UDC 004.8
DOI: 10.15587/2706-5448.2026.367900

Oleksandr Konduforov,
Kyrylo Smelyakov,
Yuri Romanenkov,
Anastasiya Chupryna,
Natalia Valenda

ENHANCING MISINFORMATION DETECTION QUALITY ON SMALL DATASETS USING SYNTHETIC LABELING

The object of research is the process of automated misinformation detection, with particular emphasis the use of large language models (LLMs) as scalable annotators to solve a problem of labeled data scarcity (typical for low-resource languages). Manual fact-checking and annotation are costly, slow, and difficult to scale, which limits the applicability of machine learning approaches. The research addresses the problem of enhancing the misinformation detection quality on small datasets by adding the LLM-generated synthetic labels to the training process and evaluates the achieved performance improvements.

The results demonstrate that incorporating LLM-generated annotations into limited training data consistently improves classification performance. The observed gains range from approximately 2 to 3 percentage points in the F1-macro score depending on the configuration. The relative improvement is higher in scenarios with smaller manual data used.

These findings can be explained by the increased volume and diversity of training data provided by synthetic annotations. LLMs provide good fact-checking across more data samples, enriching the dataset and helping the classifier generalize better. However, factors such as LLM version and size, inclusion of additional context, or the use of external knowledge retrieval influence annotation accuracy.

In practical applications, the results suggest that LLM-based annotation can serve as an efficient and scalable method for bootstrapping small datasets. This approach is especially valuable for underrepresented languages, where annotated datasets and benchmarks are scarce, as it can increase misinformation detection accuracy without expensive manual annotation.

Keywords: *misinformation detection, fake news, text classification, machine learning, Large Language Models.*

Received: 01.05.2026
Received in revised form: 11.07.2026
Accepted: 21.07.2026
Published: 31.10.2026

© The Author(s) 2026
This is an open access article under the Creative Commons CC BY license https://creativecommons.org/licenses/by/4.0/

How to cite
Konduforov, O., Smelyakov, K., Romanenkov, Y., Chupryna, A., Valenda, N. (2026). Enhancing misinformation detection quality on small datasets using synthetic labeling. Technology Audit and Production Reserves, 5 (2 (91)), 6–12. https://doi.org/10.15587/2706-5448.2026.367900

1. Introduction

The rapid growth of digital media and social platforms has significantly increased the spread of misinformation and misleading claims. Various political and social groups use disinformation tactics to manipulate elections, spread propaganda in information wars, or influence public opinion through fake news. Empirical researches show that false information spreads significantly faster and reaches larger audiences than truthful information in online networks, especially in political contexts [1].

Early research in misinformation detection primarily relied on manual fact-checking workflows performed by professional journalists and fact-checking organizations. However, the rapid growth of online content makes manual verification difficult to scale. As a result, misinformation detection began to use a machine learning approach, in which models automatically classify claims or news articles as truthful or misleading based on textual features and contextual signals [2]. Later, advances in deep learning have significantly improved the effectiveness of automated misinformation detection systems. In particular, Transformer-based language models such as *BERT* [3] and *RoBERTa* [4] have demonstrated strong performance across a wide range of natural language processing tasks. These models rely on large-scale pretraining followed by supervised fine-tuning on task-specific datasets, allowing them to capture complex semantic and contextual relationships in text. Obtaining such datasets still remains a major challenge, especially for low- and medium-resource languages. Fact-checking datasets typically require expert human annotation, which is expensive, time-consuming, and difficult to scale. Recent surveys on misinformation detection show that most of the research and available datasets focus on high-resource languages, while low-resource languages remain underrepresented [5]. In addition to annotation-based methods, recent work such as [6] demonstrates that hybrid clustering of semantically vectorized Telegram data can uncover coherent thematic structures even in small datasets, highlighting the potential of semantic representations for low-data scenarios.

To address the challenges associated with manual data annotation, researchers have explored alternative approaches for generating labeled datasets. Early work in weak supervision proposed labeling strategies that combine multiple noisy labeling sources to generate training labels. More recently, advances in large language models (LLMs) have opened new opportunities for automated data annotation. Modern LLMs demonstrate strong capabilities in reasoning, semantic understanding, and text classification, allowing them to perform complex tasks with minimal supervision. One of the recent researches demonstrate that LLMs achieved higher annotation quality compared to crowd workers and trained annotators for several annotation tasks, including relevance, stance, topics, and frame detection [7]. Another research used a few-shot chain-of-thought approach with the self-generated explanation and showed that it performs on par or even better with crowdsourced annotators for different tasks [8]. In research [9], the authors evaluated the performance of multiple LLMs and different in-context learning techniques for misinformation detection in Ukrainian language and highlighted both the strong sides and

limitations of using LLMs for this task. These findings suggest that LLMs could serve as scalable annotators capable of generating large labeled datasets at significantly lower cost.

