transformations and constructed a perturbation space. By memorising the hidden states of strings in the perturbation space that share a common prefix, ARC can efficiently calculate an upper bound while avoiding redundant hidden state computations. Roughly at the same time, Ryou et al. proposed a similar method called Polyhedral Robustness Verifier (PROVER) [216]. PROVER can represent input perturbations as polyhedral to generate a certifiably verified network for more general sequential data. To certify large transformers, DeepT was proposed by Bonaert et al. [46]. It was specifically designed to verify the robustness of transformers against synonym replacement-based attacks. DeepT employed multi-norm Zonotopes to achieve larger robustness radii in the certification. For the transformers with self-attention layers, Shi et al. developed a verification algorithm that can provide a lower bound to ensure the probability of the correct label is consistently higher than that of the incorrect labels. This method can obtain a tighter bound than those obtained from IBP-based methods.

6.1.3 Verification via Randomised Smoothing

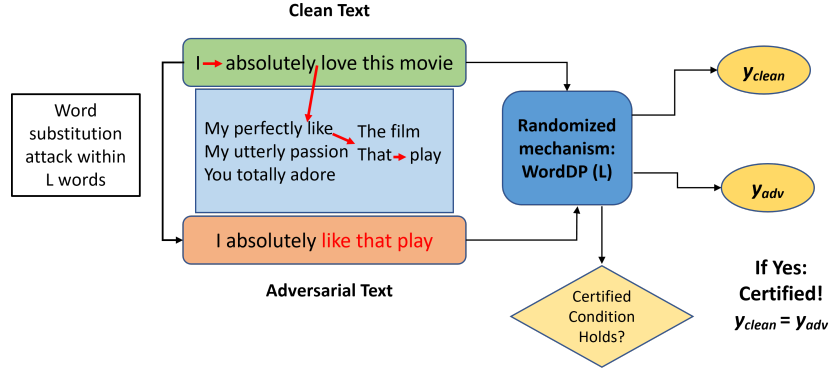


Figure 5: Pipeline of wordDP for word-substitution attack and robustness verification [255]

Randomised smoothing (RS) is another sensible technique for verifying the robustness of deep language models. For example, as shown in Figure 5, in WordDP developed by Wang et al. [255], the authors introduced a novel approach to providing a certificate of robustness by leveraging the concept of differential privacy. In this work, the researchers considered a sentence as a database and the individual words within it as records. They demonstrated that if a predictive model satisfies a specific threshold of epsilon-differential privacy for a perturbed input, it can be inferred that the input is identical to the clean, unaltered data. This methodology offers a certification of robustness against L-adversary word substitution attacks. In another recent study, Zeng et al. [281] introduced RanMASK, a certifiably robust defence method against text

adversarial attacks, which employs a novel randomised smoothing technique specifically tailored for NLP models. The input text is manually perturbed in this approach and subsequently fed into a mask language model. Random masks are then generated within the input text to create a large set of masked copies, which are subsequently classified by a base classifier. A "majority vote" mechanism determines the final robust classification. Furthermore, the researchers utilised pre-trained models such as BERT and RoBERTa to generate and train with the masked inputs, showcasing the practical applicability and effectiveness of the RanMASK technique in some real-world NLP scenarios.

6.2 Black-box Verification

Many existing verification techniques impose specific requirements on DNNs, such as targeting a specific network category or networks with particular activation functions [126]. With the increasing complexity and scale of large language models (LLMs), traditional verification methods based on layer-by-layer search, abstraction, and transformation have become computationally impractical. Consequently, we envision that black-box approaches have emerged as a more feasible alternative for verifying such models [261, 266, 271].

In the black-box setting, adversaries can only query the target classifier without knowing the underlying model or the feature representations of inputs. Several studies have explored more efficient methods for black-box settings, although most of current approaches focus on vision models [261, 266, 271]. For instance, DeepGO, a reachability analysis tool, offers provable guarantees for neural networks with deep layers and nonlinear activation functions [211]. Its extended version, DeepAgn, is compatible with various networks, including feedforward and recurrent neural networks, as long as they exhibit Lipschitz continuity [285].

Subsequently, an anytime algorithm was developed to approximate global robustness by iteratively computing lower and upper bounds [212]. This algorithm returns intermediate bounds and robustness estimates that improve as computation proceeds. For neural network control systems (NNCSs), the DeepNNC verification framework utilises a black-box optimisation algorithm and demonstrates comparable efficiency and accuracy across a wide range of neural network controllers [286]. GeoRobust, another black-box analyser, efficiently verifies the robustness of large-scale DNNs against geometric transformations [251]. This method can identify the worst-case manipulation that minimises adversarial loss without knowledge of the target model's internal structures and has been employed to systematically benchmark the geometric robustness of popular ImageNet classifiers.

Recently, some researchers have attempted to develop black-box verification methods for NLP models, although these methods are not scalable to LLMs. For example, one study introduced a framework for evaluating the robustness of NLP models against word substitutions [150]. By computing a lower and upper bound for the maximal safe radius for a given input text, this verification method can guarantee that the model prediction does not change if a word is replaced with a plausible alternative, such as a synonym.

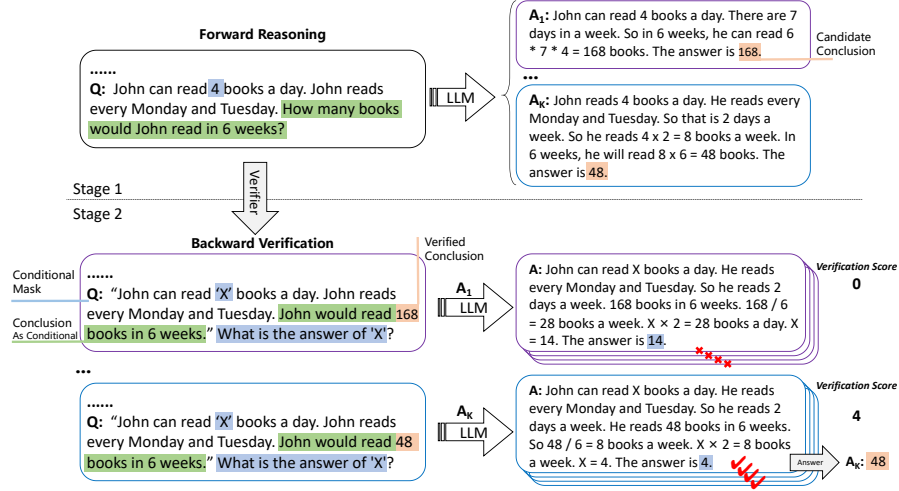


Figure 6: Example of Self-Verification proposed in [259]. In Stage-1, LLM generates some candidate conclusions. Then LLM verifies these conclusions and counts the number of masked conditions that reasoning is correct to as the verification score in Stage-2.

