MALTO at SemEval-2024 Task 6: Leveraging Synthetic Data for LLM Hallucination Detection

Federico Borra* **Claudio Savelli*** **Giacomo Rosso**

Alkis Koudounas **Flavio Giobergia**

Politecnico di Torino, Italy

Abstract

In Natural Language Generation (NLG), contemporary Large Language Models (LLMs) face several challenges, such as generating fluent yet inaccurate outputs and relying on fluency-centric metrics. This often leads to neural networks exhibiting "hallucinations". The SHROOM challenge focuses on automatically identifying these hallucinations in the generated text. To tackle these issues, we introduce two key components, a data augmentation pipeline incorporating LLM-assisted pseudo-labelling and sentence rephrasing, and a voting ensemble from three models pre-trained on Natural Language Inference (NLI) tasks and fine-tuned on diverse datasets.

1 Introduction

Natural Language Generation (NLG) models are AI systems that use neural networks to produce human-like text. They have shown significant advancements in recent years, particularly with the advent of transformer-based architectures such as GPT (Generative Pre-trained Transformer) (Radford et al., 2018). These models offered unprecedented levels of fluency and coherence in generated text (Han et al., 2021). However, a critical challenge arises: these models can produce linguistically fluent but semantically inaccurate outputs, a phenomenon referred to as *hallucination* (Ji et al., 2023). This may also lead to the generation of offensive, misleading, or factually incorrect content, as highlighted in previous studies (Engstrom and Gelbach, 2020; Bender et al., 2021). Such issues could have profound repercussions, especially for marginalized or under-resourced communities (Surden, 2020; Volokh, 2023; Koudounas et al., 2024).

To address this challenge, the Shared-task on Hallucinations and Related Observable Overgeneration Mistakes (SHROOM) has been proposed at SemEval 2024 (Mickus et al., 2024). In particular, the Shared task aims to address the existing gap in assessing the semantic correctness and meaningfulness of NLG models. The ever-increasing adoption of such models makes it necessary to automatically detect and mitigate semantic hallucinations (Huang et al., 2023).

Some examples to tackle hallucination detection tasks in literature (Ji et al., 2023) are: (i) *Information Extraction and Comparison* between a generated text and a ground truth, (ii) *Natural Language Inference Metrics* that express the entailment between generated text and a ground truth or (iii) *Faithfulness Classification Metrics* that leverage upon knowledge-grounded datasets.

In this work, we address the SHROOM shared task by introducing an automatic pipeline of hallucination detection through the comparison between a generated text and a ground truth text. We propose enriching the original data available using different augmentation techniques, including LLM-aided pseudo-labeling and sentence rephrasing. Additionally, we suggest using an ensemble of three different approaches, incorporating a simple BERT-based classifier, a model trained through Conditioned Reinforcement Learning Fine Tuning (C-RLFT) (Wang et al., 2023), and a sequential model based on iterative fine-tuning. We show how this ensemble benefits from using different, complementary approaches, particularly in recall. Our methodology obtained an accuracy of 80.07% in the SemEval-Task 6 SHROOM.

1.1 Dataset

The dataset available for the SHROOM challenge is a collection of objects. Each object represents a solution of a generative language model to either of the three tasks. The first is *Definition Modeling* (DM) (Noraset et al., 2017), the task of providing a definition for a given word. The second is *Machine Translation* (MT), i.e., translating from a source

*These authors contributed equally to this work.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1678–1684

language to a target one: this has been shown to be a challenging task that can be addressed from both statistical (Koehn, 2009) and, in recent years, neural perspectives (Bahdanau et al., 2014; Giobergia et al., 2020). Finally, the task of *Paraphrase Generation* (PG) consists of paraphrasing, i.e., producing an alternative version, of a source sentence (Zhou and Bhat, 2021). Each solution has been annotated, based on its contents, as either a *hallucination* of the generative model or *not hallucination* by 5 human annotators.

For each object, the available information includes (i) the *source (src)*, which is the input text given to the generative language model, (ii) the *hypothesis (hyp)*, which represents the generated textual output of the model, and (iii) the *target (tgt)* which is the intended reference or "gold" text that the model is supposed to generate. Additionally, the task field indicates the type of task being solved, either DM, MT, or PG. The label, either *"hallucination"* or *"not hallucination"*, is determined through majority voting among five annotators, with *p(hal)* indicating the proportion of annotators who labeled the data point as a hallucination.