Several recent researches have investigated the use of LLM-generated labels for training downstream machine learning models. They demonstrated that modern LLMs can produce annotations close to human quality for some NLP tasks and that models trained on LLM-generated labels can achieve performance comparable to models trained on human-annotated datasets or even outperforming them [10–12]. However, at the same time, these and other researches indicate that synthetic labels may introduce biases or inconsistencies that negatively affect model performance, especially when annotation quality is not carefully controlled [13].

These mixed results highlight the need for a systematic investigation into the role of synthetic annotations in training machine learning models.

The object of research is the process of automated misinformation detection, with particular emphasis the use of LLMs as scalable annotators to solve a problem of labeled data scarcity (typical for low-resource languages).

The aim of research is to enhance the misinformation detection quality on small datasets by adding the LLM-generated synthetic labels to the training process and measure the achieved performance improvements. This will make it possible to better understand the potential and limitations of using LLMs as scalable annotation tools for misinformation detection in low-data setup.

To achieve the aim, the following objectives were set:

– to generate the synthetic labels using several LLM-based approaches and evaluate the quality of the annotation;

– to test low-data training scenarios by training the classification models on mixed manual/synthetic datasets of different configurations and compare the achieved performance improvements.

2. Materials and Methods

2.1. Research methods

The task addressed in this research is binary misinformation classification, in which the goal is to determine whether a text is truthful or misleading. Given an input statement x, a classification model must predict a binary label $y \in \{True, False\}$. The task focuses exclusively on textual information, without incorporating external knowledge sources or multimodal features. This design allows isolating the impact of training data quality and quantity on the classifier performance.

The following research methods were employed in this research: supervised learning for classification; weak supervision/data augmentation using LLM annotation with multiple prompting strategies; controlled experimental design with varying proportions of manual and synthetic data; and comparative statistical evaluation using macro-averaged $F1$-score to assess both annotation quality and classification performance.

Weak supervision/data augmentation using LLM annotation with multiple prompting strategies was used to generate synthetic labels using different approaches.

Supervised learning for classification was used for training the transformer-based model for misinformation detection in different configurations. To make sure that the trained models can be compared correctly, the controlled experimental design was used with varying proportions of manual and synthetic data. The results were analyzed using comparative statistical evaluation using macro-averaged $F1$-score.

2.2. Dataset

The aim of the research requires a well-labeled dataset of significant volume to validate the approach. Because of that it was decided to use the *LIAR2* dataset [14], an improved and extended version of the well-known *LIAR* dataset [15]. The dataset consists of short political statements collected from the fact-checking website PolitiFact and labeled by professional journalists which proves its reliability for the goal of the research. Another important factor is the dataset size: it contains about 23000 samples split on train, validation and test sets in 8:1:1 proportion, which is enough to organize the planned controlled experiments and evaluate them with good level of confidence.

Each statement is assigned one of six labels: *pants-on-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. In addition to the statement text, the dataset contains metadata, including information about the speaker, the statement's context, the publication date, and calculated statistics on the speaker's trustworthiness.

Since the aim of this work is to study the effect on the binary classification task, the original six-class labeling scheme is converted into a binary one. The basic 3–3 mapping is used, where the labels *true*, *mostly-true*, and *half-true* are mapped to the *True* class, while *barely-true*, *false*, and *pants-on-fire* are mapped to the *False class*. This allows focusing on the reliability of the annotations and the impact of synthetic data augmentation rather than on multi-class decision boundaries.

2.3. LLM-based annotation

To investigate whether synthetic annotations can improve model performance in low-data settings, the large language models (LLMs) are used as automatic annotators for generating training labels. The goal of this process is to simulate a realistic scenario in which only a limited number of human-labeled samples are available, while additional data can be automatically labeled using modern generative models.

Two LLMs were used for the annotation process: *GPT-4.1 mini* (OpenAI, USA) and *Gemini 3.1 Flash-Lite* (Google, USA). These models were selected because they provide a good balance between inference cost, response latency, and reasoning capability, making them suitable candidates for large-scale annotation tasks.

For each statement in the dataset, the LLM was asked to determine whether the claim should be classified as *True*, *Mostly True*, *Mostly False* or *False*. The annotation prompt instructs the model to analyze the statement neutrally and evaluate its factual claims. The model then produces a label that is converted to binary and used as the synthetic annotation for that sample.

To analyze the influence of contextual information on annotation quality, two prompting strategies are evaluated:

1. *Statement-only prompting.* The annotation model receives only the textual claim as input and must determine whether the statement is true or false.
2. *Extended context prompting.* In addition to the statement, the annotation model receives auxiliary metadata available in the dataset, such as the speaker, context, and the statement's date. This additional information may help the model interpret the claim more accurately.

