Real, Fake, or Manipulated? Detecting Machine-Influenced Text

Yitong Wang*[1] Zhongping Zhang*[1] Margherita Piana[1] Zheng Zhou[2]
Peter Gerstoft[2] Bryan A. Plummer[1]
[1]Boston University, [2]University of California, San Diego
Correspondence: bplum@bu.edu

Abstract

Large Language Model (LLMs) can be used to write or modify documents, presenting a challenge for understanding the intent behind their use. For example, benign uses may involve using LLM on a human-written document to improve its grammar or to translate it into another language. However, a document entirely produced by a LLM may be more likely to be used to spread misinformation than simple translation (*e.g.*, from use by malicious actors or simply by hallucinating). Prior works in Machine Generated Text (MGT) detection mostly focus on simply identifying whether a document was human or machine written, ignoring these fine-grained uses. In this paper, we introduce a HiErarchical, length-RObust machine-influenced text detector (HERO), which learns to separate text samples of varying lengths from four primary types: human-written, machine-generated, machine-polished, and machine-translated. HERO accomplishes this by combining predictions from length-specialist models that have been trained with Subcategory Guidance. Specifically, for categories that are easily confused (*e.g.*, different source languages), our Subcategory Guidance module encourages separation of the fine-grained categories, boosting performance. Extensive experiments across five LLMs and six domains demonstrate the benefits of our HERO, outperforming the state-of-the-art by 2.5-3 mAP on average[1].

1 Introduction

Fine-grained Machine Generated Text (FG-MGT) detection aims to predict if a document was human-written, machine-generated, or some combination thereof. Prior work has primarily focused on separating paraphrased or machine-polished text from human and/or completely machine-generated text (Krishna et al., 2024; Li et al., 2024; Abassy et al., 2024), as these are often benign uses of a language model. In contrast, machine generated text may hallucinate (Cao et al., 2022; Parikh et al., 2020; Zhou et al., 2021; Maynez et al., 2020; Shuster et al., 2021; Gou et al., 2023; Meng et al., 2022) and is more likely to contain misinformation (Lin et al., 2022; Zellers et al., 2019), making them less trustworthy. However, prior work ignores other benign uses of LLMs, like machine translation, which may also be flagged as machine-generated by traditional MGT detectors (see Fig. 1).

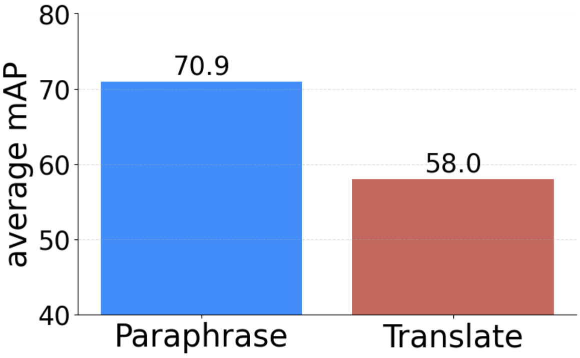


Figure 1: Paraphrasing/polishing human-written text or translating it into another language are often benign applications of an LLM that users of MGT detectors, like moderators, may wish to ignore. However, off-the-shelf methods (*e.g.*, (Hans et al., 2024)) often identify this type of data as machine-generated. In this paper, we increase the practical use of MGT detectors by separating text into fine-grained production categories, providing insight into content intent.

To address this issue, in this paper we introduce a HiErarchical, length-RObust machine-influenced text detector (HERO), which provides fine-grained labels to better understand document authorship. Specifically, as shown in Fig. 2, we expand the set of possible authorship categories to not only include machine translated text, but also the source language from which it is translated from. As we

*Denotes equal contribution
[1]Code: `https://github.com/ellywang66/HERO`

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15022–15037
November 4-9, 2025 ©2025 Association for Computational Linguistics

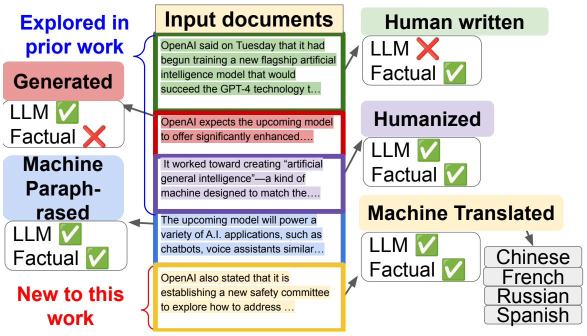

Explored in prior work
Input documents
Human written
LLM ❌
Factual ✅
Generated
LLM ✅
Factual ❌
OpenAI said on Tuesday that it had begun training a new flagship artificial intelligence model that would succeed the GPT-4 technology t…
OpenAI expects the upcoming model to offer significantly enhanced….
Humanized
LLM ✅
Factual ✅
Machine Paraph-rased
It worked toward creating "artificial general intelligence"—a kind of machine designed to match the….
LLM ✅
Factual ✅
The upcoming model will power a variety of A.I. applications, such as chatbots, voice assistants similar…
Machine Translated
LLM ✅
Factual ✅
Chinese
French
Russian
Spanish
New to this work
OpenAI also stated that it is establishing a new safety committee to explore how to address …

Figure 2: **Illustration of Fine-grained Machine Generated Text Detection (FG-MGT).** The goal of FG-MGT is to identify different types of generated text to provide some insight into the potential intent. In this paper, we extend the study of Abassy et al. (2024) to include machine-translated text.

will show, identifying the source language both provides more detailed authorship information and also improves the ability of HERO to identify translated text as a whole. However, separating similar categories of machine-influenced (*i.e.*, translated or polished) text is challenging. For example, paraphrasing and translation both originate from a human-written article, and a sophisticated actor may also try to make their generated text more human-like in an effort to make it harder to detect. This is further exacerbated during inference when documents are from different domains or language models than those seen during training.

A straightforward approach to solve our FG-MGT problem would be to use a coarse-to-fine approach (*e.g.*, (Xu et al., 2023; Yuan et al., 2023; Amit et al., 2004)), where we train a model to predict the general categories, and then refine them using specialized models. However, this approach has two drawbacks. First, it can increase inference time as both coarse and fine models must be used for each input document. Second, it introduces a tradeoff between coarse and fine model predictions that may be challenging to define for strong distribution shifts at test time (*e.g.*, documents from out-of-domain language models). Thus, as we will show, this type of naive adaptation results in poor performance in practice. Instead, we introduce Subcategory Guidance modules, where we compute a separate loss function on subsampled logits to distinguish between closely related fine-grained categories. Unlike traditional coarse-to-fine methods, this does not add any additional computational requirements at test time, enabling it to scale to large numbers of categories.

Another challenge faced in FG-MGT is the variability of input text lengths, where smaller documents prove more challenging to detect. While this challenge is shared with the traditional MGT task (Hans et al., 2024; Mitchell et al., 2023; Verma et al., 2024; Guo et al., 2023; Zhang et al., 2024; Gehrmann et al., 2019; Su et al., 2023; Tian and Cui, 2023), the introduction of fine-grained categories amplifies the issue in our setting. Inspired by work in bias mitigation (Wang et al., 2020), we train a set of expert classifiers, each specialized towards a specific text length. Following prior work (Wang et al., 2020), we use all classifiers at test time regardless of input document length. See Fig. 3 for an illustration of our approach.

Our contributions are summarized as follows:

- We introduce HERO, a robust FG-MGT detector that combines categories into a hierarchy to focus the model's ability to discriminate between fine-grained categories, which outperforms the state-of-the-art by 2.5-3 mAP on average.
- We show Subcategory Guidance modules provide an effective approach for separating similar categories without incurring test-time resource costs suffered by related work.
- We conduct an in-depth analysis on FG-MGT using HERO to identify potential manipulation and misinformation in text content to ensure the safe deployment of LLMs.

