Out-of-Distribution Detection of Machine Generated Text

by

Prodromos Kampouridis

Submitted
in partial fulfilment of the requirements for the degree of

Master of Artificial Intelligence

at the

UNIVERSITY OF PIRAEUS

February 2026

Author ..... Prodromos Kampouridis

II-MSc “Artificial Intelligence”

February 26, 2026

Certified by.....

Efstathios Stamatatos

Professor

Thesis Supervisor

Certified by.....

George Vouros

Professor

Member of Examination Committee

Certified by.....

Georgios Petasis

Research Associate

Member of Examination Committee

Out-of-Distribution Detection of Machine Generated Text

By

Prodromos Kampouridis

Submitted to the II-MSc "Artificial Intelligence" on February 2026, in partial fulfillment of the
requirements for the MSc degree

Abstract

Detecting machine generated text is increasingly important as Large Language Models (LLMs) evolve rapidly. In practice, detectors often fail to generalize out of distribution (OOD), degrading under domain shifts, topic changes, unseen generators, and paraphrasing attacks. This thesis studies whether compact machine generated text detectors can retain stronger OOD robustness through teacher-student training. The student is optimized with supervised cross-entropy, optional logit-based knowledge distillation via temperature-scaled KL divergence, and teacher-guided representation alignment using triplet loss and supervised contrastive learning.

Experiments on the MAGE benchmark follow its OOD testbed protocol across unseen domain, unseen model, combined shift, and paraphrasing settings. To support deployment-oriented evaluation, performance is reported both before and after decision-threshold calibration. The results show that triplet-based teacher guidance is the strongest distillation strategy among the distilled variants, with the best final model combining cross-entropy, knowledge distillation, and teacher-guided triplet alignment. Overall, the proposed distilled detector is competitive with the MAGE Longformer baseline on standard OOD settings and achieves substantially lower inference latency, yielding a lightweight and practically efficient detector for robust machine generated text detection beyond the training distribution.

Thesis Supervisor: Efstathios Stamatatos
Title: Professor

Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor Efstathios Stamatatos for his guidance and his valuable feedback which were essential to the completion of this work. I am also particularly grateful to Professor Georgios Vouros for his continued support and direction.

Any opinions, findings, conclusions or recommendations expressed in this material are my own and do not necessarily reflect the views of the «funding body» or the view of University of Piraeus and Inst. of Informatics and Telecom. of NCSR “Demokritos”.

Table of Contents

List of Tables

1 Introduction

1.1 Background and Motivation

Machine Generated Text Detection is the task of determining whether a given text sample was produced by a human author or generated by an automated system, most commonly an LLM. In practice, it is typically framed as a binary classification problem in which the detector outputs a label and often a score (e.g., a probability or likelihood-based statistic) indicating how likely the text is machine generated (Wu et al., 2023). It differs from generator attribution, where the goal is to identify which model produced a text rather than merely whether it is machine generated. It also differs from watermark verification, which detects a deliberate signature inserted at generation time; detection, as studied in this thesis, must often operate when no watermark is present, when the generator is unknown, or when outputs come from multiple systems with inconsistent watermark policies (Kirchenbauer et al., 2023; Wu et al., 2023).

The importance of Machine Generated Text Detection has increased sharply because LLM usage is now widespread in everyday writing workflows and institutional settings. In education and research, generative AI systems are used for drafting, summarization, translation, tutoring, and rewriting. UNESCO's guidance on generative AI in education and research emphasizes that these tools create immediate pressures around academic integrity, assessment validity, and accountable use, while also highlighting the need for careful governance and human-centered deployment practices (UNESCO, 2023). In media and information ecosystems, LLMs enable rapid creation of plausible narratives and large-scale content production. Even when the content is not explicitly malicious, the volume and apparent credibility of generated text can erode trust, overwhelm moderation capacity, and complicate provenance judgments. In security contexts, generated text can be used at scale for deception (e.g., scams or deceptive narratives), so provenance signals can be useful as part of broader defenses (Wu et al., 2023).

Further motivation arises from the long-term consequences of pervasive synthetic text on the data used to train future models. If web-scale corpora increasingly contain machine generated text, future training may incorporate synthetic artifacts and progressively lose diversity in rare or "tail" phenomena. This risk has been analyzed as model collapse, where iterative training on generated data causes degenerative behavior and distributional defects (Shumailov et al., 2023). While detection alone cannot prevent synthetic content from being published, robust detection can contribute to dataset curation and filtering pipelines intended to preserve human-authored signal.

Despite extensive recent work on machine generated text detection, existing approaches often fall short in precisely the conditions that matter most for real-world deployment: out-of-distribution (OOD) settings. The OOD problem refers to the mismatch between the training distribution used to develop a detector and the test distribution encountered after deployment. In machine generated text detection, this mismatch is especially pronounced because the text generation ecosystem changes rapidly. New model families and model versions appear frequently, and the style of generated text shifts as alignment, decoding strategies, and post-processing methods evolve. Detectors trained on a fixed set of generators can therefore overfit generator-specific artifacts and degrade when confronted with previously unseen generators or updated model behavior, a phenomenon that can be described as model drift at the level of the generator distribution (Wu et al., 2023).

OOD challenges also arise from domain and task shift. A detector trained primarily on one genre, such as news or student essays, may fail when applied to social media posts, technical writing, customer support conversations, or creative fiction because topic, register, and stylistic conventions differ. The MAGE benchmark was designed to stress this reality by constructing a broad "in-the-wild" testbed spanning diverse writing tasks and many generator models, and by evaluating detectors under progressively harder scenarios, including out-of-domain and out-of-generator settings (Li et al., 2024). Findings reported with MAGE illustrate that strong in-distribution performance does not necessarily translate to robust performance under OOD shift, reinforcing the need to treat generalization as a primary research objective rather than an afterthought.

Prompt shift is another critical source of OOD behavior. Even with a fixed generator model, changes in prompt templates and prompting intent can shift output distributions, affecting lexical choice, structure, and discourse patterns. A detector that implicitly learns prompt-dependent artifacts may therefore generalize poorly when prompts differ across settings or users. In addition, paraphrasing and editing represent a particularly damaging form of shift because they directly target detectable artifacts. Evidence from stress tests shows that paraphrasing-based attacks can substantially reduce detection rates across detector families and can even undermine watermark-based methods under adversarial conditions (Sadasivan et al., 2023).

Mixed human-AI text further complicates detection. Many documents are hybrids in which an author drafts content and uses an LLM for polishing, or an LLM produces a draft that a human edits. In such cases, binary labels may not reflect the true provenance of the final text, and detectors can become miscalibrated. This raises conceptual questions about what should count as "machine generated," as well as practical questions about how to evaluate and deploy detectors responsibly (Wu et al., 2023). Finally, fairness and language proficiency effects introduce additional failure modes. Empirical work has shown that some detectors can misclassify non-native English writing as AI-generated at higher rates, which is a serious concern for educational and evaluative settings where false positives can produce unjust outcomes (Liang et al., 2023).

These limitations motivate the central technical direction of this thesis: improving OOD generalization of machine generated text detectors through a training framework that combines knowledge distillation in a teacher-student setting with objectives that shape robust representation geometry. Knowledge distillation transfers informative soft targets from a stronger teacher model to a compact student, typically by matching temperature-scaled ("soft") output distributions using a cross-entropy loss (equivalently KL-divergence up to constants), and has been shown to improve efficiency while preserving a significant portion of the teacher's performance (Hinton et al., 2015). Metric and contrastive learning objectives, such as triplet loss and supervised contrastive loss, explicitly encourage separation between classes in embedding space. Triplet loss enforces a margin between anchor-positive pairs and anchor-negative pairs,

while supervised contrastive learning generalizes contrastive objectives to supervised settings by pulling together same-class representations and pushing apart different-class representations within a batch (Schroff et al., 2015; Khosla et al., 2020).