For the *Gemini 3.1 Flash-Lite* model, an additional configuration is tested. It enables grounded responses, allowing the model to use *Google Search* (Google, USA) to evaluate claims. This setup approximates a lightweight fact-checking workflow in which the model may retrieve relevant knowledge before generating a verdict.

The annotation process is performed prior to classifier training. Each selected statement is sent to the LLM along with the corresponding prompt, and the model's output is converted into a binary label.

To ensure transparency of the generated annotations, each LLM's annotation accuracy is measured by comparing its predicted labels with the ground-truth labels in the original dataset. This evaluation allows estimating the reliability of LLM-based annotation and to analyze how annotation quality affects the downstream performance of the misinformation detection model.

The resulting synthetic annotations are combined with human-labeled samples to form multiple training configurations with different ratios of manual and synthetic data, which are evaluated in the experiments described in the following sections.

ISSN-L 2664-9969; E-ISSN 2706-5448

2.4. Experimental setup

To evaluate the impact of synthetic annotations on misinformation detection performance, a series of controlled experiments is conducted which vary the amount of manually annotated data and LLM-generated annotations used during training. The goal is to simulate realistic low-data scenarios and analyze how synthetic labels affect classifier performance.

The experiments are based on three sizes of manually annotated training data: *500, 1000, and 2000 samples*. These subsets are randomly selected from the *LIAR2* training set using the same random seed while preserving the class balance between *True* and *False* labels. Each subset represents a different level of data scarcity. To ensure the comparability of the experiments, each bigger dataset contains a smaller one but adds additional data on top of it. So, dataset of size 1000 contains dataset of size 500 plus additional 500 samples, and same approach is used for the dataset with 2000 samples.

For mixed-label scenarios, the additional synthetic annotations are generated which use the LLM-based procedure described in Section 2.2. Two synthetic dataset sizes are considered: *2000 and 4000 samples*. Similar to manual datasets, the dataset with 4000 samples consists of a smaller dataset with 2000 samples, and then additional 2000 ones. The synthetic labeled data is combined with the manually labeled data to form the final training datasets. The process of preparing the mixed dataset with manual and synthetic labels and using it for model training is shown in Fig. 1.

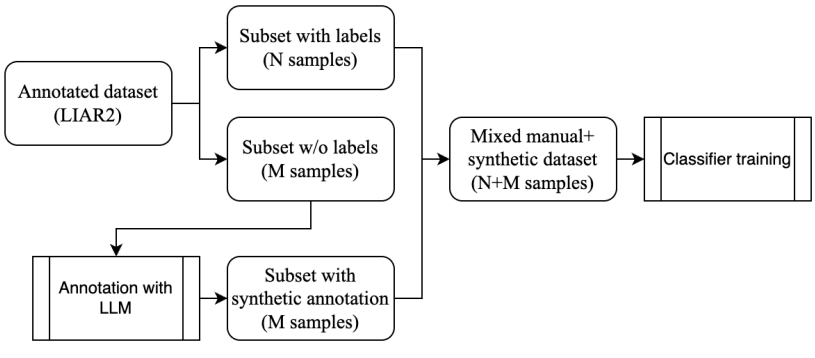


Fig. 1. The process of preparing the mixed train dataset with manual and synthetic labels

This setup allows exploring several key experimental configurations that vary the ratio between human and synthetic annotations. In particular, it is possible to evaluate scenarios in which a small number of manually labeled samples is supplemented with a larger volume of LLM-generated labels, or in which the dataset contains an equal number of manually and LLM-generated labels.

In addition to training datasets with mixed labels, the baselines and toplines are established by training models only on original manual labels to set a performance floor and ceiling for comparison with mixed-label training. Baseline results are achieved by training the classification model on only a manual part of the mixed dataset. Topline results are achieved by training the classification model on the same samples from the mixed dataset of the respective size but using the original manual labels. Table 1 summarizes the dataset configurations used in the experiments.

All models are trained and evaluated on the same validation and test sets as the original LIAR2 dataset to ensure a fair comparison across experimental conditions. This design isolates the effect of training data composition while keeping the evaluation protocol consistent across all experiments.

In addition to varying dataset size, it is also analyzed how annotation quality influences downstream performance. Synthetic labels are generated using two LLM annotators and two data inputs (pure statement texts and statement texts with additional context), with one additional annotation performed using grounding mode, enabling the comparison of the different annotation approaches.