The gold (and augmented) data cardinalities are defined in Table 1. The training dataset comprises 500 instances with gold labels, denoted as $\mathcal{D}_g$, and 30,000 unlabelled instances, referred to as $\mathcal{D}_u$ (10,000 for each of the three tasks). The evaluation split contains 1,500 labelled samples, with 500 instances used for validation ($\mathcal{D}_v$) and 1,000 for testing ($\mathcal{D}_t$). We use the validation set for fine-tuning the ensemble layer (refer to Section 2.3), while the final test set provides overall results (see Section 3).

We further rephrase the original 500 labelled sentences of the training set ($\mathcal{D}_r$ in the table, see Section 2.1.2), while applying weak labelling to the 30,000 unlabelled instances ($\mathcal{D}_{pl}$, see Section 2.1.1).

2 Methodology

The main goal of this work is to propose a binary classification model to predict whether the answer to a given query is a hallucination or not. Figure 1 presents the main architecture adopted to address this task[1]. We propose (i) using a data augmentation pipeline (see Section 2.1) consisting of Large Language Model (LLM)-aided pseudo-labelling and sentence rephrasing and (ii) adopting an ensemble model (see Section 2.3) based on the results of three models, defined as follows:

- *Baseline* model, a binary classifier based on a semantic-aware embedding (e.g. BERT-based (Devlin et al., 2019)). The baseline model is presented in Section 2.2.1
- *C-RLFT* (Conditioned Reinforcement Learning Fine Tuning (Wang et al., 2023)), based on the introduction of pseudo-labels and augmented data, with different weighting schemes based on the quality of each data point. We cover C-RLFT in more detail in Section 2.2.2
- *Sequential* model, based on the iterative fine-tuning of the baseline model with increasingly higher-quality data, as detailed in Section 2.2.3

2.1 Data Augmentation

Due to the scarcity of data, we developed an approach to extend the number of labelled samples we could use to train our models. We specifically leverage two distinct techniques: pseudo-labelling and sentence rephrasing. Both approaches are based on LLMs and, as such, may themselves be subject to hallucinations or inaccuracies. As detailed next, we mitigate this problem by (1) using the C-RLFT technique (Wang et al., 2023), which involves assigning different weights to mixed-quality samples, and (2) with a sequential training that introduces different-quality labels at different training stages.

2.1.1 Pseudo Labeling

As stated in Section 1.1, only a small fraction of the dataset available is labelled. We introduce additional pseudo labels, as obtained by querying an LLM in a few-shot learning setting. Based on the hardware available, we tested several LLM models to assess the reliability of the pseudo labels produced (in terms of accuracy). We identified SOLAR (Kim et al., 2023) as being the best-performing model among the pool of candidates. Thus, we leverage it to generate synthetic labels for unlabelled data through a few-shot learning approach. We refer to this augmented dataset as $\mathcal{D}_{pl}$.

2.1.2 Sentence Rephrasing

We utilized sentence rephrasing based on GPT-4 as an additional data augmentation technique. We

[1]The code to replicate the experiments can be found at https://github.com/MAL-TO/shroom