1.2 Risks and Potential Abuses

Misinformation and disinformation are among the most frequently cited risks of machine generated text, because generative systems enable large-scale production of plausible narratives that can be disseminated rapidly and cheaply. Machine generated text detection can support mitigation by providing provenance signals that help prioritize content for review, support investigations, and inform content moderation workflows. At the same time, detection must be interpreted carefully: a determination that text is machine generated is not evidence that the content is false, and human-authored misinformation remains common.

Plagiarism and academic dishonesty are prominent concerns in educational contexts, where the unacknowledged use of generative tools can undermine assessment validity and learning outcomes. UNESCO's guidance explicitly frames generative AI as creating immediate integrity and accountability challenges and emphasizes the need for appropriate institutional policies and pedagogical redesign (UNESCO, 2023). Detection systems are often proposed as technical safeguards, but their deployment in education carries risks, especially when detectors are brittle under OOD shift or produce biased false positives. Empirical evidence that some detectors disproportionately flag non-native English writing as machine generated demonstrates why detector outputs should not be treated as definitive proof and why fairness evaluation is essential prior to deployment (Liang et al., 2023).

Malicious uses such as online fraud (including impersonation) and social media spam are facilitated by the low marginal cost of producing fluent, context-aware messages (Wu et al., 2023). In these settings, machine generated text detection may provide useful signal for triage or filtering, particularly when combined with other security signals. However, robust deployment is complicated by the ease of

paraphrasing and post-editing, which can remove detectable artifacts. Stress tests showing that paraphrasing-based attacks can bypass many detectors highlight the need to evaluate systems under adversarial attacks and distribution shifts rather than only under clean in-distribution conditions (Sadasivan et al., 2023).

Ethical concerns include fairness, accountability, and the risk of harmful consequences from misclassification. False positives can impose reputational damage or punitive outcomes when detection outputs are used in disciplinary settings, and the risk is amplified when detector performance varies across demographic or linguistic groups. The bias findings reported by Liang et al. (2023) illustrate that detector deployment can unintentionally penalize certain writers and therefore require careful validation, transparency about limitations, and conservative operational use.

Societal implications extend beyond individual misuse cases. Pervasive synthetic text can degrade trust in information ecosystems, encourage "liar's dividend" dynamics where authentic content is dismissed as synthetic, and increase the cost of verification. It can also affect the future of machine learning itself by altering the composition of training data on the web. The model collapse phenomenon described by Shumailov et al. (2023) provides a concrete mechanism by which recursive training on generated data can degrade model behavior over time, strengthening the argument for provenance tools that support dataset hygiene and preservation of human-authored signals.

1.3 Objectives and Scope

The primary objective of this thesis is to develop and evaluate a machine generated text detection framework that remains reliable under out-of-distribution conditions, with an emphasis on generalization across unseen domains, prompts, and generator models. The thesis operationalizes this objective through a teacher-student training paradigm in which a high-capacity teacher provides informative supervision signals, and a smaller student is optimized to achieve strong detection performance at reduced computational cost. This is achieved by combining standard supervised cross-entropy with knowledge distillation via KL-divergence under temperature scaling, following

the formulation introduced by Hinton et al. (2015). The thesis further investigates whether representation learning objectives improve robustness by encouraging class-separable embedding geometry. Specifically, it studies triplet loss, which aims to enforce relative distance constraints among samples, and supervised contrastive loss, which encourages compact same-class clusters and larger margins between classes, and it evaluates systematic combinations of these objectives with cross-entropy and distillation (Schroff et al., 2015; Khosla et al., 2020).

This thesis is guided by research questions that reflect the core robustness and efficiency goals. One research question examines not only whether distillation from a stronger teacher improves the student's out-of-distribution generalization relative to a student trained only with hard-label cross-entropy, but also how closely the distilled student can approach the teacher's own out-of-distribution performance. This comparison is important because the value of distillation lies not in exceeding the teacher, but in determining whether a substantially smaller and cheaper model can retain enough of the teacher's robustness to offer a favorable efficiency-performance trade-off. A second research question examines whether triplet and supervised contrastive objectives yield representations that generalize more effectively under domain and prompt shift, and under post-processing such as paraphrasing, relative to purely discriminative cross-entropy training. A third research question examines how these objectives interact, asking which combinations produce the best trade-off among in-distribution accuracy, OOD robustness, and computational efficiency, and whether the benefits of distillation and metric/contrastive learning are complementary or redundant.

The thesis contributes a unified training framework for machine generated text detection that supports cross-entropy, temperature-scaled knowledge distillation, triplet loss, supervised contrastive loss, and principled combinations of these objectives in a teacher-student setting. It contributes an empirical study that evaluates these objectives under OOD conditions motivated by practical deployment, using benchmark designs intended to reflect "in-the-wild" detection challenges. In addition, it contributes diagnostic analyses that clarify when and why particular objectives improve robustness, including ablation studies and representation-space analyses, thereby supporting reproducible model design.

The scope of the thesis is defined by the text-only detection setting and by the evaluation resources available. Experiments are conducted using the MAGE benchmark, which is explicitly designed to measure detector performance under diverse writing tasks, diverse generator models, and OOD scenarios, and which has been used to demonstrate that OOD conditions substantially increase the difficulty of detection (Li et al., 2024). The thesis focuses on modern transformer-based architectures suitable for classification and representation learning in text, with the teacher model representing a higher-capacity detector and the student model representing a more efficient alternative suitable for deployment. The thesis does not claim universal robustness to adaptive adversaries; rather, it evaluates robustness under realistic shifts and, where available in benchmarks, under paraphrasing-oriented conditions that are known to stress detectors (Sadasivan et al., 2023; Li et al., 2024).

Several topics are intentionally out of scope to maintain a clear technical focus. Watermarking is not treated as a primary solution in this work because it requires the generator to embed a detectable signal at generation time, whereas many operational settings involve unknown generators or text produced without watermark support (Kirchenbauer et al., 2023). Open-world generator attribution is also not claimed as a primary deliverable; attribution can be explored as a constrained analysis when labels are available, but the thesis does not assert reliable identification of unseen future models. Multimodal detection, such as detecting synthetic images or audio, is excluded, and the thesis focuses on written text. Finally, the thesis does not attempt to resolve institutional policy, legal regulation, or governance questions; instead, it provides technical evidence that can inform responsible deployment in settings where provenance signals are used cautiously and in combination with human judgment (UNESCO, 2023).

1.4 Thesis Outline

This thesis is structured to move from the problem motivation and robustness challenge to the proposed training framework and its evaluation under out-of-distribution conditions.

Chapter 1 introduces the problem of out-of-distribution machine generated text detection, motivates its importance, and summarizes key risks associated with misuse.

Chapter 2 reviews prior work on detection methods, robustness challenges, and the benchmark datasets used to evaluate detectors under distribution shift.

Chapter 3 presents the proposed teacher-student framework and formalizes the training objectives studied in this thesis, including cross-entropy, KL-based knowledge distillation with temperature scaling, triplet loss, supervised contrastive loss, and their combinations (Hinton et al., 2015; Schroff et al., 2015; Khosla et al., 2020).