Table 1

List of dataset configurations used in the experiments

Manual labels	Synthetic labels	Total train size	Manual labels %	Scenario
500	0	500	100%	baseline for 500 samples
500	2000	2500	20%	mixed for 500 + 2000 samples
2500	0	2500	100%	topline for 500 + 2000 samples
500	4000	4500	11.1%	mixed for 500 + 4000 samples
4500	0	4500	100%	topline for 500 + 4000 samples
1000	0	1000	100%	baseline for 1000 samples
1000	2000	3000	33.3%	mixed for 1000 + 2000 samples
3000	0	3000	100%	topline for 1000 + 2000 samples
1000	4000	5000	20%	mixed for 1000 + 4000 samples
5000	0	5000	100%	topline for 1000 + 4000 samples
2000	0	2000	100%	baseline for 2000 samples
2000	2000	4000	50%	mixed for 2000 + 2000 samples
4000	0	4000	100%	topline for 2000 + 2000 samples
2000	4000	6000	33.3%	mixed for 2000 + 4000 samples
6000	6000	6000	100%	topline for 2000 + 4000 samples

Each experimental configuration is trained and evaluated independently using the same model architecture and training procedure. This controlled setup enables systematic analysis of three key factors: the amount of manually labeled data, the volume of synthetic annotations, and the quality of LLM-generated labels.

2.5. Classification model

To evaluate the effect of synthetic annotations on misinformation detection performance, a transformer-based text classification model is trained. As the primary classifier, it was decided to use a base *RoBERTa*, a transformer architecture that improves upon the original BERT model through optimized training procedures and larger scale pretraining.

RoBERTa was selected because it consistently achieves strong performance on a wide range of natural language understanding tasks, including text classification and misinformation detection. RoBERTa provides greater representational capacity and stronger contextual understanding, which may be beneficial for learning from noisy or partially synthetic annotations. This makes it suitable for evaluating whether LLM-generated labels can effectively support the training of more powerful classification models.

The model takes the textual claim as input and produces a probability distribution over two classes: *True* and *False*. Input statements are tokenized with the standard RoBERTa tokenizer and truncated to the model's maximum sequence length when necessary. The tokenized input is then processed by the transformer encoder, and the contextual representation of the special classification token is used as the input to the classification layer.

The classification head consists of a linear layer followed by a softmax activation that outputs class probabilities. Model parameters are optimized using the cross-entropy loss function for binary classification.

All models are trained with the same configuration across experimental setups to ensure a fair comparison of different training datasets. The training procedure uses the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 4, and gradient accumulation to stabilize training on small datasets. Each model is trained for several epochs with evaluation performed on a validation set to monitor performance and prevent overfitting.

2.6. Evaluation metrics

To assess the performance of both the annotation process and the misinformation detection model, standard evaluation metrics commonly used in binary classification tasks are employed.

The primary evaluation metric used in this research is the *macro-averaged F*1 *score (F*1*-macro)*. This metric is particularly suitable for misinformation detection tasks because it evaluates model performance equally across both classes, regardless of their relative frequency in the dataset.

To compute the macro-averaged *F*1-score, the *F*1-score is first calculated separately for each class, and the final score is obtained as the arithmetic mean of these values

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} F1_i, \quad (1)$$

where N – the number of classes.

The same evaluation metric is also used to estimate the quality of LLM-generated annotations. In this case, the labels predicted by the LLM annotators are compared with the ground-truth labels in the original dataset. This evaluation allows quantifying the reliability of synthetic annotations and analyze how annotation quality influences downstream model performance.

3. Results and Discussion

3.1. Generation of synthetic LLM-based labels

This section evaluates the quality of synthetic annotations generated by the two LLM annotators: GPT-4.1 mini and Gemini 3.1 Flash-Lite. The models were used to label the LIAR2 dataset under two input configurations: statement-only and extended-context, in which the input additionally includes the author, the statement's context, and the date. In addition, the Gemini 3.1 model was used in two modes on extended-context input: normal and grounding with Google Search.

For the data annotation purpose and further downstream model training, a subset of 6000 samples was selected randomly from the training set of the LIAR2 dataset. The original LIAR2 dataset is moderately imbalanced, with slightly more false than true statements, therefore selected subset is also imbalanced (3463 vs. 2537, or 57.7% vs. 42.3%).

The evaluation compares the LLM-generated annotations with the original dataset labels using the *F*1-macro and Accuracy metrics. In addition to standard metrics, it was decided to also analyze the distribution of predicted labels, as differences in class distributions may affect the training dynamics of downstream classifiers. Table 2 summarizes the quality of the annotation.

Table 2

Annotation results for experimental models and input types

Model	Prompting mode	Predicted false	Predicted true	Accuracy	*F*1-macro
GPT-4.1 mini	Statement only	3052	2942	69.35	69.12
Gemini 3.1 Flash Lite	Statement only	3763	2231	71.54	70.29
GPT-4.1 mini	Extended context	3032	2964	72.46	72.27
Gemini 3.1 Flash Lite	Extended context	3699	2297	74.98	74.0
Gemini 3.1 Flash Lite (grounding)	Extended context	3846	2150	84.29	83.54

In normal mode, both LLM annotators achieve similar levels of annotation accuracy, with *F*1-macro scores ranging from 69.12 to 74.0. Adding contextual information consistently improves performance for both models, resulting in higher accuracy and *F*1-macro values.