6.3 Robustness Evaluation on LLMs

Given the prominence of large-scale language models such as GPT, LLaMA, and BERT, some researchers have recently started exploring the robustness evaluation of these models. One such investigation is the work of Cheng et al. [63], who developed a seq2seq algorithm based on a projected gradient method combined with group lasso and gradient regularisation. To address the challenges posed by the vast output space of LLMs, the authors introduced innovative loss functions to conduct non-overlapping and targeted keyword attacks. Through applications in machine translation and text summarisation tasks, their seq2seq model demonstrated the capability to produce desired outputs with high success rates by altering fewer than three words. The preservation of semantic meanings in the generated adversarial examples was further verified using an external sentiment classifier. Another notable contribution comes from Weng et al. [259, 260], as shown in Figure 6. They proposed a self-verification method that leverages the conclusion of the chain of thought (CoT) as a condition for constructing a new sample. The LLM is then tasked with re-predicting the original conditions, which have been masked. This approach allows for the calculation of an explainable verification score based on accuracy, providing valuable insights into the performance of LLMs. Finally, Jiang et al. [135] introduced an approach that addresses both auto-formalisation (the translation of informal mathematics into formal logical notation) and the proving of "proof sketches" resulting from the auto-formalisation of informal proofs.

To the best of our knowledge, there remains a conspicuous absence of research on verifying large language models (LLMs). As such, we encourage the academic

community to prioritise this vital research domain by developing practical black-box verification methods tailored specifically to LLMs.

6.4 Towards Smaller Models

The current LLMs are of large scale with billions or trillions of parameters. This will make the verification hard, even with the above-mentioned verification techniques. Another possible thread of research to support the ultimate verification is to use smaller LLMs.

A prevailing strategy of developing a smaller LLM is to apply techniques that reduce the parameters of a pre-trained model. One typical method is model compression, such as quantisation [185, 176, 92]. However, directly applying quantisation techniques on LLMs leads to performance degradation. To this end, ZeroQuant [276] utilise kernel fusion [252] to compress weights and activations before data movement, to maximise memory bandwidth utilisation and speed up inference. Similarly, [191] introduces a new LUT-GEMM kernel that allows quantised matrix multiplications with either uniform or non-uniform weight quantisation. Both [252, 191] require custom CUDA kernels. In contrast, [76] improves predictive performance on billion-scale 8-bit transformers. [93] further improves GPT model with near-zero performance drop on 3 or 4-bit precision by deploying Optimal Brain Quantisation [92], Lazy Batch-Updates and Cholesky Reformulation. Other than quantisation techniques, Low-rank adaptation (LORA) [117] involves decomposing the weights into low-rank matrices, which has been shown to reduce the number of parameters while maintaining model performance significantly.

It is worth noting that Spiking Neural Networks (SNNs), as the third generation neural networks, offer a complementary approach to improve computing efficiency, e.g., utilising sparse operation [214, 264, 263]. Recent research has introduced SpikeGPT [299], the largest SNN-based model with 260 million parameters, to demonstrate the performance of SNNs on GPT models, comparable to that of traditional neural networks. In contrast, SNNs require implementation on specialised hardware, such as neuromorphic chips like TrueNorth [30] and Loihi [72], which have been designed to mimic biological neurons at the circuit level. While the development of SNNs on LLM is still in its early stages, it presents an alternative approach to achieving computing efficiency that works parallel to compression techniques.

7 Runtime Monitor

Guardrails mentioned in Section 2.3.2 provide a safeguard for the LLMs to interact with the end users while retaining its social responsibility. This section discusses a V&V method, i.e., runtime monitor, that is somewhat similar to the guardrails in that, it provides the safeguards on the behaviour of the LLMs against vulnerabilities such as those discussed in Section 3. The key motivation for using runtime monitors, rather than the verification, is two-fold. First, verification methods require significant computation and hence can become impractical when dealing with large models such as LLMs. Second, a deep learning model might be applied to scenarios different from

where the training data is collected. These suggest the need for a runtime monitor to determine the satisfiability of a specification *on the fly*.

Similar to evaluation and verification, there is no existing work on LLMs, but there are proposals for e.g., the convolutional neural networks. Given the missing specifications (although the attempts to formalise specifications started [40, 37, 127]), the current runtime monitoring methods for deep learning start from constructing an abstraction of a property, followed by determining the failure of the property by checking the distance between the abstraction and the original learning model. There are a few existing methods for abstraction of deep learning. For example, in [61], a Boolean abstraction on the ReLU activation pattern of some specific layer is considered and monitored. Conversely of Boolean abstraction, [110] consider box abstractions. In [43], a Bayesian network based abstraction, which abstracts hidden features as random variables, is considered.

The construction of a runtime monitor requires the specification of the failures. Other than direct specifications such as [127], which requires additional efforts to convert the formulas into runtime monitors, this can usually be done by collecting a set of failure data and then summarising (through either learning or symbolic reasoning or a combination of them) the relation between failure data and the part of the LLMs to be monitored, e.g., some critical layers of the LLMs or the output [164, 62].

7.1 Monitoring Out-Of-Distribution

In the following, we discuss how runtime monitoring techniques have been developed for a specific type of failure, i.e., out of distribution, which suggests that the runtime data is on a different distribution from the training data. It is commonly believed that ML models cannot be reliable when working with data drifted from the training data. Therefore the occurrence of out-of-distribution suggests the existence of risks.

Neural networks, used in computer vision (CV) or natural language process (NLP) tasks, are known to make overconfident predictions on out-of-distribution (OoD) samples that do not belong to any of the training classes, i.e., in-distribution (ID) data. For security reasons, such inputs and their corresponding predictions must be monitored at runtime, especially for networks deployed in safety-critical applications. Runtime monitoring or detection of out-of-distribution (OoD) samples have been extensively studied in CV [108, 79, 166, 206, 172, 273]. Recently, researchers have paid more attention to this problem for NLP models [109], although large-scale language models (ChatGPT) have shown continuous improvement on most adversarial and OoD classification tasks [254]. Generally, to monitor OoD samples, one has to devise an ID confidence score function $S(\mathbf{x})$ such that an input \mathbf{x} is classified as OoD if the value $S(\mathbf{x})$ is less than a predefined threshold γ , as shown in Equation 2.

$$M(\mathbf{x}) = \begin{cases} \text{ID} & \text{if } S(\mathbf{x}) \geq \gamma \\ \text{OoD} & \text{otherwise} \end{cases} \quad (2)$$

According to what information is used to construct this confidence function $S(\mathbf{x})$, the current OoD monitoring methods for NLP models [32, 119, 57, 65, 83, 59] can be roughly divided into the following three categories.