Chapter 4 describes the experimental design, datasets, evaluation metrics, and OOD protocols used to assess generalization, including held-out domains, held-out generators, and paraphrasing-oriented settings where available (Li et al., 2024).

Chapter 5 reports the empirical results and compares loss combinations under both in-distribution and OOD conditions, including efficiency comparisons between teacher and student models.

Chapter 6 provides analysis through ablation studies and error diagnostics, and it discusses limitations and deployment considerations, especially in light of rewriting-based evasion and fairness concerns (Sadasivan et al., 2023; Liang et al., 2023).

Chapter 7 concludes with a summary of findings and outlines directions for future work on robust detection under evolving generators and realistic post-processing.

2 Literature Review

This chapter reviews existing methods for OOD Detection and summarizes how they are evaluated. It also presents the datasets and benchmarks used in prior studies with commentary on their main characteristics and limitations.

2.1 Detection paradigms for Machine Generated Text

Machine generated text detection is often studied in a black-box setting where only the produced text is available, but some approaches leverage additional provenance signals, most notably watermark verification, which assumes a detectable signature was embedded at generation time (Kirchenbauer et al., 2023).

2.1.1 Stylometric (feature-based) detection

Stylometric approaches originate from authorship attribution, where the objective is to identify writers via relatively stable stylistic markers such as function-word usage, character n-grams, lexical richness, and syntactic patterns (Stamatatos, 2009).

For AI-text detection, stylometry often treats “human” and “machine” as two author classes. However, stylometry has conceptual limitations for “misinformation detection” because models may remain stylistically consistent regardless of intent. Schuster et al. show stylometry can support provenance-like discrimination but is limited for detecting machine generated misinformation as such (Schuster et al., 2020).

Evaluation implications. Because many stylometric cues correlate with domain/genre/register, stylometry is sensitive to domain shift, making it easy to confuse distribution shift with “AI-ness” when test data differs from training (Schuster et al., 2020).

2.1.2 Statistical and language-model scoring cues

A second family detects generated text via statistical irregularities under a reference LM: token-rank histograms, entropy patterns, and likelihood-based scores. GLTR operationalizes the hypothesis that many generators oversample from high-probability regions of the next-token distribution, and it provides visual analytics to help human judgment (Gehrmann et al., 2019). In a controlled human-subjects study, GLTR's annotation scheme improved human detection of fake text from 54% to 72% without prior training (Gehrmann et al., 2019).

Evaluation implications. Rank/likelihood cues are entangled with decoding strategy and excerpt length. Ippolito et al. (2020) benchmark decoding strategies and show that decoding changes can substantially affect both human and automatic detection performance, emphasizing that detectors may exploit decoding fingerprints rather than robust provenance signals.

2.1.3 Supervised neural discriminators

Supervised detectors typically fine-tune pretrained encoders (e.g., BERT, RoBERTa) on labeled human-vs-machine corpora (Devlin et al., 2019; Liu et al., 2019).

Controlled experiments demonstrate very strong **in-distribution** performance but show brittleness under shifts. In particular, Ippolito et al. (2020) show that decoding methods tuned to fool humans can introduce statistical abnormalities that make automatic detection easier; they also report that even multi-sentence excerpts can fool expert human raters over **30%** of the time.

Evaluation implications. High i.i.d. accuracy often fails to transfer when domains, generator families, decoding strategies, or attacks change, motivating OOD-oriented evaluation rather than random splits alone (Dugan et al., 2024; Li et al., 2024).

2.1.4 Training-free and zero-shot detectors

To reduce dependence on curated training data and adapt to new generators, many works propose training-free / zero-shot detection.

DetectGPT uses probability-curvature structure of a source model's log-probability function. In their ICML report, DetectGPT improves detection of GPT-NeoX fake-news samples from 0.81 AUROC (strongest zero-shot baseline) to 0.95 AUROC (Mitchell et al., 2023).

Fast-DetectGPT replaces DetectGPT's perturbation-heavy step with a more efficient sampling-based curvature estimator (Bao et al., 2023). In their reported summary table, Fast-DetectGPT improves AUROC over DetectGPT in both a "5-model" setting (from 0.9554 to 0.9887) and a ChatGPT/GPT-4 setting (from 0.7225 to 0.9338), and reports large speedups (Bao et al., 2023).

DetectLLM leverages log-rank information and reports absolute improvements of +3.9 and +1.75 AUROC points over prior state-of-the-art zero-shot baselines across their evaluated datasets/models (Su et al., 2023a).

Binoculars contrasts two closely related LMs and reports detecting **over 90%** of generated samples from ChatGPT (and other LLMs) at **0.01%** false-positive rate across a wide range of document types (Hans et al., 2024).

Evaluation implications. These methods can generalize better than supervised detectors in some OOD settings, but their performance remains context-dependent: mismatch between the scoring model and true generator family, or post-editing that removes signature artifacts, can cause sharp drops (Li et al., 2024; Dugan et al., 2024).

2.1.5 Watermarking and provenance mechanisms

Watermarking modifies generation so outputs carry a detectable signature. A canonical approach biases sampling toward a context-dependent "green list" and detects the watermark via a statistical hypothesis test, enabling detection without access to model weights (Kirchenbauer et al., 2023).

Evaluation implications. Robustness depends on both provider-side adoption and resilience to transformations (Kirchenbauer et al., 2023; Krishna et al., 2023). Stress tests show that paraphrasing/rewriting can reduce detection

success for both watermarking and standard detectors (Sadasivan et al., 2023; Krishna et al., 2023).

2.1.6 Retrieval and database-assisted defences

A more "provenance-first" defense is retrieval: if an API provider can store generations, detection can attempt to match suspect text to a database of prior generations even after paraphrasing. Krishna et al. (2023) show paraphrasing with their paragraph-level model (DIPPER) can evade several detectors; for example, they report DetectGPT accuracy dropping from 70.3% to 4.6% at 1% false-positive rate in one setting, while retrieval-based defenses recover substantial detection ability.

2.2 Robustness under distribution shift and OOD evaluation

OOD robustness in machine generated text detection is multi-factor: shifts arise from generator family/version, decoding strategy, prompt style, domain/genre/register, and post-processing pipelines (paraphrasing/human editing); some benchmarks also include multilingual test cases (Dugan et al., 2024; Li et al., 2024). Robust evaluation therefore requires separating (i) domain shift in human text, (ii) generator shift in machine text, and (iii) transformation shift induced by editing/paraphrasing (Dugan et al., 2024; Sadasivan et al., 2023; Krishna et al., 2023).

2.2.1 Decoding and sampling shifts

Decoding strategy is a primary confound. Ippolito et al. benchmark top-k, nucleus sampling, and untruncated sampling and show that "better at fooling humans" decoding can introduce measurable abnormalities that make automatic detection easier (Ippolito et al., 2020). RAID further emphasizes that variations in sampling strategies and repetition penalties can significantly degrade detector robustness (Dugan et al., 2024).

2.2.2 Post-processing and paraphrasing attacks

Paraphrasing attacks are damaging because they remove surface artifacts while preserving semantics. Sadasivan et al. (2023) introduce recursive paraphrasing

attacks and stress-test multiple detector classes, highlighting vulnerabilities under attacker-controlled rewriting. Krishna et al. (2023) similarly show paraphrasing evasion at scale and propose retrieval as a defense, arguing that evaluation should treat post-processed (paraphrased/edited) machine text as a core regime.