However, in grounding mode, the Gemini 3.1 model demonstrates substantially higher *F*1-macro score: 83.54 vs. 74 (+9.5 pp). This is an effect of additional fact-checking that the model can obtain from using Google Search for each specific sample.

It should also be noted that although the aggregate metrics for normal mode are relatively close, the prediction distributions differ substantially between the two models. The annotations generated by GPT-4.1 mini are relatively balanced between the two classes. In both prompting configurations, the model produces a similar number of *True* and *False* predictions. In contrast, Gemini 3.1 Flash Lite shows a noticeably different behavior: the model consistently produces significantly more false labels, resulting in a distribution more skewed toward the negative class. Comparing these distributions with that of the original labels, it was concluded that both models exhibit opposite biases. Interestingly, Gemini 3.1 demonstrates even higher bias in grounding mode despite better fact-checking capabilities.

The results highlight an important consideration when using LLMs as annotators. Even when models achieve similar overall performance metrics, their label distributions may differ significantly. Such differences can influence the characteristics of the resulting synthetic datasets and potentially introduce bias during classifier training.

In general, the achieved results confirm results from other researches that LLMs can be used as strong classifiers in different domains. In addition to that, this research also highlights specific biases which LLMs demonstrate for the misinformation detection task.

3.2. Simulating low-data training scenarios using mixed datasets

This section provides the analysis whether synthetic annotations generated by LLMs improve the performance of misinformation detection models. The experiments compare models trained only on manually labeled data with models trained on a mix of manual and synthetic annotations. The manually annotated labels from the LIAR2 dataset serve as the baseline and topline, as they were produced by professional fact-checkers and are assumed to be the most reliable.

The evaluation metric used for comparison is *F*1-macro, which ensures balanced evaluation across both classes as described in Section 2.6.

First, let's examine the performance of the baseline and topline models trained solely on manually labeled data. Table 3 shows the classification scores for different training dataset sizes. The first and second columns present the number of base samples and corresponding classification scores (baseline results). The third and fourth columns show the classification scores for the extended datasets: the same base samples plus an additional 2000 or 4000 samples, respectively (topline results). Baseline results represent the classification scores for configurations with limited data, while topline results show the expected maximum achievable score as additional samples with manual labels are added.

Table 3

Classification results for dataset configurations using manual labels

Number of base samples	Base only, *F*1-macro	Base + 2000, *F*1-macro	Base + 4000, *F*1-macro
500	65.30	68.28	68.94
1000	65.40	68.85	69.22
2000	66.23	68.95	70.76

These results confirm the expected trend that larger manually labeled datasets lead to better classifier performance, although some improvements remain relatively modest.

Adding synthetic annotations generated by LLMs consistently improves performance across most configurations compared to the baseline, as shown in Table 4.

ISSN-L 2664-9969; E-ISSN 2706-5448

Table 4

Experimental results for all configurations

Dataset config (manual/synthetic labels)	Baseline, manual labels	Mixed labels, GPT-4.1 mini annotation		Mixed labels, Gemini 3.1 Flash-Lite annotation			Topline, manual labels
		Statement only	Extended context	Statement only	Extended context	Extended context + grounding	
500/2000	65.30	67.56	68.22	68.10	67.87	68.09	68.28
500/4000	65.30	66.87	68.59	67.60	68.52	68.63	68.94
1000/2000	65.40	67.50	68.09	68.15	68.65	68.41	68.85
1000/4000	65.40	67.09	69.10	68.38	68.19	69.06	69.42
2000/2000	66.23	67.68	68.59	66.94	68.84	68.81	68.95
2000/4000	66.23	68.19	68.62	67.86	68.21	68.88	70.26
Average	65.64	67.48	68.54	67.83	68.38	68.65	69.12

To simplify the quality comparison with the baseline and topline results for manual samples, each table row contains the baseline result, the results on mixed labels obtained from different annotation approaches, and the topline result. Results are demonstrated as *F*1-macro scores. The topline column shows the scores achieved at training on all manual labels of the corresponding dataset size. For example, the *F*1-macro score for the "500/2000" configuration in this column represents the classification performance for the model trained on the same 2500 samples but using all manual labels. This provides a meaningful comparison of experiments that use datasets with all manual and mixed labels of the same samples. Last row (Average) represents average results across all dataset configurations.

The experimental results demonstrate that classification performance can be increased noticeably across all training dataset configurations, with gains ranging from 1.5 to 3.5 pp in *F*1-macro score.

The best performance for all manual labels is achieved with the biggest dataset size (6000), as expected. The best performance for the mixed dataset was achieved with the "1000/4000" configuration, using GPT-4.1 mini annotation with extended context.

Given the observed performance variability across mixed configurations, aggregating the results from mixed datasets across different annotation approaches provides a more stable comparison. These aggregated findings are presented in Table 5. Last row (Average) represents average results across all dataset configurations.