2 Related Work

Most prior work in detecting Machine Generated Text (MGT) treats this task as a binary classification problem (Bhattacharjee et al., 2023; Solaiman et al., 2019; Guo et al., 2023; Tian et al., 2024; Mitchell et al., 2023; Hans et al., 2024; Zhang et al., 2024; Hu et al., 2023; Kuznetsov et al., 2024), *i.e.*, detecting whether the input text is human-written or machine-generated. These include Metric-based methods (Mitchell et al., 2023; Su et al., 2023; Bao et al., 2024; Hans et al., 2024; Miralles-González et al., 2025), which extract distinguishable features from the text using the target language models. *E.g.*, Solaiman et al. (2019) apply log probability, Gehrmann et al. (2019) use the absolute rank of each token, and Verma et al. (2024) searches over a language model's feature space. Many of these methods (*e.g.*, (Mitchell et al., 2023; Su et al., 2023; Bao et al., 2024), rely on an observation that small changes to generated text typically lower its log probability under the language model, a pattern not

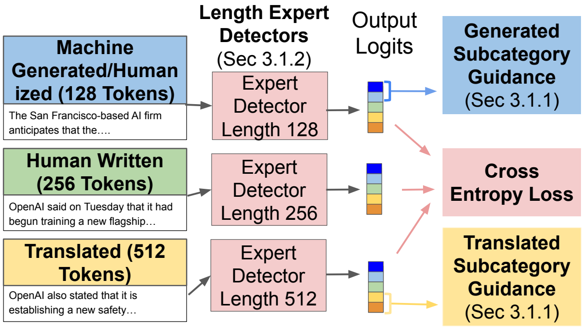

Length Expert Detectors (Sec 3.1.2)
Output Logits
Machine Generated/Humanized (128 Tokens)
The San Francisco-based AI firm anticipates that the....
Human Written (256 Tokens)
OpenAI said on Tuesday that it had begun training a new flagship...
Translated (512 Tokens)
OpenAI also stated that it is establishing a new safety...
Expert Detector Length 128
Expert Detector Length 256
Expert Detector Length 512
Generated Subcategory Guidance (Sec 3.1.1)
Cross Entropy Loss
Translated Subcategory Guidance (Sec 3.1.1)

Figure 3: **Our** HERO **framework**. Each input is processed by a specialized expert detector based on its token length. In addition to the standard cross-entropy loss, we introduce generated subcategory guidance to machine-generated and machine-humanized text, while translated subcategory guidance is used for translated text. See Sec. 3 for discussion.

seen in human-written text. Thus, these methods inject perturbations to the input text. However, these models are only defined for the binary classification, and it is unclear if they can be extended to our setting as we need to separate many types of machine influenced text.

Some recent studies have recognized the importance of detecting other categories of MGT (Krishna et al., 2024; Li et al., 2024; Nguyen-Son et al., 2021), including machine paraphrased and translated text. For example, Krishna et al. (2024) enhanced machine paraphrased text detection using retrieval methods, and Li et al. (2024) identified paraphrased sentences through article context. Nguyen-Son et al. (2021) applied round-trip translation to detect Google-translated text. Macko et al. (2023); Mao et al. (2025) explored detecting generated text in non-English languages, but not machine-translated text. Abassy et al. (2024) explored a fine-grained reason task similar to ours, but did not consider the effect of machine-translated text. The high similarity between the sub-categories can also reduce the generalization of such an approach to detect other types of manipulations.

3 Expanding Fine-grained Machine Generated Text Detection

Given an article $x_i \in \mathcal{X}$, fine-grained machine-generated text (FG-MGT) detection aims to separate samples into a set of categories $y_i \in \{0, 1, \ldots, K\}$ where $y_i = 0$ corresponds to human-written text, and $y_i = k$ where $k \in \{1, \ldots, K\}$ corresponds to one of K distinct categories of machine-influenced text. Prior work on FG-MGT explored up to four categories: **human** written, machine **generated**, **humanized** machine generated, and **paraphrased**/polished human written text (Krishna et al., 2024; Li et al., 2024; Abassy et al., 2024). However, this ignores **translated** text, another form of machine-influenced generation with often benign use, but, as shown in Fig. 1, may be detected as LLM-generated. Thus, to provide additional insight for users of FG-MGT models, we add a new category based on the source language a document was translated from. However, as we will show, we find that separating these types of similar generation types is challenging, especially on out-of-domain generators used at test time.

To address our FG-MGT task, we introduce HiErarchical, length-RObust machine-influenced text detector (HERO), which makes two improvements to FG-MGT detectors. First, Sec. 3.1.1 describes our Subcategory Guidance modules, which help construct a feature representation that can more easily separate similar categories. Second, Sec. 3.1.2 discusses our length-expert approach to improving support for varying document lengths. Sec. 3.2 discusses our data generation process that we use to train and evaluate our FG-MGT detectors.

3.1 Our HERO Approach

As discussed earlier, our objective is to create a FG-MGT model that can identify if a document is machine-generated and the specific type of machine influence. While our approach is designed to generalize across a wide range of authorship types and languages, in this paper we focus on predicting likelihoods over eight categories for English articles: human written, machine generated, paraphrased, humanized, translated (Chinese), translated (Russian), translated (Spanish), and translated (French) as defined at the beginning of Sec. 3. Our HERO model begins by taking our input document x passes it through a shared feature encoder g. To learn to identify our categories above, we use cross entropy $\mathcal{L}_{CE}$, whose classifier uses the input from $g(x)$ and estimates the likelihood that sample x was produced by one of the FG-MGT categories.

A simple approach would be to simply change an MGT detector (*e.g.*, (Hans et al., 2024; Mitchell et al., 2023; Verma et al., 2024; Guo et al., 2023; Zhang et al., 2024; Gehrmann et al., 2019; Su et al., 2023; Tian and Cui, 2023)) to produce a multi-class outputs. However, we found these models struggle to distinguish between similar generation types, especially when evaluated on out-of-distribution language models. We address this issue with a Subcategory Guidance module in the next section.

3.1.1 Fine-grained Text Classification via Subcategory Guidance

One common strategy for discriminating between fine-grained categories is to build a coarse-to-fine hierarchy (Xu et al., 2023; Yuan et al., 2023; Amit et al., 2004), where categories become more similar as you traverse down the hierarchy. However, these methods are often deployed within a single domain, *i.e.*, the distribution of the data see during training is similar to that seen at test time. This is due, in part, to the fact that these methods require careful tuning to balance the predictions of the hierarchy of classifiers being deployed. *I.e.*, they require careful calibration between the coarse and fine-grained classifiers to boost performance. In FG-MGT, this would put a significant limitation on our detectors, as it would effectively mean that we can only deploy them on seen text domains and for language models used during training.

Instead, we introduce a Subcategory Guidance module to help direct feature learning during training, which is discarded at test time. We group together semantically similar categories that specialize in separating samples in each group. Specifically, we create one module for each of the four translated categories as well as for machine-generated and humanized text. Although the machine generated and humanized text are both entirely generated, the fact that a user decided to query a language model to make the text appear more human suggests they might be trying to obfuscate a detector, providing some potential intent information. Similarly, knowing the language a document was translated from can provide clues as to where a document first appeared. Our Subcategory Guidance models aim to help our detector better discriminate between these categories.

Unlike the coarse-to-fine methods discussed earlier, these modules are discarded at test time. Thus, they do not affect computational resources at test time or require complicated calibration procedures that do not generalize well to out-of-domain samples. Instead, they boost performance by guiding the formation of the shared feature space produced by the shared encoder g during training. Each Subcategory Guidance module takes as input samples that stem only from the categories of their type. For example, the Translated Subcategory Guidance only takes features from documents from the four translated categories as input. Then it uses cross entropy to separate documents into their fine-grained categories. In effect, this simply amounts to computing a loss over a subset of predictions, making it easy to implement and deploy.

Our final objective consists of a tradeoff function balancing the task loss with our Subcategory Modules, which we define as $\mathcal{L}_{\text{GH}}$ and $\mathcal{L}_{\text{Trans}}$ for the generated/humanized and translated categories, respectively. Formally, our total loss is:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \lambda(\mathcal{L}_{\text{GH}} + \mathcal{L}_{\text{Trans}}), \qquad (1)$$

where λ is a tunable hyper-parameter.

3.1.2 Improving Support to Varying Document Lengths

Prior work has shown that short documents, which inherently have little information about authorship, are challenging to identify as machine generated (Zhang et al., 2024). Solaiman et al. (2019) found they could improve a detector's robustness to varying document lengths by randomly cropping articles during training. However, a detector for short length article has to naturally be more sensitive to distribution changes given the limited information than it does for a longer article. Training a single model to adjust for both the sensitivity as well as make fine-grained distinctions is challenging. Instead, we leverage a set of experts, each of which specializes in documents up to a set length.