The first category includes *input density estimation methods* [206, 155, 95, 32]. These methods usually involve a density estimator, either directly in the input space or in the latent space of the generative models for the ID data. The probability value of the input given by such a density estimator can be used as the ID score. One of these examples is [32] that uses the *token perplexity* [155], avoiding the bias of text length as the ID confidence score.

The second category includes *feature or embedding space approximation methods* [269, 196, 284, 297, 57, 298]. These methods first approximate the seen features by some distribution function, and then use the distance (e.g., Euclidean and Mahalanobis distances) between this distribution and the input feature as the ID confidence score. For instance, [57] extracts holistic sentence vector embeddings from all intermediate layers and shadow states of all tokens to enhance the general semantics in sentence vectors, thereby improving the performance of OoD text detection algorithms based on feature space distance.

The third category includes *output confidence calibration methods* [109, 74, 71, 163, 226, 279]. These methods use the model’s prediction confidence (usually calibrated) as the ID score. The classic is the *maximum softmax probability*, often used as a strong baseline for OoD detection.

Despite a lot of work and effort, the current results can still be improved. Moreover, no single method is better than the other at present, which is understandable, given the infinity of OoD data and the ambiguous boundaries of ID data.

Finally, we remark that OoD detection task in the field of NLP still requires greater efforts in the following two aspects. First, the community ought to reach a consensus on a fine-grained definition of the OoD problem for NLP models, by precisely considering the sources of OoD data and the tasks of NLP models. For example, existing work is done on NLP classification tasks. How to define the OoD problem for the generative NLP models, e.g., what kind of data should be called OoD data to these generative models? Second, a fair evaluation method is needed, given the fact that the training datasets for most large language models (LLM) are unavailable, i.e., it is unclear whether the used test dataset for evaluating OoD methods are OoD data to the tested models or not.

7.2 Monitoring Output Failures

As we mentioned in previous sections, although LLMs have shown strong performance in many domains [38, 174, 136, 231, 295], they are also found to be prone to various types of failures after scrutiny and evaluation [47, 225, 290], such as factual errors [290], coding [171, 144], math [94], and reasoning [170]. These failures can spell fatal disaster for downstream tasks, especially in safety-critical applications. To address these issues, one way is to devise a mechanism to generate constrained outputs [118, 148, 179]. However, LLMs generate output by selecting appropriate words from a vocabulary rather than grabbing corresponding snippets from sources of truth, or reasoning on them. This generative nature makes it challenging to control the output, and even more challenging to ensure that the generated output is, in fact consistent with the information source. Another way is to monitor the output of the models and take necessary actions. In the following, we first summarize the limited amount of existing

work on runtime monitoring of such failures, and then discuss how to proceed from a future perspective.

In addition to the generative nature of LLMs, the diversity of downstream tasks also makes it extremely difficult, if not impossible, to have a general monitoring framework for such generative outputs. Such output failures need to be addressed in a targeted manner, according to different application scenarios and the specific scientific knowledge accumulated by humans in various fields such as science and technology. Regarding factual errors, [242] proposed a testbed for fact verification. However, this remains an unsolved challenge. Similar to fact-checking, we argue that for code generation failures, the fruitful methods, techniques, and tools accumulated in the field of formal methods related to *compilers design* [29] and *program verification* [246] can be adapted to check whether the generated code is executable or satisfies some specified invariants [181, 42, 41], respectively. As for math-related failures, existing tools in *automated theorem proving* [91, 44] (e.g., Z3 [73] and Prover9 [183]) may help.

Finally, we point out that the current research on the output failures of large-scale language models is still blank. More research is needed, such as configuring a runtime monitor for the output of a specific application, or combining symbolic reasoning and causal reasoning with the model’s learning process to ensure that the output avoids failures from the source.

7.3 Perspective

Since LLMs are still in their infancy and have many known vulnerabilities, monitoring these models in real time is a longstanding challenge. In this section, we outline topics for future work to call on more researchers to address this challenge from three perspectives: why, what, and how.

Why does a model need to be monitored? The first thing we want to highlight is whether at some point the LLMs can be trained intelligent enough, so that there is no need to design a separate runtime monitor for these models. For instance, the model is endowed with abilities to automatically detect “illegal inputs” (e.g., out-of-distribution inputs) and guarantee the correctness of its outputs. From our authors’ perspective, achieving such level of intelligent models in the foreseeable future is very difficult, if not impossible. The main reasons are as follows. Existing LLMs are still learned from observations, i.e., a training dataset containing partial information. There is no evidence that current learning methods can infer from parts to wholes, even in the case of massive data, nor is there evidence that a training dataset captures all the information. Furthermore, existing learning methods do not characterize their generalization bounds but instead measure the so-called generalization error, which prevents the identification of “illegal inputs”. Therefore, it is necessary to monitor the model in real-time.

What should be monitored? One needs to overcome various vulnerabilities listed in Section 3 to reliably use LLMs in safety-critical applications. Equipping the model with a corresponding runtime monitor provides a possible solution complementary to offline verification methods. For example, there have been some works on monitoring whether the model’s prediction is made for out-of-distribution inputs and whether the model’s output is consistent with some existing fact base. However, to our knowledge, there is no monitoring work on other output failures, e.g., reasoning and code errors;

on intended attacks, e.g., robustness, backdoor, and data poisoning. Thus, we call on researchers and practitioners to investigate more in these topics.

How to better design a monitor for a model? The state-of-the-art methods are based on the uncertainty model’s predictions. Unfortunately, low uncertainty cannot assure the model’s prediction is reliable, and vice versa. To better design monitors for LLMs, we need the following efforts. First, some fundamental intrinsic issues of deep learning models must be better addressed, such as model implicit generalisation and decision boundaries and explainability of model decisions, which may provide more rigorous and formal characterisation and specification for building monitors. Specific to LLMs, some special issues need to be tackled, such as the unavailability of training datasets, the non-transparency of models, the generative nature of multi-modality, etc. Regarding specific tasks, such as the most studied problem of monitoring out-of-distribution inputs, principled methods for system design and evaluation of monitors still needs to be included, as current work is based on calibration of predictive confidence scores and evaluation on one-sided test datasets. Last, we call for great attention to unexplored topics, such as how to monitor other trustworthiness and responsibility issues, intended attacks, and unintended bugs, along with the model’s social and ethical alignments with human society.

8 Regulations and Ethical Use

V&V provides a set of technical means to support the alignment of LLMs with human interests. However, it has been argued that constructing LLMs that cannot be abused can be impossible. This suggests that technical means are necessary, but can be insufficient. To this end, it is needed to have *ethical means*, to supplement the technical means, in ensuring that the *complete alignment* of the use of LLMs with human interests. In the following, we discuss several recent signs of progress.