2.2.3 Fairness and false positives under shift

Robustness is intertwined with fairness because many detectors correlate "AI-ness" with predictability/perplexity cues that also correlate with writing proficiency and genre. Liang et al. (2023) show that several GPT detectors systematically misclassify non-native English writing as AI-generated more often than native writing, raising deployment risks in educational settings.

2.2.4 Metrics, thresholds and calibration

Most papers report AUROC because it summarizes separability independent of a single threshold (Fawcett, 2006). However, deployment requires thresholding; thus operating-point metrics (e.g., TPR at low FPR) and calibration are essential. Guo et al. show modern neural networks can be poorly calibrated and that temperature scaling can help in-distribution (Guo et al., 2017), while Ovadia et al. (2019) show calibration and uncertainty can deteriorate under dataset shift even when post-hoc methods are applied.

2.2.5 Connections to general OOD detection and domain generalization

Classic OOD detection baselines show discriminative models can be overconfident off-distribution (Hendrycks & Gimpel, 2017). Domain generalization surveys emphasize there is no universal method that dominates across shifts and highlight families of approaches (domain-invariant representations, augmentation, meta-learning, ensembling) (Zhou et al., 2021).

2.3 Datasets and benchmark suites used to evaluate detectors

Dataset design often determines whether "high accuracy" reflects realistic robustness. Earlier resources often used a single generator and narrow domain;

modern suites expand coverage across generators, domains, decoding strategies, and adversarial transformations; some suites also include multilingual evaluation (Dugan et al., 2024; Li et al., 2024).

Below is a structured catalog of widely used datasets/benchmarks, with commentary on properties that matter for evaluation.

2.3.1 News and generator-matched settings

Grover / neural fake news: Zellers et al. (2019) introduce Grover and show that generator-matched defenses can be strong; they report discriminator accuracy of **73%** and that Grover itself can reach 92% accuracy in their setting, highlighting both power and specificity of matched detection.

2.3.2 Multi-generator benchmark environments

User trust depends on communicating detector limitations, avoiding overconfident claims, and using human-in-the-loop review where appropriate (Lee & See, 2004). Because detection errors can have serious consequences, acceptance requires not only accuracy but also calibrated uncertainty and clear operational guidance (Guo et al., 2017; Lee & See, 2004).

2.3.3 ChatGPT-focused and semantic-invariant task datasets

HC3: Guo et al. (2023) introduce the Human ChatGPT Comparison Corpus with Q&A across multiple expert domains (open, finance, medical, legal, psychology) and evaluate detection under this QA framing.

HC3 Plus: Su et al. (2023b) extend HC3 toward semantic-invariant tasks (e.g., translation/summarization/paraphrasing), arguing detection is harder when outputs are constrained by the input semantics.

2.3.4 Multilingual detection benchmarks

MULTITuDE: Macko et al. (2023) provide a multilingual benchmark with 74,081 texts across 11 languages and generations from multiple multilingual LLMs, designed to test generalization to unseen languages and unseen LLMs.

2.3.5 Benchmark frameworks with robustness attacks

MGTBench: He et al. (2023) propose a benchmark framework for evaluating detection against powerful LLMs and explicitly test robustness under attacks such as paraphrasing and perturbations, reporting that attacks can significantly diminish detection effectiveness.

2.3.6 Shared tasks and community datasets

AuTexTification (IberLEF 2023): Sarvazyan et al. (2023) describe a shared task with 160K+ texts across English/Spanish and multiple domains, enabling standardized binary detection and model attribution evaluation.

SemEval-2024 Task 8: Wang et al. (2024) report the main findings of Task 8 on multidomain, multimodel, multilingual machine generated text detection, including multiple subtasks and large community participation.

2.3.7 Robustness mega-benchmarks

RAID: Dugan et al. (2024) introduce a robustness benchmark with 6M+ generations spanning 11 models, 8 domains, 11 adversarial attacks, and 4 decoding strategies, and show that many detectors degrade under adversarial edits and sampling variations.

2.3.8 MAGE: benchmark testbeds only (as required)

MAGE is best treated as an OOD benchmark suite (testbeds + findings), not as a detection method. Li et al. build a comprehensive testbed for "in the wild" machine generated text detection, organizing data into 8 testbeds with progressively higher "wildness," and explicitly evaluating robustness under unseen domains, unseen models, and paraphrasing attacks (Li et al., 2024).

Key benchmark findings include:

In OOD settings, even the best-performing detector in their evaluation misclassifies 61.95% of human-written texts from unseen domains (Li et al., 2024).

They report that OOD performance can be substantially improved by using only **0.1%** in-domain data to adjust the decision boundary, illustrating the importance of decision-threshold selection beyond AUROC (Li et al., 2024).

2.4 Knowledge distillation in text classification and detection

Knowledge distillation transfers information from a high-capacity teacher to a smaller student using soft targets (temperature-scaled distributions) alongside hard labels (Hinton et al., 2015). In NLP, distillation is widely used for compression and deployment: DistilBERT demonstrates that a smaller distilled transformer can retain strong downstream performance with reduced inference cost (Sanh et al., 2019), and TinyBERT proposes task-specific distillation procedures that compress BERT-like models while retaining accuracy (Jiao et al., 2020).

Beyond compression, distillation can act like data-dependent regularization that reduces overconfidence; Yuan et al. (2020) analyze links between distillation and label-smoothing regularization. For AI-text detection specifically, whether distillation improves OOD robustness is an empirical question that should be validated on suites like RAID and MAGE rather than only i.i.d. splits (Dugan et al., 2024; Li et al., 2024).

2.5 Metric and contrastive objectives for robust representations

A complementary robustness strategy shapes embedding geometry rather than optimizing only cross-entropy. Triplet loss is a canonical metric-learning objective encouraging anchors to be closer to positives than negatives by a margin (Schroff et al., 2015). Supervised contrastive learning generalizes contrastive objectives to labeled settings by pulling together same-class representations and pushing apart different-class representations (Khosla et al., 2020). Contrastive training has also been linked to improved OOD detection performance in standard OOD benchmarks, especially under “near-OOD” regimes (Winkens et al., 2020).

For machine generated text detection, these objectives are motivated by the observation that detectors can conflate domain style or writing proficiency with provenance leading to OOD false positives (Liang et al., 2023; Li et al., 2024).

2.6 Synthesis and open problems

Across methods, the central barrier is not achieving high accuracy on fixed i.i.d. datasets, but achieving robustness under realistic distribution shift. RAID and MAGE show that detectors can appear strong in narrow settings yet fail under unseen domains, unseen models, decoding changes, and paraphrasing often in ways that create unacceptable false-positive risks on human writing (Dugan et al., 2024; Li et al., 2024). This motivates research that (i) reports robustness across explicit shift regimes, (ii) treats calibration and operating points as first-class, and (iii) validates improvements on shared robustness suites rather than isolated datasets (Fawcett, 2006; Guo et al., 2017; Dugan et al., 2024).

3 The Proposed Method

This chapter describes a teacher-student framework that improves out-of-distribution (OOD) generalization for machine generated text detection by combining knowledge distillation with representation-shaping objectives. The core idea is to transfer "dark knowledge" from a higher-capacity teacher detector to a smaller student detector using soft targets, while simultaneously enforcing discriminative structure in the student's latent space using metric-learning and supervised contrastive losses. Knowledge distillation is implemented through a temperature-scaled KL-divergence between teacher and student predictive distributions (Hinton et al., 2015). Representation shaping is implemented through (i) triplet loss that introduces explicit margins between classes (Schroff et al., 2015) and (ii) supervised contrastive learning that uses label information to pull same-class examples together and push different-class examples apart (Khosla et al., 2020).