Formally, given an input text $\mathbf{x}$, we train a set of M expert classifiers $\{f_1, \ldots, f_M\}$, each trained with a specific maximum token length and associated parameters $\mathbf{W}_m$. Each expert is trained using Subcategory Guidance from Sec. 3.1.1. However, empirically we find that including some information from documents of lengths other than the ones targeted by an expert can help improve performance (*e.g.*, seeing some 256 token length documents can boost performance for a 512-length expert). Thus, we used length cropping, where with p_{crop}, documents of other lengths are included during training to improve the model's robustness.

Given a document at test time we can simply use the expert of the closest length. If a document is between experts, we use the larger one. However, some prior work in bias mitigation has shown that averaging experts even over settings they do not specialize in can boost performance (Wang et al., 2020). In effect, when compute is available, these experts can form a type of ensemble. Thus, in our experiments we evaluate these experts as an ensemble in addition to using them individually.

3.2 Data Preparation: Article Generation

We generate articles for a range of domains (Sec. 3.2.1) and language models (Sec. 3.2.2) to ensure FG-MGT methods generalize across many settings, which we discuss in more detail below.

3.2.1 Source Datasets

GoodNews (Biten et al., 2019) provides URLs of New York Times articles from 2010 to 2018. After filtering out broken links and non-English articles, we randomly selected 8K/2K/2K articles for train/test/validation splits.

VisualNews (Liu et al., 2021) has articles from four media sources: *Guardian*, *BBC*, *USA Today*, and *Washington Post*. We randomly selected 2K articles for evaluation.

Student essays (Essay), creative writing (WP), and news articles (Reuters) (Verma et al., 2024) represent three diverse domains with 1K articles from each dataset used for evaluation.

WikiText (Stephen et al., 2017) contains 60 test articles collected from Wikipedia.

Each source dataset above provides per-category article counts. *E.g.*, for GoodNews this results in 2K human written articles * 8 categories = 16K total test samples. Additionally, only GoodNews was used for training, so results on the remaining datasets provide insight into domain shifts of varying degrees (*e.g.*, VisualNews and Reuters being close domains, whereas the rest are far domains).

3.2.2 Generation Process

To ensure the quality of generated text we keep all prompts for each category consistent throughout the generation process. All other hyperparameters such as temperature for each LLM are also kept the same for consistency. Specifically, the language models we used include Llama-3 (Touvron et al., 2023), Qwen-1.5 (Bai et al., 2023), StableLM-2 (Bellagente et al., 2024), ChatGLM-3 (Du et al., 2022), and Qwen-2.5 (Yang et al., 2024). Llama-3 is set as our in-domain generator used for training the detector, and StableLM-2, ChatGLM-3, Qwen-1.5, and Qwen-2.5 are out-of-domain generators to evaluate the model's generalization ability. When generating articles, we used a temperature for Llama-3 of 0.6 and for StableLM-2 we used 0.7. For Qwen-1.5, Qwen-2.5, and ChatGLM-3 we use default temperatures for generating responses.

To prevent the model from leaking information about the article's category (*e.g.*, Llama-3 often responds with "Here is the polished version:"), we use the text starting from the second sentence as input to the detector. Below we further discuss category-specific generation processes.

Machine-generated articles were created by giving the LLM the title with the prompt: "Write an article on the following title, ensuring that the article consists of approximately z sentences," where z represents the number of sentences in the original article. This ensures that articles of different categories are of similar length, preventing the detector from using length as a classification feature.

Machine-paraphrased articles were generated by giving the LLM the entire human-written article as input with the prompt "Paraphrase the following article: x." We provide a study the effect of replacing only parts of the articles in App. B.

Machine-translated articles were produced using the same process as for paraphrasing, only replacing the word "paraphrase" with "translate" in the prompt. Translated articles were drawn from the following languages: Chinese, Spanish, Russian, and French (additional discussion in App. C).

Machine-humanized articles were created by giving the a machine-generated article as input to the LLM with the prompt: "Rewrite this text to make it sound more natural and human-written." We provide a specific example in App. D.

4 Experiments

Implementation Details. Our base encoder uses a DistilBERT (Sanh et al., 2020) backbone. The maximum token length of the input text is set to 512 when training all methods (including our own). The same maximum length is used evaluate the model's performance during testing except where noted. For training, we used the Adam optimizer with a maximum learning rate of 10^{-5}. Following Zhang et al. (2024); Verma et al. (2024), we fine-tuned the model for three epochs to prevent overfitting. Our experiments were conducted on a single GPU (*e.g.*, A40, L40S). For a single dataset (*e.g.*, GoodNews), data preparation takes approximately 60 hours, and training takes around 1 hour.

Metrics. Our main results use mean Average Precision (mAP) and the probability of detection (PD) at 5% false positive rate (FPR). We report mAP per generator, and then rank a detector's overall performance by averaging mAP across both in-domain

	In-domain LLMs	Out-of-domain LLMs					
Model	Llama3	Qwen1.5	Qwen2.5	ChatGLM3	StableLM2	**avg**	**PD**
Scale	-8B	-7B	-12B	-6B	-7B	**mAP**	**5%FPR**
In-domain mAP on GoodNews (Biten et al., 2019)							
OpenAI-D (large)	94.04	41.60	**41.90**	48.23	71.21	59.39	53.31
ChatGPT-D	80.83	38.79	40.43	40.00	62.74	52.56	47.08
LLM-DetectAIve	96.24	41.27	42.28	44.72	76.87	60.28	55.31
DistilBERT	96.89	38.99	40.91	42.21	74.59	58.72	53.77
HERO (ours)	**98.33**	**44.05**	41.88	**50.47**	**76.93**	**62.33**	**56.23**
Out-of-domain mAP on VisualNews (Liu et al., 2021)							
OpenAI-D (large)	60.67	32.62	**37.99**	38.51	52.53	44.46	36.64
ChatGPT-D	47.19	27.39	31.02	33.82	49.93	37.87	31.20
LLM-DetectAIve	62.41	32.54	**39.41**	36.34	**55.70**	45.28	38.51
DistilBERT	61.11	31.64	32.77	36.75	54.43	43.34	36.49
HERO (ours)	**64.17**	**38.98**	37.70	**42.17**	55.48	**47.70**	**39.09**
Out-of-domain mAP on WikiText (Stephen et al., 2017)							
OpenAI-D (large)	65.39	33.65	**36.43**	36.38	52.06	44.78	35.33
ChatGPT-D	38.78	29.16	32.76	30.78	43.08	34.91	24.38
LLM-DetectAIve	65.62	29.89	28.55	27.52	45.38	39.39	29.92
DistilBERT	66.37	33.51	28.42	30.19	49.82	41.66	33.12
HERO (ours)	**72.19**	**37.97**	31.52	**35.45**	**52.58**	**45.94**	**37.00**
Out-of-domain mAP on WP (He et al., 2023)							
OpenAI-D (large)	55.48	46.69	44.88	37.44	**55.34**	47.97	23.50
ChatGPT-D	40.57	37.71	42.22	34.86	**43.69**	39.81	17.50
LLM-DetectAIve	65.39	48.53	**51.35**	35.35	**56.32**	51.39	28.50
DistilBERT	71.65	44.61	50.88	40.51	49.74	**51.48**	29.25
HERO (ours)	**73.68**	47.14	41.01	39.12	52.58	50.71	**42.55**
Out-of-domain mAP on Reuters (He et al., 2023)							
OpenAI-D (large)	74.63	**50.72**	**51.51**	54.11	54.08	57.01	26.75
ChatGPT-D	57.42	46.03	48.28	50.65	44.05	49.29	22.00
LLM-DetectAIve	**85.92**	38.48	43.37	53.54	59.65	56.19	22.75
DistilBERT	81.04	48.66	42.95	42.14	**65.93**	56.14	31.00
HERO (ours)	84.50	48.59	41.99	50.85	58.94	**56.98**	**49.70**
Out-of-domain mAP on Essay (He et al., 2023)							
OpenAI-D (large)	51.29	29.09	31.74	**40.79**	36.42	37.87	18.25
ChatGPT-D	32.27	33.24	29.97	30.62	25.45	30.31	11.75
LLM-DetectAIve	52.56	40.22	**41.37**	34.78	37.89	**41.38**	23.75
DistilBERT	48.98	**36.13**	31.16	28.39	36.57	36.25	17.25
HERO (ours)	**60.07**	38.20	35.90	33.11	35.69	40.59	**33.76**

Table 1: Fine-grained MGT detection results on in-domain GoodNews data and five out-of-domain datasets. HERO outperforms or obtains similar results to prior work in nearly all settings and metrics, providing an overall advantage.

and out-of-domain LLMs (avg mAP). We also report F1 score in some ablations.