Table 5

Aggregated results for experiments

Dataset config (manual/synthetic labels)	Baseline, manual labels	Mixed labels, aggregated (mean, SD)	Topline, manual labels
500/2000	65.30	67.97 ± 0.26	68.28
500/4000	65.30	68.04 ± 0.78	68.94
1000/2000	65.40	68.16 ± 0.43	68.85
1000/4000	65.40	68.36 ± 0.82	69.42
2000/2000	66.23	68.17 ± 0.83	68.95
2000/4000	66.23	68.35 ± 0.40	70.26
Average	65.64	68.18 ± 0.59	69.12

When analyzing the aggregated results of the experiments, the following observations can be made:

1. *Impact of synthetic labels.* Aggregated results for mixed labels demonstrate noticeable performance improvements from adding synthetic labels across all dataset configurations, achieving average *F*1-macro score improvement about +2.5 pp. At the same time these results are lower than topline results achieved with similar amounts of purely manual labels by about –1 pp on average, which is an expected result. For easier comparison, these results are visualized in Fig. 2.

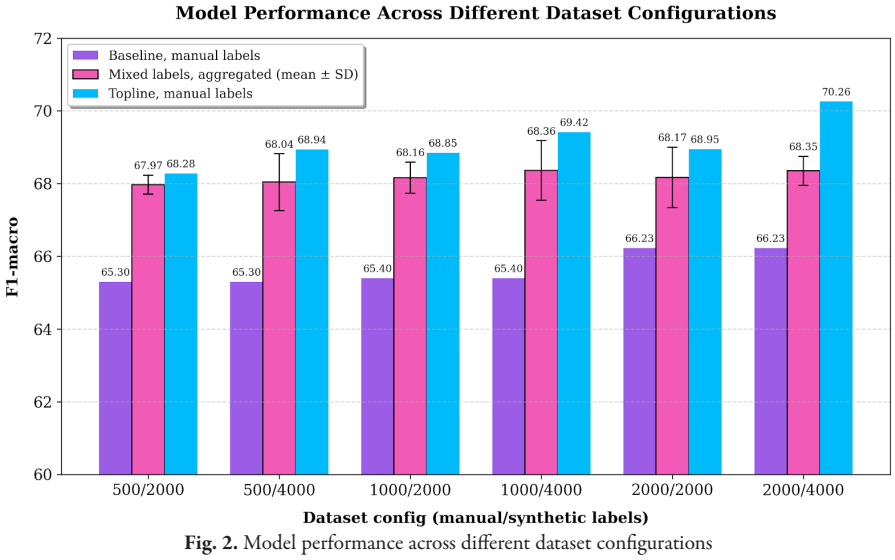


Fig. 2. Model performance across different dataset configurations

2. *Effect from quantity of manually labeled samples.* Adding synthetic labels improve performance for all manual dataset sizes, but gains drop from almost +3 pp to +2 pp, when the amount of manual data increase from 500 to 2000 samples. This suggests limited impact from adding synthetic labels when there is a higher number of manual ones.

3. *Impact of the quantity of synthetic labels.* Comparing results from adding 2000 and 4000 samples with synthetic labels shows that adding more samples slightly improves model performance (about +0.15 pp across three configuration experiments). This suggests that even though adding synthetic labels demonstrate solid improvement overall, there is likely a saturation point after which the improvement is not happening.

4. *Difference between annotation LLMs.* Both LLM annotators produce competitive results in downstream classification, which is consistent with the annotation quality results outlined in Section 3.1. In some experiments, the best results are achieved with GPT-4.1 mini using extended context annotation, while in others the Gemini-based annotations produce better performance.

5. *Effect of providing input context for annotation.* Annotations generated from extended-context inputs outperform those generated from statement-only inputs. The average improvement across configurations reaches +1 pp. This confirms that providing additional contextual information helps the LLM annotators produce more reliable labels and, as a result, improve classification accuracy.

6. *Effect of using grounding for annotation.* As outlined in Section 3.1, the Gemini 3.1 model demonstrates a substantially higher annotation quality in grounding mode compared to other annotation approaches. However, results of the downstream classification show that these improvements translate into very modest classification impact, which is similar to using extended-context inputs.

3.3. Discussion

The results of the experiments demonstrate that LLM-generated annotations can effectively augment small fact-checking datasets and improve misinformation detection performance. Across all experimental configurations, training datasets that combined original manual labels with synthetic ones achieved higher average *F*1-macro scores than models trained only on a manually annotated samples. The potential explanation for this effect is that the classifiers benefit from the additional linguistic variability and training signal provided by the synthetic labels.

The experiments show that synthetic annotations are most beneficial in low data scenarios. Best relative improvements are achieved when synthetic annotations are added to small sets of manually labeled data and the effect decreases when larger manual datasets are available. This suggests that the classification models already receive a sufficiently strong training signal, and additional synthetic data contributes less to performance gains, and indicates that LLM-based annotation may be particularly valuable for bootstrapping small datasets.