3.1 Task definition and decision rule

Let x denote a text input (a sequence of tokens) and $y \in \{0,1\}$ denote its provenance label. In this thesis, we use a binary detector that assigns $y = 1$ to human-written text and $y = 0$ to machine generated text. A detector parameterized by θ outputs logits $z_\theta(x) \in \mathbb{R}^2$and the corresponding class probabilities via softmax,

$$p_\theta(y = k \mid x) = \frac{\exp\left(z_{\theta,k}(x)\right)}{\sum_{j\in\{0,1\}} \exp\left(z_{\theta,j}(x)\right)}$$

The operational score is the probability of the human class,

$$s_\theta(x) = p_\theta(y = 1 \mid x) \in [0,1]$$

A hard prediction is produced through thresholding:

$$\hat{y}(x;\tau) = \mathbb{I}[s_\theta(x) \geq \tau]$$

Equivalently, the decision rule may be written explicitly as

$$\hat{y}(x;\tau) = \begin{cases} 1, & if\ s_\theta(x) \geq \tau, \\ 0, & if\ s_\theta(x) < \tau, \end{cases}$$

This explicit separation between a continuous score $s_\theta(x)$ and a threshold τ is crucial because OOD shifts often change the calibration and distribution of scores even when ranking quality (e.g., AUROC) remains relatively stable (Fawcett, 2006; Guo et al., 2017; Ovadia et al., 2019).

3.2 Teacher-Student Architecture

Both teacher and student detectors share the same high-level architecture: an encoder-only Transformer backbone followed by a small classification head. Given a tokenized input x with T tokens, the backbone produces contextual representations

$$H_\theta(x) \in \mathbb{R}^{T\times d},$$

where d is the hidden size of the backbone. A pooled sequence representation is extracted using the first-special-token embedding:

$$h_\theta(x) = H_\theta(x)_1 \in \mathbb{R}^d.$$

where $H_\theta(x)_1$denotes the representation of the first special token (e.g., <s> for RoBERTa and [CLS] for BERT-style models).

A linear classifier produces logits:

$$z_\theta(x) = Wh_\theta(x) + b,\ \ W \in \mathbb{R}^{2\times d}, b \in \mathbb{R}^2$$

This uniform architecture ensures that observed improvements can be attributed to training objectives (KD and auxiliary losses) and capacity differences (teacher vs. student), rather than to incompatible model heads or different feature extraction strategies.

3.3 Choice of teacher and student models

The teacher should be sufficiently expressive so that its soft predictions and internal representations are informative and stable. The student should be significantly smaller to reduce training and inference cost while maintaining strong detection performance after distillation.

In this thesis, the teacher models are RoBERTa-base (Liu et al., 2019) and DeBERTa-base (He et al., 2021). The student models are DistilBERT-base and MiniLM-L12-H384, representing two widely used compression families: layer reduction through distillation (DistilBERT) and self-attention distillation with a smaller hidden size (MiniLM) (Sanh et al., 2019; Wang et al., 2020).

Table 1: Teacher and student model architecture and size

Model	**Role**	**Layers**	**Hidden size (d)**	**Attention heads**	**Approx. parameters**
DeBERTa-base	Teacher	12	768	12	~140M
RoBERTa-base	Teacher	12	768	12	~125M
DistilBERT-base (distilbert-base-uncased)	Student	6	768	12	~66M
MiniLM-L12-H384 (MiniLMv1-L12-H384-uncased)	Student	12	384	12	~33M

A practical consequence of using MiniLM-L12-H384 is the hidden-size mismatch between the teacher models and the student model (768 vs. 384). Because several objectives in this thesis operate on representations, we introduce an alignment mapping g_ϕ from the student embedding space to the teacher embedding space:

$$\tilde{h}_S(x) = g_\phi(h_S(x)) = \begin{cases} h_S(x), & d_S = d_T, \\ Ah_S(x) + c, & d_S \neq d_T, \end{cases}$$

where $A \in \mathbb{R}^{d_T \times d_S}$ and $c \in \mathbb{R}^{d_T}$ are learned parameters. This allows student representations to be mapped into the teacher-dimensional space before applying representation-level losses, without forcing architectural changes to either backbone.

3.4 Knowledge distillation

Knowledge distillation transfers predictive behavior from teacher to student by training the student to match the teacher's output distribution, not only the hard label. This provides richer supervision because the teacher's relative probabilities across classes encode uncertainty and class similarity information that is not present in one-hot labels (Hinton et al., 2015).

Let $z_T(x)$ and $z_S(x)$ be teacher and student logits for input x. Temperature scaling with $T_{KD} > 0$ produces softened class distributions:

$$q_T^{(T_{KD})}(x) = \text{softmax}\left(\frac{z_T(x)}{T_{KD}}\right), q_S^{(T_{KD})}(x) = \text{softmax}\left(\frac{z_S(x)}{T_{KD}}\right).$$

The distillation loss is the KL-divergence from teacher to student:

$$\mathcal{L}_{KD}(x) = \text{KL}\left(q_T^{(T_{KD})}(x) \parallel q_S^{(T_{KD})}(x)\right) = \sum_{k \in \{0,1\}} q_{T,k}^{(T_{KD})}(x) \log \frac{q_{T,k}^{(T_{KD})}(x)}{q_{S,k}^{(T_{KD})}(x)}.$$

Following standard distillation practice, the term is scaled by T_{KD}^2 so that gradient magnitudes remain comparable when using larger temperatures (Hinton et al., 2015).

In the binary setting, KD can be interpreted as encouraging the student's logit difference $z_{S,1}(x) - z_{S,0}(x)$ to match the teacher's implied margin between human and machine classes. This is valuable for OOD generalization because teacher soft targets can reduce overconfident, shortcut-driven behavior that often fails under shifts.

3.5 Supervised cross-entropy loss

All Cross-entropy anchors the detector to ground-truth labels and prevents purely imitative training from inheriting teacher errors. For a labeled pair (x, y), the standard supervised loss is:

$$\mathcal{L}_{CE}(x, y) = -\log\, p_S(y \mid x).$$

Cross-entropy is also necessary to keep the decision boundary aligned with the target task when the teacher is imperfect or when the student's capacity limits exact imitation.

3.6 Triplet loss for margin-based structure

Triplet loss is a metric-learning objective that enforces a margin between same-class and different-class samples in embedding space (Schroff et al., 2015). It is designed to learn an embedding space where same-class examples are closer than different-class examples by at least a margin (Schroff et al., 2015).

A triplet consists of an anchor a, a positive $p$of the same class, and a negative $n$of the opposite class. Using a distance $d(\cdot,\cdot)$ and margin $m > 0$, the triplet loss is:

$$\mathcal{L}_{Tri} = \max\,(0, d(a, p) - d(a, n) + m).$$

In our teacher-student setting, we use a cross-view triplet construction that ties student representations to teacher representations while preserving class separation. For each example $x_i$in a mini-batch, we define the anchor as the aligned student embedding $\tilde{h}_S(x_i)$ and define the positive as the teacher embedding $h_T(x_i)$. The negative is selected from teacher embeddings of opposite-label examples in the same mini-batch. With cosine distance,

$$d(u, v) = 1 - \frac{u^\top v}{\| u \| \ \| v \|},$$

the per-sample loss becomes

$$\mathcal{L}_{Tri}(i) = \max(0, d(\tilde{h}_S(x_i), h_T(x_i)) - d(\tilde{h}_S(x_i), h_T(x_{j(i)})) + m),$$

where $j(i)$ indexes a negative example such that $y_{j(i)} \neq y_i$. This construction simultaneously pushes the student toward teacher-consistent geometry and enforces a margin between human and machine clusters.