4.1 Baselines

OpenAI-D (Solaiman et al., 2019) is a detector trained on outputs from GPT-2 (Radford et al.,

2019) series. OpenAI provides two versions: RoBERTa-base and RoBERTa-large. With fine-tuning and early stopping, OpenAI-D can also be used to detect text generated by other LLMs.

ChatGPT-D (Guo et al., 2023) is designed to identify text produced by ChatGPT-3.5 (Ouyang et al., 2022). It is trained using the HC3 (Guo et al., 2023) dataset, which includes 40,000 questions along with both human-written and ChatGPT-generated answers, before finetuning on our task.

LLM-DetectAIve (Abassy et al., 2024) distinguishes between machine-generated, machine-paraphrased, and human-written text by fine-tuning RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) models. We apply the DeBERTa backbone of LLM-DetectAIve in our experiments.

DistilBERT (Sanh et al., 2020) is a distilled version of BERT (Devlin et al., 2019). Since the model is pre-trained using knowledge distillation, it is smaller and faster at inference time.

4.2 Results

Tab. 1 compares the performance of our HERO approach to prior work on in-domain and out-of-domain datasets, respectively. Note that both tables have results on in-domain and out-of-domain LLMs used for generation. HERO achieves a boost in both mAP and PD 5%FPT in nearly all settings (*e.g.*, a 2% boost in avg mAP on GoodNews as seen in Tab. 1). In cases where we underperform prior work on avg mAP, *e.g.*, 1 point worse on WP and Essay in Tab. 1, we greatly outperform in PD 5%FPT (a 12.5% and 10% gain, respectively). Thus, our approach demonstrates significant benefits over the methods from prior work.

Fig. 4 reports per-class performance on GoodNews and Reuters as representatives of in-domain and out-of-domain data, respectively. We make two major observations about these results. First, in-domain data and LLMs gets nearly perfect performance, highlighting the significant role shifts in both has on performance. For example, while nearly all methods get perfect performance on humanized data in Fig. 4(a), when we shift domains (but not LLMs) in Fig. 4(b) performance drops significantly. Second, no method gets the best performance consistently across all categories. For example, while HERO gets best performance identifying out-of-domain human-written articles, LLM-DetectAIive performs best on identifying Russian-sourced translations on Reuters (but performs relatively poorly on paraphrased data). Thus, our performance improvements come from having more consistent results rather than being strictly better for all categories.

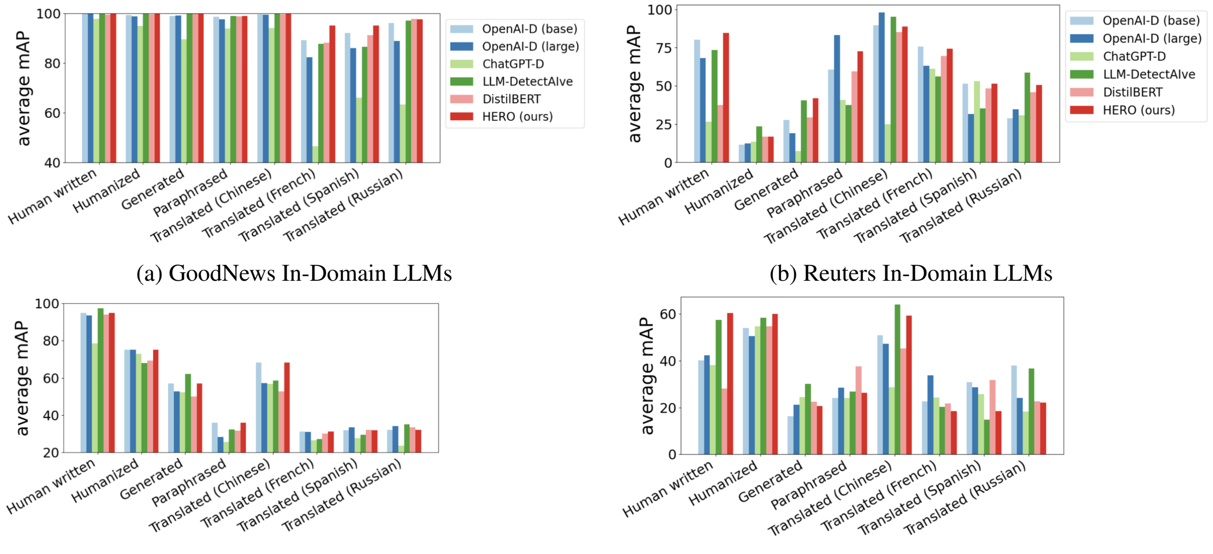
	Llam a3-8B	Qwen 1.5-7B	StableL M2-12B	ChatGL M3-6B	Qwen 2.5-7B	**avg** **mAP**
(a) DistilBERT (2020)						
L=32	37.97	25.38	22.75	24.10	26.42	27.33
L=50	45.49	31.01	27.47	28.05	31.52	32.71
L=128	58.09	37.97	33.98	33.68	39.64	40.67
L=256	46.41	29.46	37.60	32.01	34.19	35.93
L=500	66.71	32.12	58.50	59.37	32.47	49.84
L=512	61.13	31.18	54.49	35.02	32.88	42.94
(b) HERO (ours) - Single Length Specialist Only						
L=32	37.51	32.15	31.18	26.22	29.61	31.33
L=50	44.62	36.68	34.08	30.04	33.72	35.83
L=128	62.23	43.94	38.53	36.81	43.79	45.06
L=256	50.98	33.75	43.22	34.67	39.25	40.37
L=500	63.79	37.73	56.44	40.53	38.42	47.38
L=512	60.87	35.36	54.51	38.00	35.50	44.85
(c) HERO (ours) - All Length Specialists						
L=32	36.39	31.78	30.92	25.79	29.22	30.82
L=50	42.65	35.75	33.66	29.04	32.93	34.81
L=128	60.23	42.51	37.20	35.01	42.08	43.40
L=256	51.92	32.50	39.27	34.46	38.63	39.36
L=500	69.91	47.39	58.98	38.47	47.07	52.36
L=512	64.07	38.22	56.73	41.00	38.04	47.61

Table 2: Comparison of mAP scores on VisualNews (Liu et al., 2021) across different input lengths for DistilBERT (2020) and HERO. HERO consistently outperforms DistilBERT across all lengths and generators. For length-specialist models, we use the expert closest in length, defaulting to the longer one when in between.

Tab. 2 reports performance on various input document lengths using our FG-MGT detectors. Across all token length settings, performance generally improves with longer token lengths with the best results consistently observed at 500 and 512 tokens. Compared to DistilBERT (2020), both the individual length specialist and HERO demonstrate improved performance. The Length Specialist approach shows especially strong performance on short lengths, with the single specialists outperforming the ensemble, validating that such documents require special care.

4.3 HERO Model Analysis

Tab. 4 provides an ablation study to show the contribution of each component of HERO. We see Subcategory Guidance provides a 2.5 average mAP gain over the baseline DistilBERT (Sanh et al., 2020). We also compare to a naive coarse-to-fine approach that first tries to predict if an input document is human-written, machine-generated, para-



(a) GoodNews In-Domain LLMs

(b) Reuters In-Domain LLMs

(c) GoodNews Out-of-Domain LLMs

(d) Reuters Out-of-Domain LLMs

Figure 4: Per-class mAP results on the GoodNews (Biten et al., 2019) and Reuters (Verma et al., 2024) datasets. Top row: In-domain LLMs. Bottom row: Out-of-domain LLMs. Our method shows more robust performance, especially on human-written and translated categories.