Experiments on improving annotation quality to enhance downstream classification performance demonstrate mixed results. Adding contextual information to statements slightly improves both annotation and classification accuracy. At the same time, using grounding with the Google Search significantly improves annotation accuracy, but it translates into minor improvements in classification accuracy. Most likely, this is happening because classification accuracy for the non-grounding annotation is already comparable to the topline accuracy achieved in training on purely manual labels, leaving little room for improvement.

The results from this research generally correlate with the ones from similar researches: first, that LLMs can be used as reliable annotators, and second, that classifiers trained on purely LLM-generated or mixed annotation, can produce strong results comparing to manual annotation. The important addition of this research is that it confirms these effects specifically for misinformation detection task and suggests and successfully verifies a method to use synthetic LLM-generated annotation to improve the classification results on small datasets.

Several limitations of this research should be acknowledged. First, the research is based on a single dataset, which the authors used as a reliable benchmark, whereas experiments on other datasets can bring different results. Second, the experiments use a relatively simple classification model. While RoBERTa provides a strong baseline for text classification, the impact of synthetic annotations may differ when using more sophisticated model architectures.

Another limitation is that the research evaluates a limited set of experimental configurations. It would be interesting to extend it to larger amounts of manual and synthetic labels to check when a saturation from manual training data is achieved and there is no improvement from adding any synthetic labels.

An obvious extension of this work would be to apply the proposed approach to misinformation datasets in low-resource languages, such as Ukrainian. This would allow evaluation of whether LLM-based annotation can effectively bootstrap datasets in languages with scarce manually annotated resources.

Another promising direction is to investigate quality filtering techniques for synthetic annotations. For example, labels could be accepted only when multiple LLM annotators agree, or when the model expresses high confidence in its prediction.

Finally, further experiments could examine the potential of more sophisticated classification models that better leverage additional signals in higher-quality synthetic annotations. This could unlock the potential of techniques like grounding with Google Search or stronger reasoning models for annotation, achieving higher accuracy for mixed labels.

4. Conclusions

1. Five sets of synthetic labels were generated using two different LLMs (GPT-4.1 mini and Gemini 3.1 Flash Lite) from the LIAR2 dataset. Two different prompting modes were used: using only statements and using statements with extended context. In addition to that, a special grounding mode was used for Gemini 3.1 Flash Lite which allows LLM to use Google Search to evaluate claims before making a conclusion. The performance of the annotation was evaluated for each generated set of labels. Both LLMs achieved similar levels of annotation quality, with *F*1-macro scores ranging from 69.12 to 74.0 for different prompting modes. The highest quality was achieved for Gemini 3.1 Flash Lite in grounding mode where *F*1-macro score reached 83.54, which clearly shows that additional search context can significantly improve the fact-checking capabilities of the LLM. The evaluation also demonstrated the opposite biases of the used LLMs toward making more true or false predictions.

2. A series of controlled experiments of training the classification models using mixed manual/synthetic datasets was conducted to test multiple low-data training scenarios. The same transformer-based classifier (RoBERTa) and same training configurations were used in all experiments. The models were trained on mixed datasets with different proportion of manually annotated (original) and LLM-generated synthetic samples. For each dataset configuration, all five sets of synthetic labels generated at the previous step, were used. In addition to that, a set of baseline and topline models were trained to estimate the range of the expected model performances for each dataset size. The achieved performances for each model were evaluated using the test set of the LIAR2 dataset. The experiments confirmed that LLM-generated labels improve classification performance across almost all tested configurations. They also revealed that providing additional contextual information and using built-in search techniques (grounding) improves annotation quality, but the effect of improving the quality of annotation on the downstream classification accuracy is limited. In addition to that, the results also indicate that simply increasing the amount of synthetic data does not always guarantee further gains. Overall, the findings suggest that LLM-based annotation is a promising approach for augmenting small misinformation datasets. This approach may be particularly useful for augmenting fact-checking datasets in low-resource languages, where large-scale manual annotation is not present.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research, and its results presented in this paper.

Financing

The research was performed without financial support.

Data availability

Data will be made available on reasonable request.

ISSN-L 2664-9969; E-ISSN 2706-5448

Use of artificial intelligence

The authors declare that they used AI tools in the following cases:
– ChatGPT (GPT-5.1) was used as a search tool to explore the research area and find the relevant papers in Deep Research mode. The authors carefully reviewed each paper manually and selected those relevant to the article's topic. The relevant papers were cited in the Introduction and References sections, but did not affect the experimental results.
– Gemini 3.5 Flash was used to generate a Python code to draw a chart for model performance across different dataset configurations (Fig. 2) using the data from Table 5.
– Grammarly was used to fix mistakes and improve the text clarity and style. The authors carefully reviewed the recommendations provided by the tool to ensure they didn't alter the meaning of the study or the experimental results.
The authors take full responsibility for the final version of the article. Declaration submitted by Oleksandr Konduforov.