A diagnostic ablation sometimes used in practice is a logit-space triplet, where anchors/positives/negatives are logits rather than embeddings:

$$a = z_S(x_i), p = z_T(x_i), n = z_T(x_{j(i)}),$$

which tests whether explicit margins in decision space alone can improve robustness without relying on deeper representational alignment.

3.7 Supervised Contrastive Loss

This Supervised contrastive learning generalizes contrastive objectives by using labels to define positives and negatives within the batch, explicitly clustering same-class examples and separating different classes (Khosla et al., 2020).

Let a batch contain Bfeature vectors $v_i$with labels y_i. Define normalized vectors $\bar{v}_i = v_i / \| v_i \|$and similarity

$$s_{i,j} = \frac{\bar{v}_i^{\top} \bar{v}_j}{\tau_c},$$

where $\tau_c > 0$is the contrastive temperature. For each anchor i, define the set of positives $P(i) = \{p \neq i : y_p = y_i\}$. The supervised contrastive loss is:

$$\mathcal{L}_{SupCon} = \sum_{i=1}^{B} \frac{-1}{| P(i) |} \sum_{p \in P(i)} \log \frac{\exp(s_{i,p})}{\sum_{a \neq i} \exp(s_{i,a})}.$$

Unlike triplet loss, which relies on a specific negative, supervised contrastive loss leverages all valid positives and negatives within the batch, often providing more stable optimization and stronger clustering effects (Khosla et al., 2020).

To directly connect student features to teacher features, we employ a multi-view construction in which teacher and student representations act as two "views" of the same labeled example. For a batch of size B, we create a $2B$-sized set of features by concatenating student-aligned embeddings and teacher embeddings:

$$v = \left[\tilde{h}_S(x_1), \dots, \tilde{h}_S(x_B), h_T(x_1), \dots, h_T(x_B)\right], \qquad y = [y_1, \dots, y_B, y_1, \dots, y_B].$$

This makes teacher and student embeddings of the same class contribute as additional positives, encouraging teacher-consistent clustering while preserving label-based discrimination.

3.8 Combined objective and loss configurations

The proposed method defines a family of student training objectives obtained by combining cross-entropy, distillation, and representation-shaping losses. The base distillation-supervision mixture is a weighted sum of hard-label cross-entropy and temperature-scaled KL distillation:

$$\mathcal{L}_{CE\text{-}KD}(x, y) = \alpha\, \mathcal{L}_{CE}(x, y) + (1 - \alpha)\, T_{KD}^2\, \mathcal{L}_{KD}(x),$$

where $\alpha \in [0,1]$controls the trade-off between matching labels and matching teacher soft outputs (Hinton et al., 2015).

When triplet and supervised contrastive losses are enabled, the full student objective becomes:

$$\mathcal{L}_S = \alpha\, \mathcal{L}_{CE} + (1 - \alpha)\, T_{KD}^2\, \mathcal{L}_{KD} + \lambda_{Tri}\, \mathcal{L}_{Tri} + \lambda_{SupCon}\, \mathcal{L}_{SupCon},$$

with nonnegative weights $\lambda_{Tri}$and λ_{SupCon}controlling the strength of geometry constraints relative to classification/distillation terms.

In practice, each loss component targets a different failure mode that is common under OOD shifts. Cross-entropy ensures correctness with respect to labels, KD encourages smoother, teacher-consistent decision behavior, triplet loss

introduces explicit margins that can make the boundary less sensitive to small distributional changes, and supervised contrastive loss promotes compact same-class clusters that can reduce overlap between human-like and machine-like regions in representation space (Hinton et al., 2015; Schroff et al., 2015; Khosla et al., 2020).

Teacher training is typically performed with cross-entropy to produce a strong and stable detector, and then the student is trained using one of the above combined objectives. This two-stage design is consistent with the standard teacher-student paradigm in which the teacher remains fixed during student training, so that the student's learning signal remains stationary (Hinton et al., 2015).

3.9 Threshold determination

Because the final output is a continuous score $s_\theta(x)$, threshold selection is required to deploy a binary detector. Threshold choice is especially important under OOD conditions because score calibration can change across domains, generators, or writing styles, shifting the optimal operating point even when the ranking quality measured by AUROC remains similar (Fawcett, 2006; Guo et al., 2017; Ovadia et al., 2019).

Let $V = \{(x_i, y_i)\}_{i=1}^{M}$ be a validation set intended to reflect deployment conditions as closely as possible. For any threshold τ, define predictions $\hat{y}_i(\tau) = \mathbb{I}[s_\theta(x_i) \geq \tau]$. Class-wise recalls are:

$$\mathrm{Rec}_{human}(\tau) = \frac{\sum_{i=1}^{M} \mathbb{I}[\, y_i = 1 \wedge \hat{y}_i(\tau) = 1]}{\sum_{i=1}^{M} \mathbb{I}[\, y_i = 1]}, \mathrm{Rec}_{machine}(\tau)$$

$$= \frac{\sum_{i=1}^{M} \mathbb{I}[\, y_i = 0 \wedge \hat{y}_i(\tau) = 0]}{\sum_{i=1}^{M} \mathbb{I}[\, y_i = 0]}.$$

A balanced operating criterion that weighs both classes equally is the average recall:

$$\mathrm{AvgRec}(\tau) = \frac{1}{2}(\mathrm{Rec}_{human}(\tau) + \mathrm{Rec}_{machine}(\tau)).$$

A general threshold-selection rule consistent with this metric is:

$$\tau^{\star} = \arg \max_{\tau \in [0,1]} \mathrm{AvgRec}(\tau).$$

In practice, $\tau^{\star}$is found by evaluating $\mathrm{AvgRec}(\tau)$on a dense grid of thresholds or on the set of unique validation scores. This procedure is model-agnostic and does not depend on any particular benchmark.

ROC-based operating points provide an alternative, widely used view of thresholding. Each threshold corresponds to a point on the ROC curve; ROC analysis explicitly characterizes the trade-off between true positive rate and false positive rate across thresholds (Fawcett, 2006). A classic scalar criterion is Youden's Jstatistic (Youden, 1950),

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau),$$

and choosing $\tau$to maximize $J(\tau)$corresponds to maximizing the sum of sensitivity and specificity minus one, giving equal weight to both error types.

In real deployments, misclassification costs may be asymmetric; for instance, falsely accusing a human author (false positive for "machine") can be more harmful than missing a machine generated text, or vice versa. A cost-sensitive threshold can be selected by minimizing empirical expected risk on validation data, or when probabilities are well calibrated by using a Bayes-optimal threshold derived from costs. Because neural networks are often miscalibrated, post-hoc calibration can be applied before thresholding. Temperature scaling is a standard calibration method that fits a single parameter T_{cal}on a held-out set and rescales logits $z(x) \mapsto z(x)/T_{cal}$to improve probability calibration without changing ranking much (Guo et al., 2017).

Finally, the threshold criterion should match the evaluation or deployment objective. If one wishes to optimize an F_1measure rather than balanced recall, the optimal threshold can behave differently and can depend strongly on prevalence and calibration; theoretical analysis highlights that "optimal threshold for F_1" is not generally equivalent to $\tau = 0.5$and can have unintuitive behavior when the classifier is weak (Lipton et al., 2014). This reinforces the general principle used in this thesis: threshold selection is treated as a formal optimization step that is

explicitly tied to the chosen operating metric and validated on a held-out set, rather than fixed by convention.