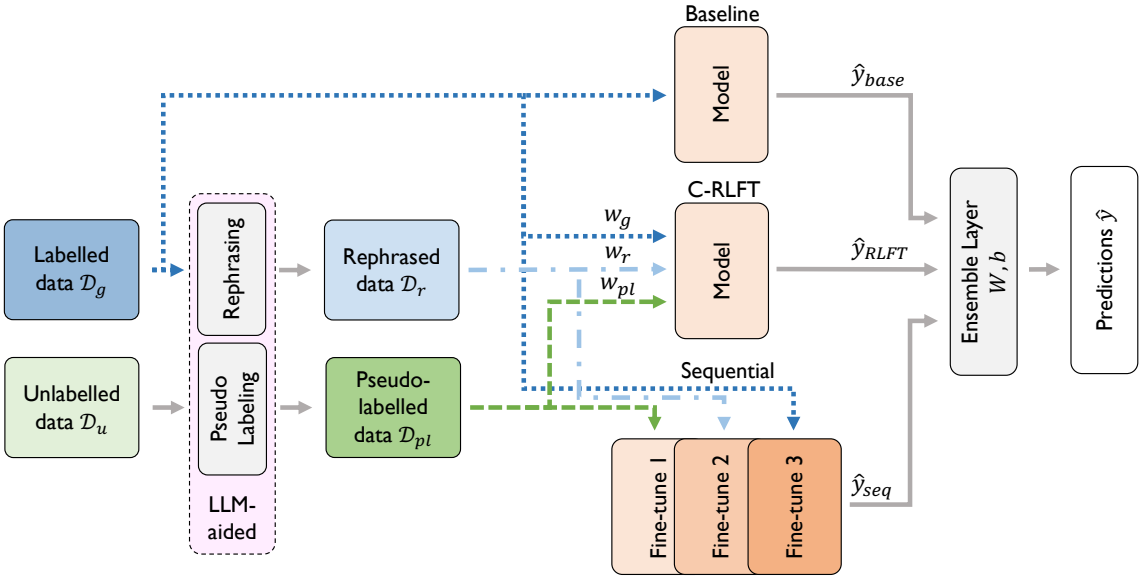

Figure 1: Pipeline architecture depicting data augmentation techniques and weighted ensemble of strategies.

do so by rephrasing both the model output and the target output of each gold sample. This approach aims to provide the model with diverse data while maintaining the reliability of the labels. We refer to this dataset as $\mathcal{D}_r$.

2.2 Models

We adopt an ensemble of three models, as described below. All models are based on DeBERTa (**?**). More specifically, we use a baseline model that has been fine-tuned in different ways.

2.2.1 Baseline

We employed a baseline model utilizing the DeBERTa encoder pre-trained on the Natural Language Inference (NLI) task, with a binary classification head. We fine-tune this model on the provided classification task using only data with the gold labels available, referred to as $\mathcal{D}_g$. The training approach involved minimizing the Binary Cross Entropy (BCE) loss.

We use the probability $p(hal)$ as the ground truth instead of the binary label. This is done to better reflect the distribution of votes of the human annotators in the output logits of the model.

2.2.2 C-RLFT

Conditioned Reinforcement Learning Fine Tuning (C-RLFT) is a technique that refines models using coarse-grained reward labels, allowing fine-tuning with both expert and sub-optimal data lacking preference labels. In our specific scenario, we fine-tuned the model by assigning different weights to data based on their label type, i.e., synthetic or gold. The weight assigned to each data sample influences the contribution to the final BCE loss.

We define a weighting scheme for the gold dataset $\mathcal{D}_g$, the pseudo-labelled dataset $\mathcal{D}_{pl}$ and the rephrased dataset $\mathcal{D}_r$, as follows:

$$w(x_i) = \begin{cases} w_g & \text{if } x_i \in \mathcal{D}_g \\ w_r & \text{if } x_i \in \mathcal{D}_r \\ w_{pl} & \text{if } x_i \in \mathcal{D}_{pl} \end{cases}$$

We choose weights $w_g > w_r > w_{pl}$. In this way, we aim to assign a higher importance to gold labels due to their reliability. The lowest weight is assigned to the pseudo-labelled points because of the lower quality of the automatically assigned labels. An intermediate weight is given to rephrased sentences due to the higher quality of the ground truth w.r.t. the pseudo-labelled points. The weighted loss is thus defined as follows, for a point x_i with ground truth y_i, as computed for a binary classifier $f(\cdot)$:

$$\begin{aligned} wBCE(x_i, y_i) = &- w(x_i) \cdot (y_i \; log \; f(x_i) \\ &+ (1 - y_i) \; \log(1 - f(x_i))) \end{aligned}$$

2.2.3 Sequential

The third model used is based on a sequential strategy that uses both generated and augmented data. We introduce three fine-tuning steps, performed sequentially on the initial model. The model initially underwent fine-tuning using the pseudo-labelled dataset $\mathcal{D}_{pl}$, which is the lowest-quality dataset

among the three available. Subsequently, we fine-tuned the resulting model on the rephrased data $\mathcal{D}_r$, which benefits from the original, correct, labels. The final fine-tuning step is then executed on the golden truth dataset $\mathcal{D}_g$. This approach is inspired by curriculum learning (Soviany et al., 2022), with data being ordered by veracity instead of difficulty.

This strategy aims to enhance the model's understanding of the task by starting with a substantial amount of data, including the less reliable synthetic labels, and progressively updating the model parameters with increasingly consistent data. This sequential approach allows the model to first adapt to the task using a broader dataset and then refine its knowledge with the highest quality data available.