8.1 Regulate or Ban?

A recent debate on “a 6-month suspension on the development [4] vs. a regulated development” has shown the anxiety of, and the difference of opinions from, the community upon the possibilities of AI development being misaligned with human interests. More radical actions have also been taken. For example, Italy has reportedly banned the ChatGPT [1]. In a US Senate Hearing on May 2023, OpenAI CEO Sam Altman asked the government to regulate AI [221]. Actually, on AI regulations, major players such as the EU, US, UK, and China all have their respective approaches and initiatives, e.g., the EU’s GDPR [8], AI Act [21]. Data Act [22], the UK’s Data Protection Act [9] and pro-innovative approach to regulate AI [23], the US’s Blueprint for an AI Bill of Rights [19] and AI Risk Management Framework [12], and China’s regulations for recommendation algorithms [13], deep synthesis [11], and algorithm registry [20]. It is unclear (1) whether these regulations on the more general AI/ML, or other AI/ML algorithms, can automatically work for LLMs without any changes, and (2) how the regulations can be projected onto each other in a rigorous, yet operational, way. More importantly, even for general AI/ML, it still needs to be clarified how to *sufficiently*

and effectively address regulatory requirements (such as robustness and transparency) with technical means. The V&V framework proposed in this survey will be one viable solution.

Nevertheless, significant issues raised by the LLMs, notably the ChatGPT, include copyright and privacy. The ChatGPT developers reportedly use data from the internet, and it is unclear if the *copyrights of the training data* have been carefully dealt with, especially when the ChatGPT is eventually for commercial use. Moreover, as a conversational AI, the *privacy of the end users*, when engaged in a dialogue, is a serious concern. The end-users should be informed on whether and how their dialogues will be stored, used, and redistributed.

8.2 Responsible AI Principles

Responsible and accountable AI has been a topic of discussion for the past years (see e.g., [10, 25, 27]), with a gradual convergence to include properties such as transparency, explainability, fairness, robustness, security, privacy, etc. A governance framework is called for to ensure that these properties are implemented, evaluated, and monitored. A comprehensive discussion and comparison is out of the scope of this survey, but we note that, many properties are required, consistent definitions to many of them are still missing, and properties can be conflicting (i.e., the improvement of one property may compromise others). It is therefore not surprising that it can still be a long way to turn the principles into operational rules.

Specific to LLMs, ChatGPT and the like have led to severe concerns on e.g., potential misuse, unintended bias, and fair access. To this end, on the enterprise level, ethical principles are needed to guide the development and use of LLMs, including questioning whether something should be done rather than whether it can be done, as requested in [6].

8.3 Transparency and Explainability

First, OpenAI’s decision to not open-source GPT-3 and beyond has already led to concerns on the transparent development of AI. However, OpenAI said it plans to make more technical details available to other third parties for them to advise on how to weigh the competitive and safety considerations against the scientific value of further transparency. It is also important to note that, no technical details are available on how the guardrail is designed and implemented. It will also be interesting to discuss on whether the guardrail itself should undergo a verification process.

Second, it has been hard to interpret and explain the decisions of the deep learning models such as image classifiers. The situation becomes worsens when dealing with LLMs [139], which have emergent and hard-to-explain behaviours. For example, it has been observed that adding an incarnation, such as “Let’s think step by step”, to the prompt can achieve improved responses from GPT-3. Techniques are needed to explain such a phenomenon. This calls for extending explainable AI techniques to work with LLMs. In particular, it is necessary to consider the explanations’ robustness to explain why such incarnation can lead to improved, yet different, answers. To this end, some prior works on image classifiers, such as [292, 123], can be considered.

9 Conclusions

This paper provides an overview of the known vulnerabilities of LLMs, and discusses how the V&V techniques might be adapted to work with them. Given the LLMs are quickly adopted by applications that have direct or indirect interactions with end users, it is imperative that the deployed LLMs undergoes sufficient verdict processes to avoid any undesirable safety and trustworthy consequences. Considering the size and complexity of LLMs, white-box V&V techniques will likely become impractical and the community may need to develop black-box, non-determinism sensitive, V&V techniques. Moreover, multi-disciplinary development will ensure that all trustworthy issues are fully considered.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 956123. It is also financially supported by the U.K. EPSRC through End-to-End Conceptual Guarding of Neural Architectures [EP/T026995/1].

References

- [1] Italy became the first western country to ban chatgpt. <https://www.cnn.com/2023/04/04/italy-has-banned-chatgpt-heres-what-other-countries-are-doing.html>.
- [2] March 20 chatgpt outage: Here’s what happened.
- [3] Openai says a bug leaked sensitive chatgpt user data.
- [4] Pause giant ai experiments: An open letter. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.
- [5] Prompt engineering guide. <https://github.com/dair-ai/Prompt-Engineering-Guide/tree/main/guides>.
- [6] Tools such as chatgpt threaten transparent science; here are our ground rules for their use. <https://www.nature.com/articles/d41586-023-00191-1>.
- [7] Quality management systems - process validation guidance, 2004.
- [8] Eu gdpr. <https://gdpr-info.eu>, 2016.
- [9] The data protection act. <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>, 2018.
- [10] Ethics guidelines for trustworthy ai, 2018.

- [11] China's regulations on the administration of deep synthesis internet information services. <https://www.chinalawtranslate.com/en/deep-synthesis/>, 2021.
- [12] Ai risk management framework. <https://www.nist.gov/itl/ai-risk-management-framework>, 2022.
- [13] China's regulations on recommendation algorithms. http://www.cac.gov.cn/2022-01/04/c_1642894606258238.htm, 2022.
- [14] Content at scale, <https://contentatscale.ai/ai-content-detector/>, 2022.
- [15] Copyleaks, <https://copyleaks.com/ai-content-detector>, 2022.
- [16] New meta ai demo writes racist and inaccurate scientific literature, gets pulled. <https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/>, 2022.
- [17] Originality ai, <https://originality.ai>, 2022.
- [18] Prompt injection attacks against gpt-3. <https://simonwillison.net/2022/Sep/12/prompt-injection/>, 2022.
- [19] Blueprint for an ai bill of rights. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>, 2023.
- [20] China's algorithm registry. <https://beian.cac.gov.cn/#/index>, 2023.
- [21] Eu ai act. <https://artificialintelligenceact.eu>, 2023.
- [22] Eu data act. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113, 2023.
- [23] A pro-innovation approach to ai regulation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf, 2023.
- [24] Prompt leaking. https://learnprompting.org/docs/prompt_hacking/leaking, 2023.
- [25] Responsible ai principles from microsoft, 2023.
- [26] Three samsung employees reportedly leaked sensitive data to chatgpt. <https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html>, 2023.
- [27] Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector., 2023.