4 Experiments

This chapter presents the experimental setup used to evaluate out-of-distribution (OOD) detection of machine generated text. It describes the MAGE benchmark and the specific OOD testbeds used in this thesis, the model variants examined (with and without knowledge distillation and with different training objectives), the baselines used for comparison, the compute-aware workflow adopted to manage computationally expensive testbeds, and the resulting quantitative and qualitative findings.

4.1 Benchmark: MAGE (Machine Generated Text Detection in the Wild)

All experiments are conducted on the MAGE benchmark (Li et al., 2024), which evaluates machine generated text detection under progressively harder distribution shifts. MAGE consists of eight testbeds that vary the domain, the generator (LLM), and additional transformations (e.g., paraphrasing). This thesis focuses on the OOD testbeds that best reflect real deployment conditions.

Table 2: MAGE OOD Testbeds and shift types considered

Testbed	Shift type	What is held out?	OOD relevance
TB5	Unseen models	Generator/model family	Evaluates robustness to new generators not seen during training.
TB6	Unseen domains	Text domain/task	Evaluates robustness to unseen writing styles and topical distributions.
TB7	Unseen domains + unseen model	New domains and a new generator	Represents a practical combined-shift setting (harder

			'in-the-wild' scenario).
TB8	Paraphrasing attack	Paraphrased outputs	Stress-test robustness under distribution manipulation designed to hide surface-level cues.

The evaluation metrics follow MAGE: HumanRec (recall on human-written text), MachineRec (recall on machine generated text), AvgRec (the mean of HumanRec and MachineRec), and AUROC. AvgRec reflects performance at a specific operating threshold and is therefore threshold-dependent, whereas AUROC measures threshold-independent separability between the two classes.

4.2 Models, Baselines and Training Objectives

This subchapter describes the detector architectures and training objectives used in the experimental study. The design is intended to evaluate whether compact student detectors can retain strong out-of-distribution (OOD) detection performance and whether this performance can be improved through teacher guidance at the logit level, at the representation level, or through a combination of both. Across the full study, RoBERTa-base and DeBERTa-base are used as teacher models, while MiniLM and DistilBERT are used as student models.

4.2.1 Detector Architectures

All detectors in this thesis share the same overall architecture. A pretrained Transformer encoder is followed by a dropout layer and a linear classification head that predicts two classes: human-written and machine generated text. The classifier operates on the contextual representation of the first token, which functions as a pooled sequence representation. Inputs are truncated to a maximum length of 512 tokens, ensuring a common computational setting across all teacher and student experiments.

The teacher models serve as high-capacity reference detectors. Their role is twofold: first, they provide a direct upper-bound baseline without compression; second, they provide training signals for the student models in the distillation settings. The student models are deliberately smaller and computationally cheaper, allowing the thesis to examine whether OOD robustness can be preserved under stricter efficiency constraints.

4.2.2 Experimental families

The experiments are organized into three main families. The first family consists of teacher-only baselines. These runs evaluate the teacher models directly, without any student compression. They establish the reference performance of the larger encoders and are essential for determining how much performance is lost, preserved, or reshaped after distillation.

The second family consists of student-only baselines. These runs train the student model without any teacher information. They provide the cleanest measure of what a compact detector can achieve using only supervision from the ground-truth labels.

The third family consists of teacher-guided student variants. In these runs, the student is trained with additional information derived from the teacher. Depending on the variant, this guidance may come from the teacher's softened output distribution, from the teacher's learned representations, or from both simultaneously.

4.2.3 Core training objectives

Four main loss components are used in different combinations. Cross-entropy (CE) is the standard supervised classification objective. It uses the ground-truth label as the primary learning signal and serves as the backbone objective for both teacher-only and student-only baselines.

Knowledge distillation (KD) transfers information from the teacher's output distribution to the student. In the implementation, KD is realized through KL-divergence between the teacher's softened logits and the student's softened logits, with temperature scaling. In the hybrid variants, this distillation term is

combined with CE, so that the student learns simultaneously from hard labels and from teacher soft targets.

Triplet loss is used to shape the representation space. In the main teacher-guided triplet variants, the student representation is trained to align with the teacher representation through a cross-view triplet objective. In the student-only triplet variants, triplet loss is applied in a supervised in-batch manner directly on the student's own embeddings.

Supervised contrastive loss (SupCon) also acts at the representation level. It encourages examples from the same class to cluster together and examples from different classes to separate. In the teacher-guided contrastive variants, teacher and student representations are paired in the contrastive space; in the student-only contrastive variants, the objective is applied only to student embeddings.

4.2.4 Main model variants

For clarity, the thesis refers to experimental variants using descriptive names rather than implementation-specific identifiers. The principal configurations are summarized below.

Table 3: Main Model Variants

Variant family	**Model variant**	**Teacher used during training?**	**Main objective**
Teacher-only baseline	Teacher-only supervised detector	No	CE
Teacher-only enhanced baseline	Teacher-only CE + triplet detector	No	CE + triplet
Student-only baseline	Student-only supervised detector	No	CE
Student-only metric baseline	Student-only CE + triplet detector	No	CE + triplet

Variant family	Model variant	Teacher used during training?	Main objective
Student-only contrastive baseline	Student-only CE + supervised contrastive detector	No	CE + SupCon
Student-only multi-loss baseline	Student-only CE + triplet + supervised contrastive detector	No	CE + triplet + SupCon
Logit-distilled student	Student CE + KD detector	Yes	CE + KD
Feature-distilled student	Student CE + teacher-triplet detector	Yes	CE + teacher-guided triplet
Feature-distilled student	Student CE + teacher-SupCon detector	Yes	CE + teacher-guided SupCon
Feature-distilled student	Student CE + teacher-triplet + teacher-SupCon detector	Yes	CE + two teacher-guided feature losses
Hybrid-distilled student	Student CE + KD + teacher-triplet detector	Yes	CE + KD + teacher-guided triplet
Hybrid-distilled student	Student CE + KD + teacher-SupCon detector	Yes	CE + KD + teacher-guided SupCon
Full hybrid-distilled student	Student CE + KD + teacher-triplet + teacher-SupCon detector	Yes	CE + KD + two teacher-guided feature losses

4.2.5 Teacher only and student only baselines

The teacher-only supervised detector is the primary uncompressed baseline. It shows how well a high-capacity encoder performs without any student compression. A second teacher-only variant combines CE with triplet loss in order to test whether representation shaping also benefits the teacher itself.

The student-only supervised detector is the primary compact baseline. It indicates what a lightweight model can achieve without teacher guidance. The student-only metric and contrastive baselines extend this comparison by adding triplet loss, supervised contrastive loss, or both. These variants are scientifically important because they distinguish the effect of representation shaping itself from the effect of teacher guidance.

4.2.6 Teacher-guided variants

The teacher-guided student variants are divided into two categories. The first category consists of feature-distillation variants. These models do not use logit-level KD, but they do use teacher representations to shape the student embedding space. In the triplet-based version, the student is pulled toward the teacher representation while separating from mismatched examples through a margin-based metric objective. In the contrastive version, the student and teacher representations are paired within a supervised contrastive space. A third feature-distillation variant combines both mechanisms.

The second category consists of hybrid distillation variants. These combine CE + KD with one or more teacher-guided feature losses. They are the most complete teacher-guided settings in the study because they transfer information at both the decision level and the representation level.

