Supplementary Material for CoALFake: Collaborative Active Learning with Human-LLM Co-Annotation for Cross-Domain Fake News Detection

A Algorithm

Algorithm 1 *CoALFake* Pipeline

Input: Unlabelled Data (D_{Pool}), Demonstration Set (D_{demo}), LLM ($\mathcal{P}$), Human Annotators ($\mathcal{H}$)
Output: Trained Domain-Agnostic Classifier (C)
Initialize $D_{\text{labelled}} \leftarrow \emptyset$, round $\leftarrow 1$
while not convergent **do**
    if round $= 1$ **then**
        Select D_{Sample} using Eq. 1.
    else
        Select D_{Sample} using Eq. 2.
    end if
    Retrieve demonstration examples $S = \{(x_i, y_i)\}_{i=1}^{k}$ from D_{demo} using k-NN
    Perform in-context learning as Eq. 3 to label D_{Sample}
    Perform label verification as Eq. 4.
    Send D_{Noisy} to human annotators for re-annotation
    Update $D_{\text{labelled}} \leftarrow D_{\text{labelled}} \cup D_{\text{Sample}}$
    Compute domain embeddings as Eq. 5.
    Train Domain-Agnostic Classifier C using D_{labelled}
    Update D_{demo} with new high-confidence samples
    round $\leftarrow$ round $+1$
end while
Return: Trained Classifier C

B Baselines

We compare our model with several widely used baselines:

- **HPNF** [Shu *et al.*, 2019]: This model extracts multiple features, including structural and temporal characteristics, from a news article's propagation network to form its feature representation. A Logistic Regression classifier is then applied to distinguish between fake and real news. The **HPNF+LIWC** variant enhances this approach by integrating feature vectors from HPNF with those derived from LIWC.
- **AE** [Silva *et al.*, 2020]: This method employs an Auto-Encoder architecture to learn latent representations of news records based on their propagation networks. These representations are subsequently used to identify fake news.
- **SAFE** [Zhou *et al.*, 2020]: A multimodal method for fake news detection, SAFE learns separate latent representations for each modality of a news record while also constructing a joint representation that captures cross-modality insights.
- **EDDFN** [Silva *et al.*, 2021]: This model leverages both domain-specific and cross-domain knowledge in news records to improve fake news detection across diverse domains. Additionally, an unsupervised technique selects a subset of unlabelled but informative news records for manual annotation.
- **MDFEND** [Zhu *et al.*, 2022]: This approach employs domain-specific experts to extract features from news articles, while domain gates assign varying weights to these experts based on their relevance. The final feature vector is obtained by aggregating the outputs of these experts.
- **FuDFEND** [Liang *et al.*, 2022]: This model begins by utilizing the final layer of the BERT Transformer block to transform news articles into word embedding vectors. A GRU module then generates multi-domain tags for each news item. The feature extraction module integrates these multi-domain tag features to construct the final comprehensive feature vector.
- **DITFEND** [Nan *et al.*, 2022]: This model transfers coarse-grained domain-level knowledge by training a general model on data from all domains using a meta-learning approach. To facilitate fine-grained instance-level knowledge transfer, a language model is trained specifically on the target domain. This model evaluates the transferability of each data instance from the source domains and re-weights their contributions accordingly.
- **SLFEND** [Wang *et al.*, 2023]: This model enhances feature extraction through the use of soft labels. A novel Leap GRU mechanism filters out irrelevant words, allowing the membership function module to generate soft labels for each news item. These soft labels aid in extracting multi-domain features, leading to a comprehensive feature representation.

C Parameter Settings

This section examines how modifying the model's hyperparameters influences its performance on fake news detection. Figure 1 shows the model's behaviour for various values of $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5, each representing the weight assigned to a specific loss term. We observe that setting λ_1 either too high or too low degrades performance, suggesting that the L_{recon} loss term should be given a moderate weight relative to the others. Similarly, performance declines when $\lambda_2 < 1$ and when $\lambda_3 > 1$. Additionally, increasing λ_4 beyond 0.1 reduces the F1 score, and keeping λ_5 small is essential to avoid over-clustering.