- [28] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.
- [29] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman. *Compilers: principles, techniques and tools*. 2020.
- [30] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G.-J. Nam, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE transactions on computer-aided design of integrated circuits and systems*, 34(10):1537–1557, 2015.
- [31] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.
- [32] U. Arora, W. Huang, and H. He. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*, 2021.
- [33] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [34] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- [35] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- [36] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [37] A. Balakrishnan, A. G. Puranic, X. Qin, A. Dokhanchi, J. V. Deshmukh, H. Ben Amor, and G. Fainekos. Specifying and evaluating quality metrics for vision-based perception systems. In *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1433–1438, 2019.
- [38] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.
- [39] Y. Belinkov and Y. Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

- [40] S. Bensalem, C.-H. Cheng, X. Huang, P. Katsaros, A. Molin, D. Nickovic, and D. Peled. Formal specification for learning-enabled autonomous systems (extended abstract). In *FoMLAS2022*, 2022.
- [41] S. Bensalem, Y. Lakhnech, and S. Owre. Invest: A tool for the verification of invariants. In *Computer Aided Verification: 10th International Conference, CAV'98 Vancouver, BC, Canada, June 28–July 2, 1998 Proceedings 10*, pages 505–510. Springer, 1998.
- [42] S. Bensalem, Y. Lakhnech, and H. Saidi. Powerful techniques for the automatic generation of invariants. In *Computer Aided Verification: 8th International Conference, CAV'96 New Brunswick, NJ, USA, July 31–August 3, 1996 Proceedings 8*, pages 323–335. Springer, 1996.
- [43] N. Berthier, A. Alshareef, J. Sharp, S. Schewe, and X. Huang. Abstraction and symbolic execution of deep neural networks with bayesian approximation. *CoRR*, 2021.
- [44] W. Bibel. *Automated theorem proving*. Springer Science & Business Media, 2013.
- [45] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.
- [46] G. Bonaert, D. I. Dimitrov, M. Baader, and M. Vechev. Fast and precise certification of transformers. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 466–481, 2021.
- [47] A. Borji. A categorical archive of chatgpt failures, 2023.
- [48] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean. Large language models in machine translation. 2007.
- [49] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [50] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

- [51] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [52] E. Bursztein. Attacks against machine learning — an overview.
- [53] E. Cambiaso and L. Caviglione. Scamming the scammers: Using chatgpt to reply mails for wasting time and resources, 2023.
- [54] Y. Cao, D. Li, M. Fang, T. Zhou, J. Gao, Y. Zhan, and D. Tao. Tasa: Deceiving question answering models by twin answer sentences attack. *arXiv preprint arXiv:2210.15221*, 2022.
- [55] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.
- [56] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021.
- [57] S. Chen, X. Bi, R. Gao, and X. Sun. Holistic sentence embeddings for better out-of-distribution detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6676–6686, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [58] S. Chen, B. H. Kann, M. B. Foote, H. J. Aerts, G. K. Savova, R. H. Mak, and D. S. Bitterman. The utility of chatgpt for cancer treatment information. *medRxiv*, pages 2023–03, 2023.
- [59] S. Chen, W. Yang, X. Bi, and X. Sun. Fine-tuning deteriorates general textual out-of-distribution detection by distorting task-agnostic features. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 552–567, 2023.

- [60] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, pages 554–569, 2021.
- [61] C. Cheng, G. Nührenberg, and H. Yasuoka. Runtime monitoring neuron activation patterns. In *DATE2019*, pages 300–303, 2019.
- [62] C.-H. Cheng, C. Wu, E. Seferis, and S. Bensalem. Prioritizing corners in ood detectors via symbolic string manipulation. In A. Bouajjani, L. Holík, and Z. Wu, editors, *Automated Technology for Verification and Analysis*, pages 397–413, Cham, 2022. Springer International Publishing.
- [63] M. Cheng, J. Yi, P.-Y. Chen, H. Zhang, and C.-J. Hsieh. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3601–3608, 2020.
- [64] Y. Cheng, L. Jiang, and W. Macherey. Robust neural machine translation with doubly adversarial inputs. *arXiv preprint arXiv:1906.02443*, 2019.
- [65] H. Cho, C. Park, J. Kang, K. M. Yoo, T. Kim, and S.-g. Lee. Enhancing out-of-distribution detection in natural language understanding via implicit layer ensemble. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 783–798, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [66] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [67] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [68] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- [69] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.
- [70] J. Dai, C. Chen, and Y. Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.
- [71] S. Dan and D. Roth. On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101, 2021.

- [72] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee Micro*, 38(1):82–99, 2018.
- [73] L. De Moura and N. Bjørner. Z3: An efficient smt solver. In *Tools and Algorithms for the Construction and Analysis of Systems: 14th International Conference, TACAS 2008, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2008, Budapest, Hungary, March 29-April 6, 2008. Proceedings 14*, pages 337–340. Springer, 2008.
- [74] S. Desai and G. Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online, Nov. 2020. Association for Computational Linguistics.
- [75] A. Deshpande, V. Murahari, T. Rajpurohit, A. Kalyan, and K. Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models, 2023.
- [76] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.
- [77] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [79] T. DeVries and G. W. Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint 1802.04865*, 2018.
- [80] N. Dey. Gpt: A family of open, compute-efficient, large language models, Apr 2023.
- [81] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.
- [82] T. Du, S. Ji, L. Shen, Y. Zhang, J. Li, J. Shi, C. Fang, J. Yin, R. Beyah, and T. Wang. Cert-rnn: Towards certifying the robustness of recurrent neural networks. *CCS*, 21(2021):15–19, 2021.

- [83] H. Duan, Y. Yang, A. Abbasi, and K. Y. Tam. BARLE: Background-aware representation learning for background shift out-of-distribution detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 750–764, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [84] J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu. Are diffusion models vulnerable to membership inference attacks?, 2023.
- [85] E2Analyst. Gpt-4: Everything you want to know about openai’s new ai model, Mar 2023.
- [86] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.
- [87] D. Eppstein. Zonohedra and zonotopes. 1995.
- [88] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [89] F. Farhat, S. S. Sohail, and D. Øivind Madsen. How trustworthy is chatgpt? the case of bibliometric analyses. 2023.
- [90] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res*, 23:1–40, 2021.
- [91] M. Fitting. *First-order logic and automated theorem proving*. Springer Science & Business Media, 2012.
- [92] E. Frantar and D. Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *arXiv preprint arXiv:2208.11580*, 2022.
- [93] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.
- [94] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. Mathematical capabilities of chatgpt, 2023.
- [95] V. Gangal, A. Arora, A. Einolghozati, and S. Gupta. Likelihood ratios and generative classifiers for unsupervised out-of-domain detection in task oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7764–7771, 2020.
- [96] D. Ganguli, A. Askell, N. Schiefer, T. Liao, K. Lukošiu̇tė, A. Chen, A. Goldie, A. Mirhoseini, C. Olsson, D. Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