	ID LLMs			OOD LLMs		
Model	Low	Median	High	Low	Median	High
In-domain mAP on GoodNews (Biten et al., 2019)						
OpenAI-D (base)	95.23	94.53	92.71	52.58	**27.30**	29.06
OpenAI-D (large)	92.36	90.87	84.25	58.44	25.97	28.16
ChatGPT-D	72.81	69.87	65.68	46.41	26.40	25.74
LLM-DetectAIve	93.45	95.03	89.59	**65.90**	25.99	**31.86**
DistilBERT	94.85	95.49	92.46	62.97	26.55	27.38
HERO (ours)	**97.41**	**97.36**	**95.50**	64.61	26.44	27.31
Out-of-domain mAP on WikiText (Stephen et al., 2017)						
OpenAI-D (base)	67.02	69.39	64.81	31.29	29.40	32.49
OpenAI-D (large)	72.88	69.68	67.56	**42.19**	30.92	33.11
ChatGPT-D	53.40	44.53	40.96	28.55	29.01	27.66
LLM-DetectAIve	67.20	71.23	63.40	37.48	31.11	31.52
DistilBERT	70.50	65.74	57.95	34.24	29.04	32.18
HERO (ours)	**73.98**	**69.19**	**63.08**	36.76	**32.06**	**34.13**

Table 3: Fine-grained MGT detection results by BLEU translation quality. We find that HERO performs especially well across quality groups on out-of-domain data.

phrased, or translated. If it is machine-generated or translated, we use a separate detector to separate it into the subcategories. Comparing the 2nd and 3rd row of Tab. 4, we see the naive approach underperforms our Subcategory Guidance approach by 14.5 average mAP, highlighting the challenges of generalizing beyond the training domain in our task. We also show that Length Cropping and our expert models from Sec. 3.1.2 both individually boost performance, but when we combine all components we see the best performance.

Tab. 3 reports the effect of translation quality binned into low, medium, and high based on BLEU scores. Similar to our per-category results discussed earlier, HERO's benefits stem from performing better across the varying degrees of translation quality. Notably, our approach performs especially well on out-of-domain data (WikiText results), obtaining the best performance in all but one setting. We also evaluate how the extent of paraphrasing used affects performance in App. B, where HERO typically reports at least a 2 average mAP gain over the baseline DisilBERT model. These results demonstrate our approach's robustness to a wide range of applications.

Is HERO still effective if subcategory information is not required? Tab. 5 we evaluate a setting where the goal is only to predict one of four categories: human written, machine generated, machine paraphrased, and translated (effectively eliminating the subcategories). We compare a DistilBERT trained to predict these four categories with HERO, where we take the highest subcategory score to represent our confidence in that category. We see that HERO still obtains 3 mAP gain on average, demonstrating the benefits of leveraging subcategory information even if the fine-grained category predictions are not necessary.

Fig. 5 shows the effect of training on different combinations of languages. As we increase the number of languages beyond two, we start to see

	In-domain LLMs	Out-of-domain LLMs						
Model	Llama3	Qwen1.5	Qwen2.5	ChatGLM3	StableLM2	**avg**	**PD**	**F1**
Scale	-8B	-7B	-12B	-6B	-7B	**mAP**	**5%FPR**	**Score**
DistilBERT (2020)	61.11	31.64	54.43	36.75	32.77	43.34	36.49	32.30
+Naive Coarse-to-Fine	42.62	29.12	28.36	26.89	30.66	31.53	10.67	22.62
+Subcategory Guidance	61.20	37.09	54.83	39.16	37.56	45.97	39.09	33.87
+Length Cropping (2019)	60.69	38.24	53.84	40.06	34.20	45.40	37.66	31.22
+Length Specialists	63.21	34.44	54.05	39.72	**38.10**	45.90	38.86	33.85
HERO (ours)	**64.17**	**38.98**	**55.48**	**42.17**	37.70	**47.70**	**39.09**	**33.99**

Table 4: Ablation Study on Visualnews (Liu et al., 2021). Each component contributes to model performance. Additionally, our Subcategory Guidance outperforms alternatives like a Naive Coarse-to-Fine approach.

Model	Llama3	Qwen1.5	StableLM2	ChatGLM3	Qwen2.5	**avg mAP**
Scale	-8B	-7B	-12B	-6B	-7B	
DistilBERT (2020)	49.58	45.24	41.66	44.30	45.68	45.29
HERO (Ours)	**52.46**	**47.29**	**45.59**	**47.23**	**48.19**	**48.15**

Table 5: Comparison of mAP scores on VisualNews (Liu et al., 2021) for DistilBERT (2020) and HERO on human-written, machine-generated, machine paraphrased, and machine translated categories. Identifying source languages can still boost performance even when all translations are treated as a single category.

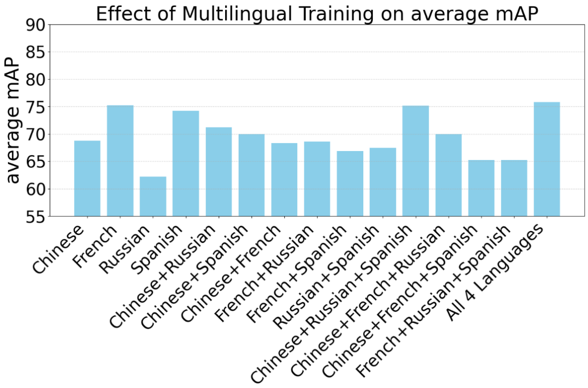


Figure 5: Effect of multilingual training on average mAP across different language combinations evaluated on the four class VisualNews (Liu et al., 2021) setting also reported in Tab. 5. Models trained on multiple languages generally outperform those trained on a single language, with the highest average mAP observed when training on all four languages.

some saturation, where there are smaller differences between models, suggesting that a very large number of languages may not be necessary to recognize a document as originating from another language.

5 Conclusion

In this paper, we conduct an in-depth study of fine-grained MGT detection, aiming to further distinguish between machine translated and machine paraphrased texts from MGT. We introduced HERO, a fine-grained machine-influenced text detection framework that goes beyond the classical binary classification approach. Our hierarchical structure, combined with length-specialist models, enables strong generalization across diverse LLMs and varying input lengths, making it suitable for real-world applications. Our extensive experiments across multiple LLMs and different datasets show that HERO consistently outperforms the state-of-the-art by 2.5-3 mAP, and does especially well in out of domain settings. We also show that identifying source languages can boost a model's ability to identify translated text. Overall, HERO enables more accurate detection of machine-influenced content, which is essential for future works in discerning between benign and malicious uses of LLMs.

Acknowledgments. This material is based upon work supported by DARPA under agreement number HR00112020054. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

6 Limitations

In this paper, we have investigated the FG-MGT task and our proposed HERO shows improved performance over existing detectors. Despite the improved performance, our method still has several limitations, discussed further below.

While our proposed method improves performance for zero-shot evaluations in our experiments, our approach does not guarantee 100% accuracy on other LLMs and datasets. Therefore, we strongly discourage the use of our approach without proper human supervision (*e.g.*, for plagiarism detection or similar formal applications). A more appropriate application of HERO is to introduce human-supervision for more reliable detection against LLM-generated misinformation.

We also notice the performance difference between in-domain LLM and out-of-domain LLMs. As shown in Sec. 4.2, the performance of HERO on out-of-domain generators (StableLM-2, ChatGLM-3, Qwen-2.5, Qwen-1.5) is still lower than that on in-domain generators (Llama-3). Therefore, out-of-domain evaluations remain a challenge for future research in this topic.

Our translated data also utilized a round-trip strategy (discussed in App. C) to control for content consistency. However, these translations may also introduce some noise into the articles that may make them easier to detect. Thus, our results should be seen as only an approximation of the model's true performance.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, and 1 others. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. *arXiv preprint arXiv:2408.04284*.

Yali Amit, Donald Geman, and Xiaodong Fan. 2004. A coarse-to-fine strategy for multiclass shape detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1606–1621.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations (ICLR)*.

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, and 1 others. 2024. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*.

Amrita Bhattacharjee, Tharindu Kumarage, Raha Moraffah, and Huan Liu. 2023. Conda: Contrastive domain adaptation for ai-generated text detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 598–610.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3340–3354.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning (ICML)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. Mgtbench: Benchmarking machine-generated text detection. *arXiv preprint arXiv:2303.14822*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-text detection via adversarial learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2024. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems (NeurIPS)*.