We examine the sensitivity of the model's performance to other parameters: the latent dimension (d), the number of epochs and the batch size in Figure 2. Overall, the model yields consistent performance for $d > 512$, epochs > 300 and batch size < 128.

D Prompting

We access OpenAI APIs through the Azure service, utilizing GPT-3.5 Turbo as the LLM annotator for our experiments. The temperature is set to 0. Below the prompt used for annotation:

Listing 1: Annotation Prompt.

```
I need your assistance in evaluating the
    authenticity of a news article.
I will provide you the news article. You
    have to answer only with Fake or
   Real.
I will give you some examples of news.
   Your answer after [output] should be
   consistent with the following
   examples:

[example 1]:
[input news]: [news text: {...}]
[output]: [This is {...} news]

[example 2]:
[input news]: [news text: {...}]
[output]: [This is {...} news]

[target news]:
[input news]: [news text: {...}]
[output]
```

E Analysis of Labelling Costs

In this section, we compare the labelling costs of GPT-3 and crowdsourced labelling. For simplicity, we exclude costs related to GPT-3 template selection, human labeler selection and other factors, focusing solely on the cost per label charged by the API or crowdsourcing platform.

The GPT-3.5-turbo API offered by OpenAI charges based on the number of tokens used for encoding and generation. According to OpenAI's pricing[1], the cost is \$3.00 for input and \$6.00 for output per 1 million tokens. Since the sequence length can vary significantly between different datasets, the cost of labelling a single instance with GPT-3.5-turbo also varies. Additionally, various few-shot labelling strategies with GPT-3.5-turbo incur different costs, with more shots leading to a higher labelling cost due to the longer prompt. For our experiments, we track the number of tokens used in each API call.

We estimate the crowdsourcing labelling price from Google Cloud Platform[2]. For labeling classification tasks, it charges 1000 units (50 tokens per unit) for \$129 in Tier 1 and \$90 in Tier 2. We adopt the average cost from Tier 1 and Tier 2 as the human labelling cost. For generation tasks, there is no detailed instruction, as the rate can vary significantly based on task difficulty. Therefore, we follow the cost of classification tasks by charging \$0.11 per 50 tokens. It is important to note that actual human labelling is often more expensive. For example, the same instance is labelled by multiple labelers for majority voting and some datasets are labelled by experts rather than through crowdsourcing.

F Sampling strategy baselines

We compare our proposed sampling strategy with some common strategies for thorough comparisons:

- **Random Selection** We use random selection as a baseline, which samples uniformly from D_{Pool}. Since the pool data and test data generally share the same distribution, the sampled batch is expected to be i.i.d. (independent and identically distributed) with the test data.

- **Maximum Entropy** Entropy is a widely used measure of uncertainty [Settles, 2009]. Data points with the highest entropy according to the model M are selected for annotation. The selection is based on the following criterion:

$$\arg\max_{x \in D_{\text{Pool}}} \left(- \sum_{y \in Y} P_M(y|x) \log P_M(y|x) \right).$$

- **Least Confidence** [Culotta and McCallum, 2005] propose a method where examples are ranked based on the probability assigned by M to the predicted class $\hat{y}$. The data point with the highest confidence (i.e., the model's least uncertainty) is chosen for annotation. The selection is made according to:

$$\arg\max_{x \in D_{\text{Pool}}} \left(1 - P_M(\hat{y}|x)\right).$$

- **K-Means Diversity Sampling** Diversity sampling aims to select batches of data that are heterogeneous in the feature space. Following [Yuan *et al.*, 2020], we apply k-means clustering to the L2-normalized embeddings of M_4, and then sample the nearest neighbours of the k cluster centres.