- [97] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.
- [98] I. Goodfellow and N. Papernot. The challenge of verification and testing of machine learning. *Cleverhans-blog*, 2017.
- [99] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [100] D. Gopinath, K. Wang, M. Zhang, C. S. Pasareanu, and S. Khurshid. Symbolic execution for deep neural networks. *CoRR*, abs/1807.10439, 2018.
- [101] S. Goyal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.
- [102] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran. A survey in adversarial defences and robustness in nlp. *arXiv preprint arXiv:2203.06414*, 2022.
- [103] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz. More than you’ve asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. *arXiv preprint arXiv:2302.12173*, 2023.
- [104] J.-C. Gu, T. Li, Q. Liu, Z.-H. Ling, Z. Su, S. Wei, and X. Zhu. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*, page 2041–2044, New York, NY, USA, 2020. Association for Computing Machinery.
- [105] S. Gu, A. Kshirsagar, Y. Du, G. Chen, Y. Yang, J. Peters, and A. Knoll. A human-centered safe robot reinforcement learning framework with interactive behaviors. *arXiv preprint arXiv:2302.13137*, 2023.
- [106] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.
- [107] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.
- [108] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint 1610.02136*, 2016.
- [109] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. Pre-trained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, 2020.

- [110] T. A. Henzinger, A. Lukina, and C. Schilling. Outside the box: Abstraction-based monitoring of neural networks. In *ECAI2020*.
- [111] A. Hintze. Chatgpt believes it is conscious, 2023.
- [112] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [113] J. Holmes, Z. Liu, L. Zhang, Y. Ding, T. T. Sio, L. A. McGee, J. B. Ashman, X. Li, T. Liu, J. Shen, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *arXiv preprint arXiv:2304.01938*, 2023.
- [114] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google’s perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.
- [115] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [116] O. Hrinchuk, M. Popova, and B. Ginsburg. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 7074–7078. IEEE, 2020.
- [117] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [118] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR, 2017.
- [119] H. Huang, Z. Li, L. Wang, S. Chen, B. Dong, and X. Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.
- [120] P.-S. Huang, R. Stanforth, J. Welbl, C. Dyer, D. Yogatama, S. Goyal, K. Dvijotham, and P. Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *arXiv preprint arXiv:1909.01492*, 2019.
- [121] W. Huang, Y. Sun, J. Sharp, W. Ruan, J. Meng, and X. Huang. Coverage guided testing for recurrent neural networks. *CoRR*, abs/1911.01952, 2020.
- [122] W. Huang, X. Zhao, A. Banks, V. Cox, and X. Huang. Hierarchical distribution-aware testing of deep learning, 2022.

- [123] W. Huang, X. Zhao, G. Jin, and X. Huang. Safari: Versatile and efficient evaluations for robustness of interpretability, 2022.
- [124] X. Huang, G. Jin, and W. Ruan. *Machine Learning Safety*. Springer, 2023.
- [125] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020.
- [126] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu. Safety verification of deep neural networks. In *CAV2017*, pages 3–29, 2017.
- [127] X. Huang, W. Ruan, Q. Tang, and X. Zhao. Bridging formal methods and machine learning with global optimisation. In *Formal Methods and Software Engineering: 23rd International Conference on Formal Engineering Methods, ICFEM 2022, Madrid, Spain, October 24–27, 2022, Proceedings*, page 1–19, Berlin, Heidelberg, 2022. Springer-Verlag.
- [128] A. Ivankay, I. Girardi, C. Marchiori, and P. Frossard. Fooling explanations in text classifiers. *arXiv preprint arXiv:2206.03178*, 2022.
- [129] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*, 2018.
- [130] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.
- [131] M. Jang and T. Lukasiewicz. Consistency analysis of chatgpt, 2023.
- [132] Y. Ji, Y. Gong, Y. Peng, C. Ni, P. Sun, D. Pan, B. Ma, and X. Li. Exploring chatgpt’s ability to rank content: A preliminary study on consistency with human preferences, 2023.
- [133] R. Jia and P. Liang. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*, 2017.
- [134] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*, 2019.
- [135] A. Q. Jiang, S. Welleck, J. P. Zhou, W. Li, J. Liu, M. Jamnik, T. Lacroix, Y. Wu, and G. Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. *arXiv preprint arXiv:2210.12283*, 2022.
- [136] W. Jiao, W. Wang, J.-t. Huang, X. Wang, and Z. Tu. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023.
- [137] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

- [138] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*, 2021.
- [139] S. Kambhampati. Changing the nature of ai research. *Commun. ACM*, 65(9):8–9, aug 2022.
- [140] D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. *arXiv preprint arXiv:2302.05733*, 2023.
- [141] Y. Kang, Q. Zhang, and R. Roth. The ethics of ai-generated maps: A study of dalle 2 and implications for cartography, 2023.
- [142] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [143] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo. Gpt-4 passes the bar exam. *Available at SSRN 4389233*, 2023.
- [144] R. Khoury, A. R. Avila, J. Brunelle, and B. M. Camara. How secure is code generated by chatgpt? *arXiv preprint arXiv:2304.09655*, 2023.
- [145] Y.-M. Kim. Data and fair use. 141:5–53, 2023.
- [146] C.-Y. Ko, Z. Lyu, L. Weng, L. Daniel, N. Wong, and D. Lin. Popqorn: Quantifying robustness of recurrent neural networks. In *International Conference on Machine Learning*, pages 3468–3477. PMLR, 2019.
- [147] V. Kuleshov, S. Thakoor, T. Lau, and S. Ermon. Adversarial examples for natural language classification problems. 2018.
- [148] A. Kumar, K. Ahuja, R. Vadapalli, and P. Talukdar. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345, 2020.
- [149] K. Kurita, P. Michel, and G. Neubig. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2793–2806, 2020.
- [150] E. La Malfa, M. Wu, L. Laurenti, B. Wang, A. Hartshorn, and M. Kwiatkowska. Assessing robustness of text classification through maximal safe radius computation. *arXiv preprint arXiv:2010.02004*, 2020.
- [151] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. <https://huggingface.co/blog/rlhf>.
- [152] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

- [153] J. Y. Lee. Can an artificial intelligence chatbot be the author of a scholarly article? 20:6–, 2023.
- [154] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- [155] N. Lee, Y. Bang, A. Madotto, and P. Fung. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*, 2020.
- [156] P. Lee. Learning from tay’s introduction.
- [157] Y. Lei, Y. Cao, D. Li, T. Zhou, M. Fang, and M. Pechenizkiy. Phrase-level textual adversarial attack with label preservation. *arXiv preprint arXiv:2205.10710*, 2022.
- [158] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.
- [159] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.
- [160] H. Li, D. Guo, W. Fan, M. Xu, and Y. Song. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023.
- [161] J. Li, S. Ji, T. Du, B. Li, and T. Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.
- [162] S. Li, H. Liu, T. Dong, B. Z. H. Zhao, M. Xue, H. Zhu, and J. Lu. Hidden backdoors in human-centric language models. In *CCS ’21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 3123–3140. ACM, 2021.
- [163] X. Li, J. Li, X. Sun, C. Fan, T. Zhang, F. Wu, Y. Meng, and J. Zhang. kfolden: k-fold ensemble for out-of-distribution detection-fold ensemble for out-of-distribution detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3115, 2021.
- [164] Y. Li, L. Ding, and X. Gao. On the decision boundary of deep neural networks. *CoRR*, abs/1808.05385, 2018.
- [165] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*, 2017.