Kristian Kuznetsov, Eduard Tulchinskii, Laida Kushnareva, German Magai, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. Robust AI-generated text detection by restricted embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Yafu Li, Zhilin Wang, Leyang Cui, Wei Bi, Shuming Shi, and Yue Zhang. 2024. Spotting ai's touch: Identifying llm-paraphrased spans in text. In *Findings of the Annual Meeting of the Association for Computational Linguistics: ACL*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visualnews: Benchmark and challenges in entity-aware image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Dianhui Mao, Denghui Zhang, Ao Zhang, and Zhihua Zhao. 2025. Mlsdet: Multi-llm statistical deep ensemble for chinese ai-generated text detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Pablo Miralles-González, Javier Huertas-Tato, Alejandro Martín, and David Camacho. 2025. Not all tokens are created equal: Perplexity attention weighted networks for ai generated text detection. *Preprint*, arXiv:2501.03940.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning (ICML)*.

Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. 2021. Machine translated text detection through text similarity with round-trip translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5792–5797.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Merity Stephen, Xiong Caiming, Bradbury James, and Richard Socher. 2017. Pointer sentinel mixture models. *Proceedings of ICLR*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP*.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale positive-unlabeled detection of AI-generated texts. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. 2023. Dynamic coarse-to-fine learning for oriented tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7318–7328.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. 2023. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6317–6327.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems (NeurIPS)*, 32.

Zhongping Zhang, Wenda Qin, and Bryan A. Plummer. 2024. Machine-generated text localization. In *Findings of the Annual Meeting of the Association for Computational Linguistics: ACL*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

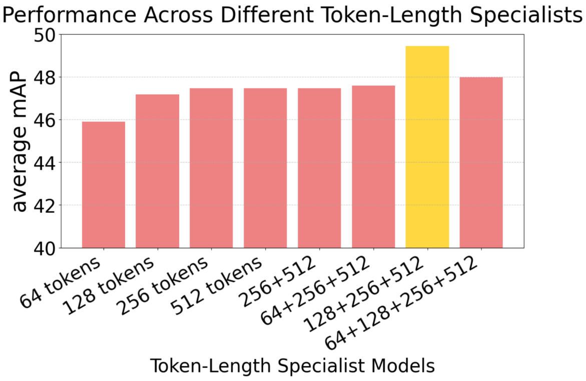


Figure 6: Average mAP across different token-length specialist models evaluated on VisualNews (Liu et al., 2021). Models trained with a single token length achieve moderate performance, while combining specialists across multiple token lengths significantly improves detection accuracy. The highest average mAP is observed when using specialists for 128, 256, and 512 tokens.

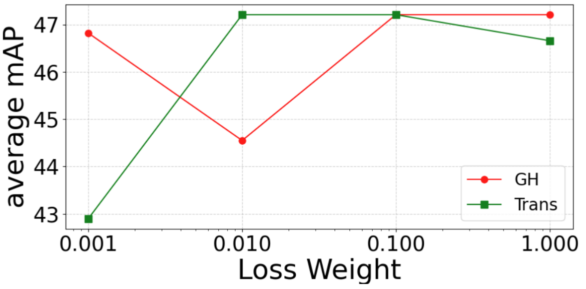


Figure 7: Effect of GH (Generate-Humanized) and Trans loss weights for guided learning on average mAP performance evaluated on VisualNews (Liu et al., 2021). The model achieves the highest mAP when both the GH and Trans loss weights are set to 0.01.

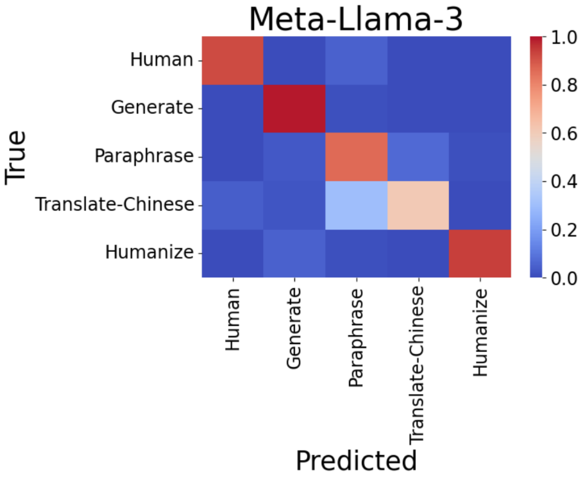


Figure 8: Confusion Matrix for in-domain LLMs on VisualNews. HERO performs well in most categories, especially on the machine-translated articles.

Appendix

A Additional Results

Fig. 7 shows the effect of changing the loss weight λ from Eq. 1. The same value of λ performs best for both, reducing the number of hyperparameters that need to be tuned for our model.

Fig. 6 ablates the number and size of experts to train. We find that three experts generally provide enough coverage to perform well on a diverse set of lengths. That said, the number of experts likely would vary depending on the maximum input sequence a model can support. However, very long documents are easier to detect as machine-generated (see Tab. 2), so support for very long sequences may not be necessary as a model may be able to effectively detect a language model was used on just part of a document.

Class confusion matrix. To provide a more intuitive understanding of HERO, we provide the visualization for HERO's performance across different FG-MGT categories on VisualNews using confusion matrices as shown in Fig. 8 and Fig. 9. The results show that HERO can accurately distinguish translated text from different source languages, even when evaluated on out-of-domain LLMs. However, the model continues to struggle with distinguishing between generated and humanized content. This challenge may stem from the fact that both types are produced by LLMs using human written input, resulting in similar surface-level characteristics.

B Effect of Paraphrasing Extent

To examine how the extent of paraphrasing can affect the performance of HERO, we paraphrased 20%, 40%, 60%, 80%, and 100% of the input text. The resulting performance is shown in Fig. 10. As paraphrasing extent increases, the detection accuracy also improves, suggesting that higher levels of paraphrasing make manipulation patterns more discernible to the model.

C Round-trip Translation Strategy

To create translated versions of the same documents, we adopt the strategy of round-trip translation to generate translated data for our FG-MGT task. Fig. 11 provides a specific example: we first translate the original article into target languages (Chinese, Spanish, French, Russian), and then translate these articles back into English, obtaining machine-translated articles for detection.

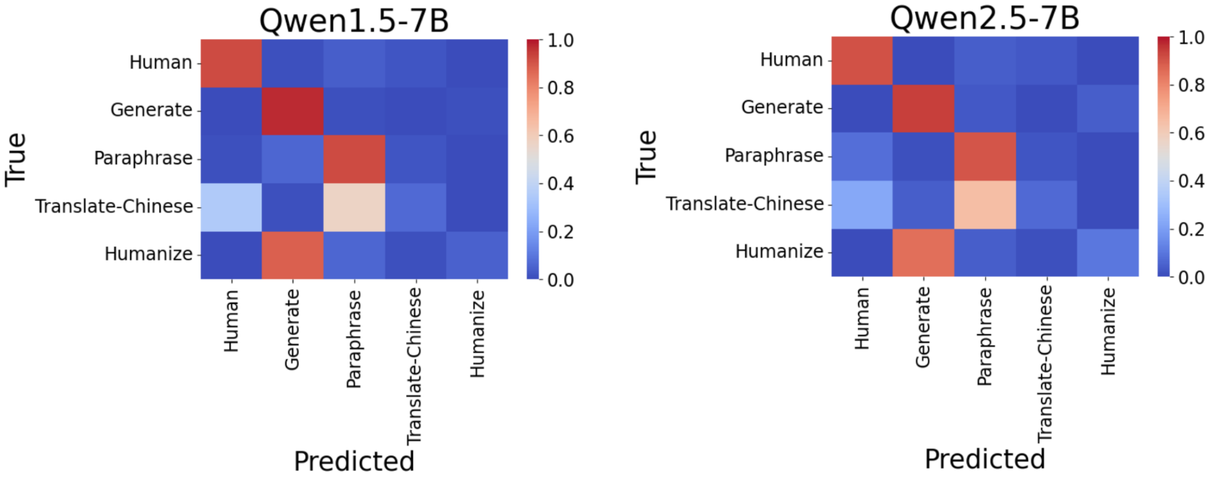


Figure 9: Confusion Matrix on out-of-domain LLMs on VisualNews. Our method can still accurately distinguish between human-written and machine-generated categories. However, when compared to in-domain evaluations in Fig. 8, detecting machine-humanized text becomes more challenging.