2.3 Ensemble

The final step in the proposed pipeline involves creating an ensemble of results from the previously-introduced techniques, which has already proven to be effective in other NLP tasks (Jia et al., 2023; Koudounas et al., 2023). We trained three distinct models (*baseline*, *C-RLFT*, *sequential*) with specific strategies, and we generated their outputs ($\hat{y}_{base}$, $\hat{y}_{RLFT}$, $\hat{y}_{seq}$) on a validation set of previously unseen gold data. We obtain a single result $\hat{y}$ from the previous ones by using a single-layer network ($W \in \mathbb{R}^3$ and $b \in \mathbb{R}$), as follows:

$$\begin{aligned} \hat{y}_{models} &= (\hat{y}_{base}, \hat{y}_{RLFT}, \hat{y}_{seq}) \\ \hat{y} &= \sigma(W^\intercal \hat{y}_{models} + b) \end{aligned} \tag{1}$$

This network is trained to predict a single output from the three models' predicted probabilities. We trained this network by minimizing a BCE function.

3 Experimental Results

This section presents the experimental setup used, and the main results obtained.

3.1 Experimental Setup

The dataset used to train and validate the model is the one made available in the SHROOM challenge's model agnostic track (refer to Section 1.1). The augmentations are specified in Section 2.1.

For the model backbone and synthetic labelling we leverage Huggingface pre-trained models[2]. We also leverage *GPT-4* for sentence rephrasing. All the experiments' results are obtained based on 5 different runs. For C-RLFT, we identified the best performance for weights $w_g = 1.01, w_r = 0.4, w_{pl} = 0.1$.

Dataset Type	Label	Split	#Samples
$\mathcal{D}_g$	yes	Train	500
$\mathcal{D}_r$	yes	Train	500
$\mathcal{D}_u$	no	Train	30,000
$\mathcal{D}_{pl}$	weak	Train	30,000
$\mathcal{D}_v$	yes	Val	500
$\mathcal{D}_t$	yes	Test	1000

Table 1: Dataset type, labelling, and number of instances for each considered split.

3.2 Model performance

We summarize the results obtained on the test set in Table 2. We report the results in terms of F_1 score, precision, and recall on the "Hallucination" class, as well as overall accuracy.

We use as backbone both DeBERTa and a version of DeBERTa that has been fine-tuned on the Machine Natural Language Inference (MNLI) task. Further discussions on the choice of the backbone are presented in Section 3.3.

Section 3.4 highlights the result differences for each of the considered strategies and includes additional considerations on the ensemble of the approaches. Finally, we provide qualitative examples of the results in Section 3.5.

3.3 Backbone impact

We start by examining the differences between two backbone models, both fine-tuned on the gold data only – these are referred to as the *baseline* models in Table 2.

There is a notable increase of 0.09 in the F_1 score for the MNLI-fine-tuned model compared to the original DeBERTa. Interestingly, all the proposed DeBERTa-based approaches are still outperformed by the baseline DeBERTa+MNLI-based model (although to a lesser extent). This highlights the close relationship between the tasks of *Hallucination Detection* and *Natural Language Inference*.

3.4 Strategies comparisons

The *baseline* strategy, which utilizes all available labelled gold data, establishes a lower bound in the expected performance. Both *C-RLFT* and *sequential training* exhibit substantial performance

[2]We use *deberta-xlarge* and *deberta-xlarge-mnli* as encoders, *TheBloke/SOLAR-10.7B-Instruct-v1.0-GPTQ* for pseudo labelling.

Model	Method	F_1 score	Precision	Recall	Accuracy
DeBERTa	Baseline	0.6207±0.0808	0.7112±0.0661	0.5588±0.1562	0.7254±0.0231
DeBERTa	C-RLFT	0.6182±0.0857	0.8081±0.0939	0.5089±0.1574	0.7476±0.0245
DeBERTa	Sequential	0.7075±0.0394	**0.8169±0.0396**	0.6253±0.0690	**0.7898±0.0194**
DeBERTa	Ensemble	**0.7119±0.0272**	0.7918±0.0402	**0.6474±0.0466**	0.7867±0.0171
D.+MNLI	Baseline	0.7138±0.0253	0.7420±0.0319	**0.6882±0.0372**	0.7753±0.0178
D.+MNLI	C-RLFT	0.6146±0.0917	**0.8410±0.0706**	0.4900±0.1376	0.7528±0.0302
D.+MNLI	Sequential	0.7320±0.0229	0.8177±0.0233	0.6628±0.0329	**0.8024±0.0141**
D.+MNLI	Ensemble	**0.7371±0.0223**	0.8016±0.0347	0.6829±0.0425	0.8017±0.0143