Authors' contributions

Oleksandr Konduforov: Conceptualization, Methodology, Investigation, Writing – original draft; ***Kyrylo Smelyakov***: Formal analysis, Supervision, Project administration; ***Yuri Romanenkov***: Formal analysis, Validation; ***Anastasiya Chupryna***: Data curation, Writing – review and editing; ***Natalia Valenda***: Data curation, Validation, Writing – review and editing.

All authors have read and agreed to the published version of the article and are jointly responsible for the reliability, accuracy, and integrity of the data and research results presented in the article.

References

1. Vosoughi, S., Roy, D., Aral, S. (2018). The spread of true and false news online. *Science, 359 (6380),* 1146–1151. https://doi.org/10.1126/science.aap9559
2. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H. (2017). Fake News Detection on Social Media. *ACM SIGKDD Explorations Newsletter, 19 (1),* 22–36. https://doi.org/10.1145/3137597.3137600
3. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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).* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692.* https://doi.org/10.48550/arXiv.1907.11692
5. Rananga, S., Isong, B., Modupe, A., Marivate, V. (2024). Misinformation Detection: A Review for High and Low-Resource Languages. *Journal of Information Systems and Informatics, 6 (4),* 2892–2922. https://doi.org/10.51519/journalisi.v6i4.931
6. Lynnyk, R., Vysotska, V., Hu, Z., Uhryn, D., Diachenko, L., Smelyakov, K. (2025). Information Technology for Modelling Social Trends in Telegram Using E5 Vectors and Hybrid Cluster Analysis. *International Journal of Information Technology and Computer Science, 17 (4),* 80–119. https://doi.org/10.5815/ijitcs.2025.04.07
7. Gilardi, F., Alizadeh, M., Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120 (30).* https://doi.org/10.1073/pnas.2305016120
8. He, X., Lin, Z., Gong, Y., Jin, A. L., Zhang, H., Lin, C. et al. (2024). AnnoLLM: Making large language models to be better crowdsourced annotators. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track).* Association for Computational Linguistics, 165–190. https://doi.org/10.18653/v1/2024.naacl-industry.15
9. Vysotska, V., Chupryna, A., Valenda, N., Konduforov, O. (2025). Evaluating the in-context learning capabilities of large language models for misinformation detection for Ukrainian news. *Proceedings of the Computational Intelligence Application Workshop (CIAW 2025),* 4110, 181–197. Available at: https://ceur-ws.org/Vol-4110/paper15.pdf
10. Pangakis, N., Wolken, S. (2024). Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels. *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024).* Association for Computational Linguistics, 113–131. https://doi.org/10.18653/v1/2024.nlpcss-1.9
11. Tong, Z., Gu, Y., Liu, H., Liu, Q., Wu, S., Shi, H., Zhang, X.-Y. (2025). Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 24276–24290. https://doi.org/10.18653/v1/2025.acl-long.1182
12. Horych, T., Mandl, C., Ruas, T., Greiner-Petter, A., Gipp, B., Aizawa, A., Spinde, T. (2025). The promises and pitfalls of LLM annotations in dataset labeling: A case study on media bias detection. *Findings of the Association for Computational Linguistics: NAACL 2025.* Association for Computational Linguistics, 1370–1386. https://doi.org/10.18653/v1/2025.findings-naacl.75
13. Lu, Y., Smith, K. (2025). Feeding LLM annotations to BERT classifiers at your own risk (Version 2). *arXiv:2504.15432.* https://doi.org/10.48550/arXiv.2504.15432
14. Xu, C., Kechadi, M.-T. (2024). An Enhanced Fake News Detection System With Fuzzy Deep Learning. *IEEE Access, 12,* 88006–88021. https://doi.org/10.1109/access.2024.3418340
15. Wang, W. Y. (2017). "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Association for Computational Linguistics, 422–426. https://doi.org/10.18653/v1/P17-2067

✉ ***Oleksandr Konduforov,*** *Engineering Manager, Superhuman, San Francisco, USA, e-mail: oleksandr.konduforov@grammarly.com, ORCID: https://orcid.org/0009-0000-3942-2444*

Kyrylo Smelyakov, *Doctor of Technical Sciences, Professor, Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0001-9938-5489*

Yuri Romanenkov, *Doctor of Technical Sciences, Professor, Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0002-6544-5348*

Anastasiya Chupryna, *Candidate of Technical Sciences, Associate Professor, Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0003-0394-9900*

Natalia Valenda, *Candidate of Technical Sciences, Associate Professor, Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, ORCID: https://orcid.org/0000-0003-3250-6172*

✉ *Corresponding author*