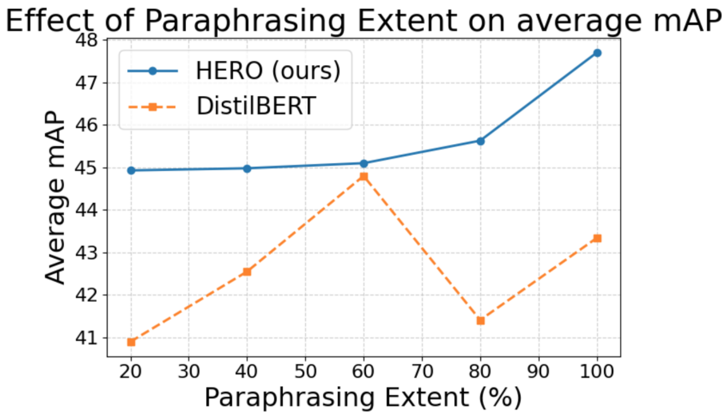
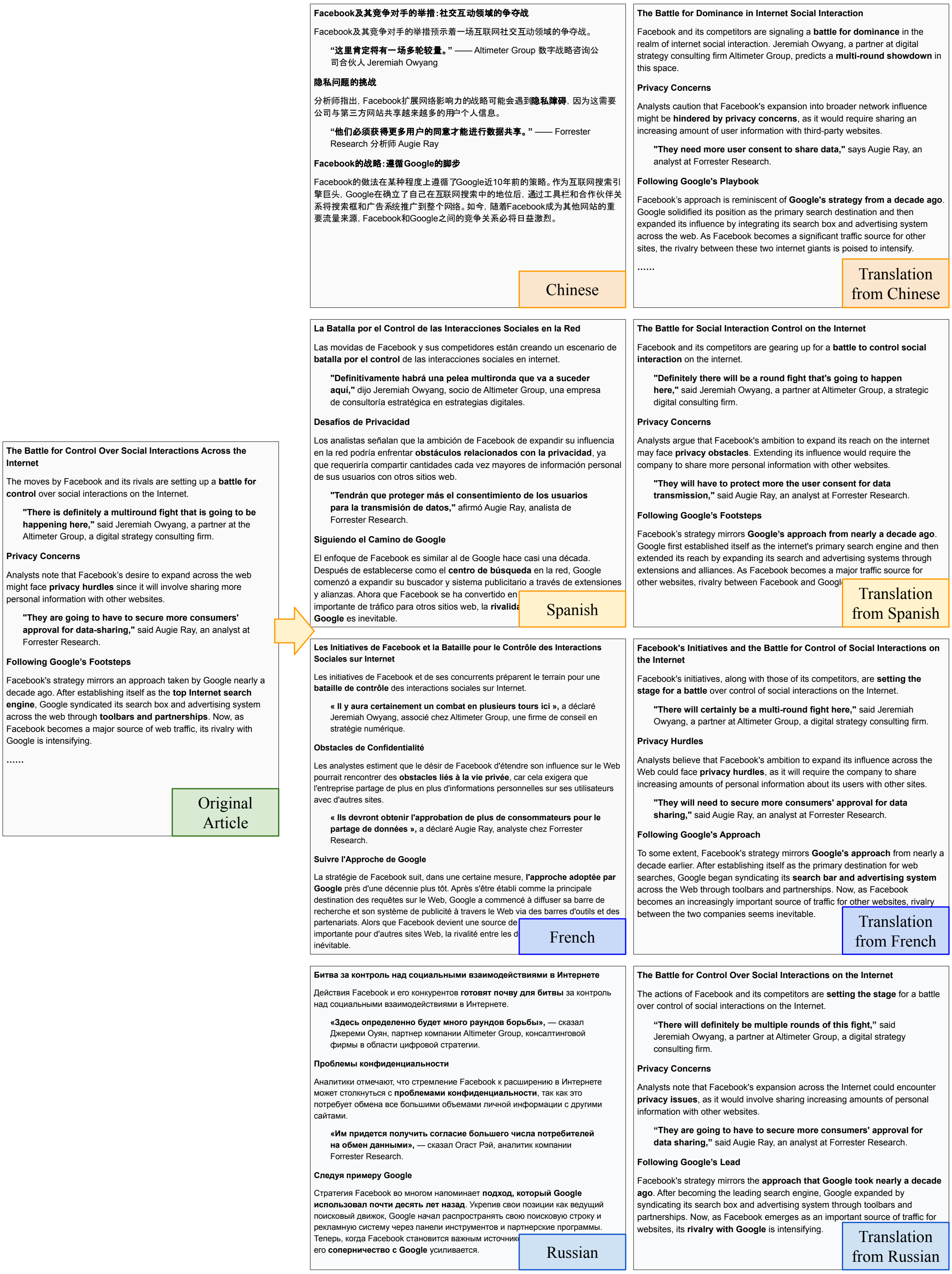


Figure 10: We illustrate the effect of paraphrasing extent on average mAP. Higher levels of paraphrasing improve the model's performance.

D Humanized example

The purpose of machine-humanized data is to simulate a setting where a bad actor may attempt to make their generated text harder to detect. It accomplishes this by querying a LLM with a request to make the input article sound more human using processes based on those from Abassy et al. (2024) (see Sec. 3.2.2 for additional discussion). Fig. 12 shows an example of the differences produced by machine-humanized data.



The Battle for Control Over Social Interactions Across the Internet

The moves by Facebook and its rivals are setting up a **battle for control** over social interactions on the Internet.

> **"There is definitely a multiround fight that is going to be happening here,"** said Jeremiah Owyang, a partner at the Altimeter Group, a digital strategy consulting firm.

Privacy Concerns

Analysts note that Facebook’s desire to expand across the web might face **privacy hurdles** since it will involve sharing more personal information with other websites.

> **"They are going to have to secure more consumers' approval for data-sharing,"** said Augie Ray, an analyst at Forrester Research.

Following Google’s Footsteps

Facebook's strategy mirrors an approach taken by Google nearly a decade ago. After establishing itself as the **top Internet search engine**, Google syndicated its search box and advertising system across the web through **toolbars and partnerships**. Now, as Facebook becomes a major source of web traffic, its rivalry with Google is intensifying.

......

Facebook及其竞争对手的举措:社交互动领域的争夺战

Facebook及其竞争对手的举措预示着一场互联网社交互动领域的争夺战。

> **“这里肯定将有一场多轮较量。”** —— Altimeter Group 数字战略咨询公司合伙人 Jeremiah Owyang

隐私问题的挑战

分析师指出，Facebook扩展网络影响力的战略可能会遇到**隐私障碍**，因为这需要公司与第三方网站共享越来越多的用户个人信息。

> **“他们必须获得更多用户的同意才能进行数据共享。”** —— Forrester Research 分析师 Augie Ray

Facebook的战略:遵循Google的脚步

Facebook的做法在某种程度上遵循了Google近10年前的策略。作为互联网搜索引擎巨头，Google在确立了自己在互联网搜索中的地位后，通过工具栏和合作伙伴关系将搜索框和广告系统推广到整个网络。如今，随着Facebook成为其他网站的重要流量来源，Facebook和Google之间的竞争关系必将日益激烈。

The Battle for Dominance in Internet Social Interaction

Facebook and its competitors are signaling a **battle for dominance** in the realm of internet social interaction. Jeremiah Owyang, a partner at digital strategy consulting firm Altimeter Group, predicts a **multi-round showdown** in this space.

Privacy Concerns

Analysts caution that Facebook's expansion into broader network influence might be **hindered by privacy concerns**, as it would require sharing an increasing amount of user information with third-party websites.

> **"They need more user consent to share data,"** says Augie Ray, an analyst at Forrester Research.

Following Google's Playbook

Facebook’s approach is reminiscent of **Google's strategy from a decade ago**. Google solidified its position as the primary search destination and then expanded its influence by integrating its search box and advertising system across the web. As Facebook becomes a significant traffic source for other sites, the rivalry between these two internet giants is poised to intensify.

......

La Batalla por el Control de las Interacciones Sociales en la Red

Las movidas de Facebook y sus competidores están creando un escenario de **batalla por el control** de las interacciones sociales en internet.

> **"Definitivamente habrá una pelea multironda que va a suceder aquí,"** dijo Jeremiah Owyang, socio de Altimeter Group, una empresa de consultoría estratégica en estrategias digitales.