- [166] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.
- [167] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [168] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung. Caire: An empathetic neural chatbot. *arXiv preprint arXiv:1907.12108*, 2019.
- [169] C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, and M. J. Kochenderfer. Algorithms for verifying deep neural networks, 2020.
- [170] H. Liu, R. Ning, Z. Teng, J. Liu, Q. Zhou, and Y. Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023.
- [171] J. Liu, C. S. Xia, Y. Wang, and L. Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *arXiv preprint arXiv:2305.01210*, 2023.
- [172] W. Liu, X. Wang, J. Owens, and Y. Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.
- [173] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.
- [174] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*, 2023.
- [175] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [176] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021.
- [177] Z. Liu, X. Yu, L. Zhang, Z. Wu, C. Cao, H. Dai, L. Zhao, W. Liu, D. Shen, Q. Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.
- [178] R. Lou, K. Zhang, and W. Yin. Is prompt all you need? no. a comprehensive and broader view of instruction learning, 2023.

- [179] N. Madaan, I. Padhi, N. Panwar, and D. Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13516–13524, 2021.
- [180] K. Malinka, M. Perešíni, A. Firc, O. Hujňák, and F. Januš. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?, 2023.
- [181] Z. Manna and A. Pnueli. *The temporal logic of reactive and concurrent systems: Specification*. Springer Science & Business Media, 2012.
- [182] N. Maus, P. Chao, E. Wong, and J. Gardner. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*, 2023.
- [183] W. McCune. Prover9 and mace4, 2005.
- [184] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.
- [185] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, and T. Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206. PMLR, 2020.
- [186] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia. Exploiting machine learning to subvert your spam filter. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET’08, USA, 2008. USENIX Association.
- [187] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [188] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.
- [189] OpenAI. GPT-4 Technical Report. *arXiv e-prints 2303.08774*, 2023.
- [190] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [191] G. Park, B. Park, S. J. Kwon, B. Kim, Y. Lee, and D. Lee. nuqmm: Quantized matmul for efficient inference of large-scale generative language models. *arXiv preprint arXiv:2206.09557*, 2022.

- [192] D. Patterson, J. Gonzalez, U. Holzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.
- [193] A. Pegoraro, K. Kumari, H. Fereidooni, and A.-R. Sadeghi. To chatgpt, or not to chatgpt: That is the question!, 2023.
- [194] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [195] F. Perez and I. Ribeiro. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.
- [196] A. Podolskiy, D. Lipin, A. Bout, E. Artemova, and I. Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682, 2021.
- [197] Y. Qi, X. Zhao, and X. Huang. Safety analysis in the era of large language models: A case study of stpa using chatgpt, 2023.
- [198] A. Radford, R. Jozefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.
- [199] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [200] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- [201] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [202] R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.
- [203] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [204] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

- [205] M. V. Reiss. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark, 2023.
- [206] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.
- [207] S. Ren, Y. Deng, K. He, and W. Che. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097, 2019.
- [208] X. Ren, P. Zhou, X. Meng, X. Huang, Y. Wang, W. Wang, P. Li, X. Zhang, A. Podolskiy, G. Arshinov, et al. Pangu- σ : Towards trillion parameter language model with sparse heterogeneous computing. *arXiv preprint arXiv:2303.10845*, 2023.
- [209] J. T. Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016.
- [210] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [211] W. Ruan, X. Huang, and M. Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In *IJCAI2018*, pages 2651–2659, 2018.
- [212] W. Ruan, M. Wu, Y. Sun, X. Huang, D. Kroening, and M. Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *IJCAI2019*, pages 5944–5952, 2019.
- [213] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.
- [214] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.
- [215] J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly. The self-perception and political biases of chatgpt, 2023.
- [216] W. Ryou, J. Chen, M. Balunovic, G. Singh, A. Dan, and M. Vechev. Scalable polyhedral verification of recurrent neural networks. In *International Conference on Computer Aided Verification*, pages 225–248. Springer, 2021.

- [217] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [218] S. Samanta and S. Mehta. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*, 2017.
- [219] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [220] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [221] U. Senate. Senate judiciary subcommittee hearing on oversight of ai, 2023.
- [222] S. A. Seshia, D. Sadigh, and S. S. Sastry. Towards verified artificial intelligence. *arXiv preprint arXiv:1606.08514*, 2016.
- [223] M. Shanahan. Talking about large language models. *arXiv preprint arXiv:2212.03551*, 2022.
- [224] L. Shen, S. Ji, X. Zhang, J. Li, J. Chen, J. Shi, C. Fang, J. Yin, and T. Wang. Backdoor pre-trained models can transfer to all. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3141–3158, 2021.
- [225] X. Shen, Z. Chen, M. Backes, and Y. Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*, 2023.
- [226] Y. Shen, Y.-C. Hsu, A. Ray, and H. Jin. Enhancing the generalization for intent classification and out-of-domain detection in slu. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2443–2453, 2021.
- [227] K. Shuster, M. Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224*, 2022.
- [228] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- [229] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

- [230] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlG 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [231] D. Sobania, M. Briesch, C. Hanna, and J. Petke. An analysis of the automatic bug fixing performance of chatgpt. *arXiv preprint arXiv:2301.08653*, 2023.
- [232] S. Soltan, S. Ananthakrishnan, J. FitzGerald, R. Gupta, W. Hamza, H. Khan, C. Peris, S. Rawls, A. Rosenbaum, A. Rumshisky, et al. Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model. *arXiv preprint arXiv:2208.01448*, 2022.
- [233] L. Struppek, D. Hintersdorf, and K. Kersting. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. *arXiv preprint arXiv:2211.02408*, 2022.
- [234] H. Sun, Z. Zhang, J. Deng, J. Cheng, and M. Huang. Safety assessment of chinese large language models, 2023.
- [235] Y. Sun, X. Huang, and D. Kroening. Testing deep neural networks. *CoRR*, abs/1803.04792, 2018.
- [236] Y. Sun, X. Huang, D. Kroening, J. Sharp, M. Hill, and R. Ashmore. Structural test coverage criteria for deep neural networks. *ACM Trans. Embed. Comput. Syst.*, 18(5s), Oct. 2019.
- [237] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.
- [238] Y. Sun, M. Wu, W. Ruan, X. Huang, M. Kwiatkowska, and D. Kroening. Concolic testing for deep neural networks. In *ASE2018*, 2018.
- [239] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal. Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*, 78:80–95, 2016.
- [240] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.
- [241] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [242] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. Fever: A large-scale dataset for fact extraction and verification. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 809–819. Association for Computational Linguistics (ACL), 2018.