Table 2: Performance metrics for DeBERTa and DeBERTa + MNLI models. Best scores are highlighted in bold, and second-best are underlined.

src	hyp	tgt	Target $\mathbf{p(hal)}$	$\hat{\mathbf{p}}\mathbf{(hal)}$
Король Харальд Гормссон, более известный как Харальд Синезубый, ввёл в Дании христианство.	King Harald Hormsson, better known as Harald Sinezubii, introduced Christianity to Denmark.	King Harald Gormsson, better known as “Harald Bluetooth", introduced Christianity to Denmark.	0.40	0.40
Why’d you got to go and do that?	Why did you have to go do that?	Why would you say that?	0.00	0.91

Table 3: Examples of correctly and wrongly predicted as *“Hallucination”* or *“Not Hallucination”*. The model output is *p(hal)* and must be confronted with gold *p(hal)*. The first example proposed is a Russian to English *Machine Translation* (MT), and the second is an English *Paraphrase Generation* (PG).

improvements.

Regarding the *ensemble* strategy, the results in terms of F_1 score outperform individual techniques. We observe a trade-off where the precision of the final result is slightly compromised in exchange for an improved recall. This suggests that the *ensemble* effectively identifies instances of hallucination overlooked by the standalone approaches. These advantages are consistent across both backbones implementations, with and without the additional MNLI fine-tuning.

In a setting where detected hallucinations are shown to the final user with a warning, we argue that the recall is a metric of greater interest (w.r.t. precision). A false negative could be potentially harmful since final users are not warned of the presence of possible hallucinations. A false positive would raise a warning that may be inspected by the final user and safely ignored.

The weights learned for the ensemble layer, based on Equation 1, are $W = (0.52, 1.7, 1.82)$, $b = -1.7$. This shows how both C-RLFT and the sequential models are weighted similarly and more heavily w.r.t. the baseline. The baseline is assigned a non-zero weight: it is considered, although to a lesser extent, in the final vote. The negative bias implies a learned prior: without further knowledge, the initial prediction is of a negative one (i.e., the majority class).

3.5 Qualitative Example

Table 3 demonstrates the effectiveness of the applied strategies through some qualitative examples. We specifically showcase the sentences with the minimum (first row) and maximum (second row) errors.

The first instance depicts a partial hallucination, attributed to the transliteration of “Sinezubii” instead of the translation “Bluetooth,” which is absent from the translation hypothesis. In the second example, despite a paraphrased similarity between the source and hypothesis, the target introduces an action (“saying”) not present in the source (“doing”). As such, we argue that this might be a case of incorrectly labelled ground truth.

4 Conclusions

This work tackles the SHROOM Task 6 challenge at SemEval 2024, focusing on semantic hallucination in NLG models. We propose an automatic

pipeline for hallucination detection, utilizing data augmentation and an ensemble of three different methodologies. The ensemble of the approaches obtained an accuracy of 80.07% in the task's leaderboard. Particular attention should also be paid to the results obtained with the novelty method *sequential*, which was able to outperform the results of the other two methods due to the proposed sequential training.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

David Freeman Engstrom and Jonah B Gelbach. 2020. Legal tech, civil procedure, and the future of adversarialism. *U. Pa. L. Rev.*, 169:1001.

Flavio Giobergia, Luca Cagliero, Paolo Garza, Elena Baralis, et al. 2020. Cross-lingual propagation of sentiment information based on bilingual vector space alignment. In *EDBT/ICDT Workshops*, pages 8–10.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jianguo Jia, Wen Liang, and Youzhi Liang. 2023. A review of hybrid and ensemble in deep learning for natural language processing. *arXiv preprint arXiv:2312.05589*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti, Elena Baralis, et al. 2023. baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy. In *CEUR Workshop Proceedings*. CEUR.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2024. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 Task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey.

Harry Surden. 2020. The ethics of artificial intelligence in law: Basic questions. *Forthcoming chapter in Oxford Handbook of Ethics of AI*, pages 19–29.

Eugene Volokh. 2023. Chatgpt coming to court, by way of self-represented litigants. *The Volokh Conspiracy*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5075–5086.