Desafíos de Privacidad

Los analistas señalan que la ambición de Facebook de expandir su influencia en la red podría enfrentar **obstáculos relacionados con la privacidad**, ya que requeriría compartir cantidades cada vez mayores de información personal de sus usuarios con otros sitios web.

> **"Tendrán que proteger más el consentimiento de los usuarios para la transmisión de datos,"** afirmó Augie Ray, analista de Forrester Research.

Siguiendo el Camino de Google

El enfoque de Facebook es similar al de Google hace casi una década. Después de establecerse como el **centro de búsqueda** en la red, Google comenzó a expandir su buscador y sistema publicitario a través de extensiones y alianzas. Ahora que Facebook se ha convertido en importante de tráfico para otros sitios web, la **rivalida** **Google** es inevitable.

The Battle for Social Interaction Control on the Internet

Facebook and its competitors are gearing up for a **battle to control social interaction** on the internet.

> **"Definitely there will be a round fight that's going to happen here,"** said Jeremiah Owyang, a partner at Altimeter Group, a strategic digital consulting firm.

Privacy Concerns

Analysts argue that Facebook's ambition to expand its reach on the internet may face **privacy obstacles**. Extending its influence would require the company to share more personal information with other websites.

> **"They will have to protect more the user consent for data transmission,"** said Augie Ray, an analyst at Forrester Research.

Following Google’s Footsteps

Facebook’s strategy mirrors **Google’s approach from nearly a decade ago**. Google first established itself as the internet's primary search engine and then extended its reach by expanding its search and advertising systems through extensions and alliances. As Facebook becomes a major traffic source for other websites, rivalry between Facebook and Googl

Les Initiatives de Facebook et la Bataille pour le Contrôle des Interactions Sociales sur Internet

Les initiatives de Facebook et de ses concurrents préparent le terrain pour une **bataille de contrôle** des interactions sociales sur Internet.

> **« Il y aura certainement un combat en plusieurs tours ici »,** a déclaré Jeremiah Owyang, associé chez Altimeter Group, une firme de conseil en stratégie numérique.

Obstacles de Confidentialité

Les analystes estiment que le désir de Facebook d'étendre son influence sur le Web pourrait rencontrer des **obstacles liés à la vie privée**, car cela exigera que l'entreprise partage de plus en plus d'informations personnelles sur ses utilisateurs avec d'autres sites.

> **« Ils devront obtenir l'approbation de plus de consommateurs pour le partage de données »,** a déclaré Augie Ray, analyste chez Forrester Research.

Suivre l'Approche de Google

La stratégie de Facebook suit, dans une certaine mesure, **l'approche adoptée par Google** près d'une décennie plus tôt. Après s'être établi comme la principale destination des requêtes sur le Web, Google a commencé à diffuser sa barre de recherche et son système de publicité à travers le Web via des barres d'outils et des partenariats. Alors que Facebook devient une source de importante pour d'autres sites Web, la rivalité entre les d inévitable.

Facebook's Initiatives and the Battle for Control of Social Interactions on the Internet

Facebook's initiatives, along with those of its competitors, are **setting the stage for a battle** over control of social interactions on the Internet.

> **"There will certainly be a multi-round fight here,"** said Jeremiah Owyang, a partner at Altimeter Group, a digital strategy consulting firm.

Privacy Hurdles

Analysts believe that Facebook's ambition to expand its influence across the Web could face **privacy hurdles**, as it will require the company to share increasing amounts of personal information about its users with other sites.

> **"They will need to secure more consumers' approval for data sharing,"** said Augie Ray, an analyst at Forrester Research.

Following Google's Approach

To some extent, Facebook's strategy mirrors **Google's approach** from nearly a decade earlier. After establishing itself as the primary destination for web searches, Google began syndicating its **search bar and advertising system** across the Web through toolbars and partnerships. Now, as Facebook becomes an increasingly important source of traffic for other websites, rivalry between the two companies seems inevitable.

Битва за контроль над социальными взаимодействиями в Интернете

Действия Facebook и его конкурентов **готовят почву для битвы** за контроль над социальными взаимодействиями в Интернете.

> **«Здесь определенно будет много раундов борьбы»,** — сказал Джереми Оуян, партнер компании Altimeter Group, консалтинговой фирмы в области цифровой стратегии.

Проблемы конфиденциальности

Аналитики отмечают, что стремление Facebook к расширению в Интернете может столкнуться с **проблемами конфиденциальности**, так как это потребует обмена все большими объемами личной информации с другими сайтами.

> **«Им придется получить согласие большего числа потребителей на обмен данными»,** — сказал Огаст Рэй, аналитик компании Forrester Research.

Следуя примеру Google

Стратегия Facebook во многом напоминает **подход, который Google использовал почти десять лет назад**. Укрепив свои позиции как ведущий поисковый движок, Google начал распространять свою поисковую строку и рекламную систему через панели инструментов и партнерские программы. Теперь, когда Facebook становится важным источник его **соперничество с Google** усиливается.

The Battle for Control Over Social Interactions on the Internet

The actions of Facebook and its competitors are **setting the stage** for a battle over control of social interactions on the Internet.

> **“There will definitely be multiple rounds of this fight,”** said Jeremiah Owyang, a partner at Altimeter Group, a digital strategy consulting firm.

Privacy Concerns

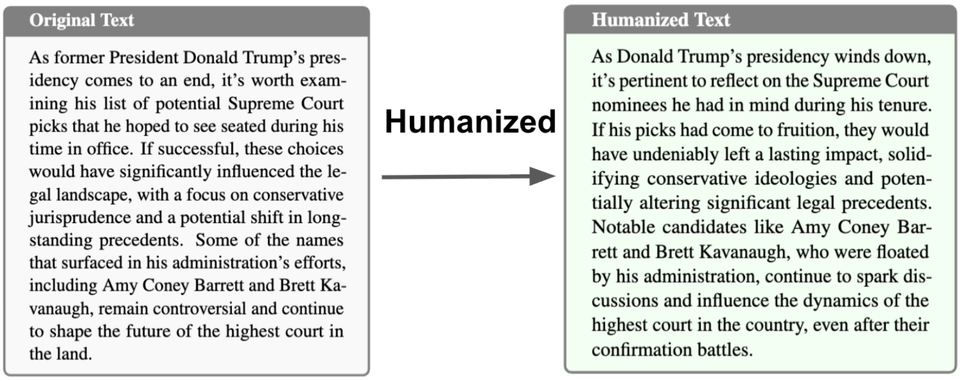
Analysts note that Facebook's expansion across the Internet could encounter **privacy issues**, as it would involve sharing increasing amounts of personal information with other websites.

> **“They are going to have to secure more consumers' approval for data sharing,”** said Augie Ray, an analyst at Forrester Research.

Following Google’s Lead

Facebook's strategy mirrors the **approach that Google took nearly a decade ago**. After becoming the leading search engine, Google expanded by syndicating its search box and advertising system through toolbars and partnerships. Now, as Facebook emerges as an important source of traffic for websites, its **rivalry with Google** is intensifying.

Figure 11: **Round-trip Strategy for Generating Translated Articles.**This strategy allows us to automatically produce translated articles from existing datasets, eliminating the need for additional data collection. See Sec. C for discussion.



Original Text

As former President Donald Trump's presidency comes to an end, it's worth examining his list of potential Supreme Court picks that he hoped to see seated during his time in office. If successful, these choices would have significantly influenced the legal landscape, with a focus on conservative jurisprudence and a potential shift in longstanding precedents. Some of the names that surfaced in his administration's efforts, including Amy Coney Barrett and Brett Kavanaugh, remain controversial and continue to shape the future of the highest court in the land.

Humanized Text

As Donald Trump's presidency winds down, it's pertinent to reflect on the Supreme Court nominees he had in mind during his tenure. If his picks had come to fruition, they would have undeniably left a lasting impact, solidifying conservative ideologies and potentially altering significant legal precedents. Notable candidates like Amy Coney Barrett and Brett Kavanaugh, who were floated by his administration, continue to spark discussions and influence the dynamics of the highest court in the country, even after their confirmation battles.

Figure 12: **Humanized text example**. We utilize machine-generated text and ask the LLMs to rewrite it to sound more natural and human-like, while maintaining the same level of detail and length.