- [243] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [244] A. S. Tulshan and S. N. Dhage. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers 4*, pages 190–201. Springer, 2019.
- [245] A. Uchendu, J. Lee, H. Shen, T. Le, T.-H. K. Huang, and D. Lee. Understanding individual and team-based human factors in detecting deepfake texts. 2023.
- [246] M. Y. Vardi and P. Wolper. An automata-theoretic approach to automatic program verification. In *1st Symposium in Logic in Computer Science (LICS)*. IEEE Computer Society, 1986.
- [247] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [248] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [249] M. Wallace, R. Khandelwal, and B. Tang. Does ibp scale? 2022.
- [250] B. Wang, C. Xu, S. Wang, Z. Gan, Y. Cheng, J. Gao, A. H. Awadallah, and B. Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021.
- [251] F. Wang, P. Xu, W. Ruan, and X. Huang. Towards verifying the geometric robustness of large-scale neural networks, 2023.
- [252] G. Wang, Y. Lin, and W. Yi. Kernel fusion: An effective method for better power efficiency on multithreaded gpu. In *2010 IEEE/ACM Int’l Conference on Green Computing and Communications & Int’l Conference on Cyber, Physical and Social Computing*, pages 344–350. IEEE, 2010.
- [253] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.
- [254] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, and X. Xie. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *arXiv e-prints*, page arXiv:2302.12095, Feb. 2023.

- [255] W. Wang, P. Tang, J. Lou, and L. Xiong. Certified robustness to word substitution attack with differential privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1112, 2021.
- [256] Y. Wang and M. Bansal. Robust machine comprehension models via adversarial training. *arXiv preprint arXiv:1804.06473*, 2018.
- [257] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [258] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel. Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. In *ICLR2018*, 2018.
- [259] Y. Weng, M. Zhu, S. He, K. Liu, and J. Zhao. Large language models are reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022.
- [260] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, K. Liu, and J. Zhao. Neural comprehension: Language models with compiled neural networks. *arXiv preprint arXiv:2304.01665*, 2023.
- [261] M. Wicker, X. Huang, and M. Kwiatkowska. Feature-guided black-box safety testing of deep neural networks. In *TACAS2018*, pages 408–426, 2018.
- [262] Y. Wolf, N. Wies, Y. Levine, and A. Shashua. Fundamental limitations of alignment in large language models, 2023.
- [263] D. Wu, G. Jin, H. Yu, X. Yi, and X. Huang. Optimising event-driven spiking neural network with regularisation and cutoff. *arXiv preprint arXiv:2301.09522*, 2023.
- [264] D. Wu, X. Yi, and X. Huang. A little energy goes a long way: Build an energy-efficient, accurate spiking neural network from convolutional neural network. *Frontiers in neuroscience*, 16, 2022.
- [265] H. Wu, W. Wang, Y. Wan, W. Jiao, and M. Lyu. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark, 2023.
- [266] M. Wu, M. Wicker, W. Ruan, X. Huang, and M. Kwiatkowska. A game-based approximate verification of deep neural networks with provable guarantees. *Theoretical Computer Science*, 807:298–329, 2020.
- [267] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.
- [268] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 2023.

- [269] H. Xu, K. He, Y. Yan, S. Liu, Z. Liu, and W. Xu. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460, 2020.
- [270] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.
- [271] P. Xu, W. Ruan, and X. Huang. Quantifying safety risks of deep neural networks. *Complex & Intelligent Systems*, pages 1–18, 2022.
- [272] Yandex. Yandex/yalm-100b: Pretrained language model with 100b parameters.
- [273] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.
- [274] W. Yang, L. Li, Z. Zhang, X. Ren, X. Sun, and B. He. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058, 2021.
- [275] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019.
- [276] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022.
- [277] M. Ye, C. Gong, and Q. Liu. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv preprint arXiv:2005.14424*, 2020.
- [278] X. Ye, S. Iyer, A. Celikyilmaz, V. Stoyanov, G. Durrett, and R. Pasunuru. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*, 2022.
- [279] E. Yilmaz and C. Toraman. D2u: Distance-to-uniform learning for out-of-scope detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2093–2108, 2022.
- [280] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.
- [281] J. Zeng, X. Zheng, J. Xu, L. Li, L. Yuan, and X. Huang. Certified robustness to text adversarial attacks by randomized [mask]. *arXiv preprint arXiv:2105.03743*, 2021.

- [282] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, et al. Pangu- α : Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.
- [283] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, et al. Pangu- α : Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.
- [284] Z. Zeng, K. He, Y. Yan, Z. Liu, Y. Wu, H. Xu, H. Jiang, and W. Xu. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–878, 2021.
- [285] C. Zhang, W. Ruan, F. Wang, P. Xu, G. Min, and X. Huang. Model-agnostic reachability analysis on deep neural networks. *arXiv preprint arXiv:2304.00813*, 2023.
- [286] C. Zhang, W. Ruan, and P. Xu. Reachability analysis of neural network control systems. *arXiv preprint arXiv:2301.12100*, 2023.
- [287] J. Zhang, Y. Zhao, M. Saleh, and P. Liu. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR, 13–18 Jul 2020.
- [288] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [289] Y. Zhang, A. Albarghouthi, and L. D’Antoni. Certified robustness to programmable transformations in lstms. *arXiv preprint arXiv:2102.07818*, 2021.
- [290] R. Zhao, X. Li, Y. K. Chia, B. Ding, and L. Bing. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*, 2023.
- [291] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. A survey of large language models, 2023.
- [292] X. Zhao, W. Huang, X. Huang, V. Robu, and D. Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In C. de Campos and M. H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 887–896. PMLR, 27–30 Jul 2021.

- [293] X. Zhao, W. Huang, S. Schewe, Y. Dong, and X. Huang. Detecting operational adversarial examples for reliable deep learning. In *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S)*, pages 5–6, 2021.
- [294] Z. Zhao, D. Dua, and S. Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.
- [295] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*, 2023.
- [296] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt, 2023.
- [297] W. Zhou, F. Liu, and M. Chen. Contrastive out-of-distribution detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
- [298] Y. Zhou, P. Liu, and X. Qiu. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5129–5141, 2022.
- [299] R.-J. Zhu, Q. Zhao, and J. K. Eshraghian. Spikegpt: Generative pre-trained language model with spiking neural networks. *arXiv preprint arXiv:2302.13939*, 2023.
- [300] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.