[1] https://platform.openai.com/docs/pricing

[2] https://cloud.google.com/ai-platform/data-labeling/pricing#labeling_costs

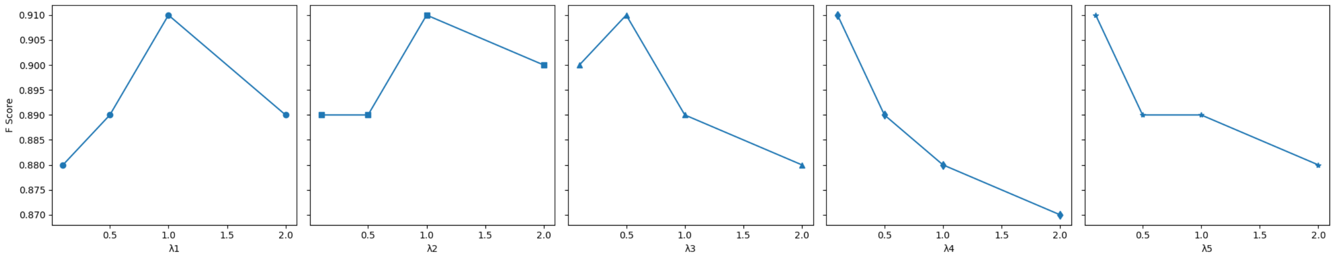

Figure 1: Overall F1-Scores with different hyperparameters: $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and$\lambda_5$.

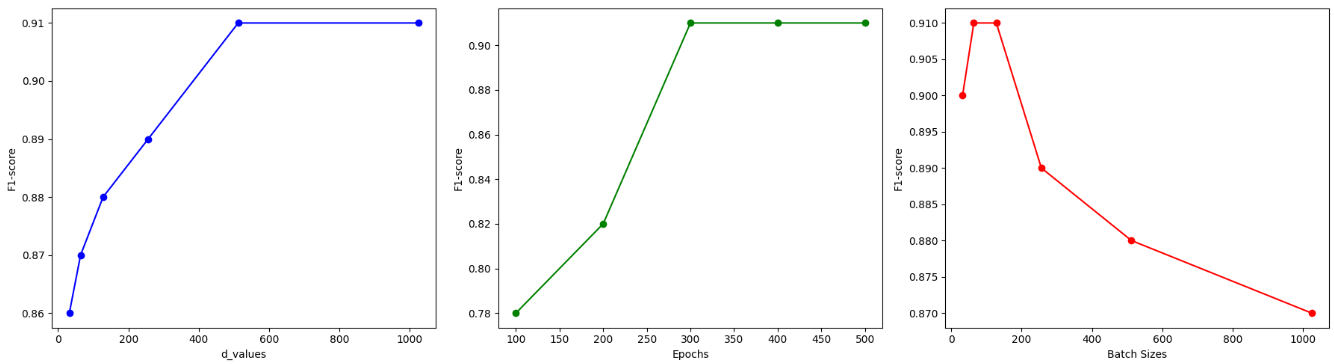

Figure 2: Overall F1-Scores with different hyperparameters: Embedding Dimension (d), Epochs and Batch Size.

References

[Culotta and McCallum, 2005] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

[Liang *et al.*, 2022] Chaoqi Liang, Yu Zhang, Xinyuan Li, Jinyu Zhang, and Yongqi Yu. Fudfend: fuzzy-domain for multi-domain fake news detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 45–57. Springer, 2022.

[Nan *et al.*, 2022] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. Improving fake news detection of influential domain via domain-and instance-level transfer. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2834–2848, 2022.

[Settles, 2009] Burr Settles. Active learning literature survey. 2009.

[Shu *et al.*, 2019] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405, 2019.

[Silva *et al.*, 2020] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embedding partial propagation network for fake news early detection. In *CIKM (Workshops)*, volume 2699, 2020.

[Silva *et al.*, 2021] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 557–565, 2021.

[Wang *et al.*, 2023] Daokang Wang, Wubo Zhang, Wenhuan Wu, and Xiaolei Guo. Soft-label for multi-domain fake news detection. *IEEE Access*, 2023.

[Yuan *et al.*, 2020] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, 2020.

[Zhou *et al.*, 2020] Xinyi Zhou, Jindi Wu, and Reza Zafarani. : Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer, 2020.

[Zhu *et al.*, 2022] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191, 2022.