4.2.7 Implementation details

An important implementation detail concerns representation alignment across heterogeneous backbones. When teacher-guided representation losses are used and the teacher and student hidden sizes differ, the student representation is projected into the teacher representation space before alignment is applied. In addition, the implementation contains a small number of legacy diagnostic variants in which triplet loss is used without the standard supervised CE term, and in some cases on logits rather than on the main embedding space. These auxiliary variants are not central to the main thesis conclusions, but they help explain a small number of unstable or strongly asymmetric results.

4.3 Experimental Workflow

Testbeds 5 and 6 are computationally expensive because they require multiple folds (cross-validation over unseen models or unseen domains). To control computational cost while keeping the selection process explicit, experiments follow a staged workflow:

1. TB7 (default boundary): evaluate teacher-only, student-only, and teacher-guided configurations, and select the best-performing distilled configuration according to AvgRec.
2. TB6: calibrate the decision boundary by sweeping thresholds and selecting the value that maximizes AvgRec in the unseen-domain setting.
3. TB5: evaluate the selected model before and after applying the calibrated threshold in the unseen-model setting.
4. TB7 (after adjustment): re-evaluate the selected configuration using the calibrated threshold to report the final operating point under the main OOD setting.
5. TB8: measure paraphrasing robustness before and after applying the same threshold adjustment.

Threshold calibration follows the MAGE boundary-adjustment protocol by using 0.1% in-domain calibration data within Testbed 6. For each unseen-domain fold k, a threshold t_k^* is selected to maximize AvgRec on the corresponding calibration subset, and the final refined threshold is obtained by averaging these fold-specific thresholds, $t^* = \frac{1}{K}\sum_{k=1}^{K} t_k^*$. Because AUROC is threshold-independent, calibration changes the operating point (i.e., the balance between HumanRec and MachineRec) without changing AUROC.

4.4 Quantitative Results

This section reports the quantitative results of the proposed detectors on the MAGE testbeds. Following the staged workflow of the study, Testbed 7 is first used to select the best performing distilled configuration under the default decision boundary, Testbed 6 is then used for threshold refinement, and the refined threshold is subsequently transferred to Testbeds 5, 7, and 8. To keep the presentation concise and focused, only the most informative configurations are reported.

4.4.1 Results on Testbeds 7 & 8 with RoBERTa-Base as teacher

The following table summarizes the most relevant RoBERTa-based configurations on the two main downstream OOD testbeds using the default decision boundary.

Table 4: Selected RoBERTa-based results On Testbeds 7 & 8 (before refinement)

Configuration	Testbed 7 AvgRec	Testbed 7 AUROC	Testbed 8 AvgRec	Testbed 8 AUROC
Teacher-only supervised CE	72.26%	0.9483	65.98%	0.7450
Teacher-only CE + triplet	70.88%	0.9543	65.75%	0.7508
MiniLM, student-only supervised CE	73.27%	0.9359	65.83%	0.7392
MiniLM, student-only CE + triplet	79.70%	0.9389	66.39%	0.7283
DistilBERT, student-only supervised CE	78.46%	0.9471	71.59%	0.8123
DistilBERT, student-only CE + SupCon	80.59%	0.9086	69.46%	0.7562
MiniLM, student CE + KD	69.64%	0.9479	65.92%	0.7844
MiniLM, student CE + teacher-triplet distillation	79.50%	0.9402	66.34%	0.7349
MiniLM, student CE + teacher-triplet + teacher-SupCon distillation	78.65%	0.9411	70.37%	0.7784

For the RoBERTa teacher, the best overall result on Testbed 7 under the default boundary is the student-only DistilBERT model trained with CE + SupCon, which reaches 80.59% AvgRec. The strongest distilled RoBERTa-based configuration is RoBERTa-base - MiniLM with CE + teacher-triplet distillation, which reaches 79.50% AvgRec on Testbed 7 and is therefore selected for threshold refinement. On Testbed 8, the strongest RoBERTa-based distilled default result is RoBERTa-base - MiniLM with CE + teacher-triplet + teacher-SupCon distillation at 70.37%

AvgRec, while the best overall RoBERTa-based default result is DistilBERT student-only CE at 71.59%. Thus, under the RoBERTa teacher, distillation produces competitive but not uniformly dominant performance.

4.4.2 Threshold refinement for the selected RoBERTa-based distilled model

Table 5: Threshold refinement for the selected RoBERTa-based distilled model

Testbed	Default AvgRec	Refined AvgRec	AUROC	Refined threshold
Testbed 5	84.76%	84.81%	0.9379	0.0114
Testbed 6	67.03%	77.87%	0.8492	0.0114
Testbed 7	79.50%	85.81%	0.9402	0.0114
Testbed 8	66.34%	64.38%	0.7349	0.0114

Threshold refinement improves the selected RoBERTa-based distilled detector substantially on Testbed 6 and Testbed 7, with the largest gain observed on Testbed 7, where AvgRec increases from 79.50% to 85.81%. On Testbed 5, the effect is essentially neutral, since AvgRec changes only from 84.76% to 84.81%. By contrast, the same refined threshold does not transfer well to Testbed 8, where AvgRec decreases from 66.34% to 64.38%.

4.4.3 Results on Testbeds 7 & 8 with DeBERTa-base as teacher

Table 6: Selected DeBERTa-based results On Testbeds 7 & 8 (before refinement)

Configuration	Testbed 7 AvgRec	Testbed 7 AUROC	Testbed 8 AvgRec	Testbed 8 AUROC
Teacher-only supervised CE	73.58%	0.8839	65.42%	0.7082
Teacher-only CE + triplet	76.08%	0.9427	62.77%	0.6697
MiniLM, student-only supervised CE	75.23%	0.9192	63.48%	0.6752
MiniLM, student-only CE + SupCon	79.64%	0.9323	66.39%	0.7116

Configuration	Testbed 7 AvgRec	Testbed 7 AUROC	Testbed 8 AvgRec	Testbed 8 AUROC
DistilBERT, student-only supervised CE	72.80%	0.8918	65.92%	0.7525
DistilBERT, student-only CE + triplet	80.15%	0.9470	69.52%	0.7788
DistilBERT, student CE + KD	72.93%	0.8974	67.22%	0.7527
DistilBERT, student CE + teacher-triplet distillation	79.12%	0.9049	64.93%	0.7139
DistilBERT, student CE + KD + teacher-triplet distillation	79.65%	0.9380	72.77%	0.8057

For the DeBERTa-base teacher, the best overall result on Testbed 7 under the default boundary is again a student-only model, namely DistilBERT with CE + triplet, which reaches 80.15% AvgRec. However, the strongest distilled DeBERTa-based configuration is DeBERTa-base - DistilBERT with CE + KD + teacher-triplet distillation, which reaches 79.65% AvgRec on Testbed 7 and becomes the selected distilled candidate for threshold refinement. As in the RoBERTa-based setting, this choice reflects the distillation-focused selection criterion of the staged workflow rather than the strongest overall result across all evaluated models. On Testbed 8, this same hybrid-distilled model gives the strongest default result, namely 72.77% AvgRec with 0.8057 AUROC, making it the strongest final distilled detector in the study.

4.4.4 Threshold refinement for the selected DeBERTa-based distilled model

As before, the refined threshold is estimated on Testbed 6 and transferred to Testbeds 5, 7, and 8.

Table 7: Threshold refinement for the selected DeBERTa-based distilled model

Testbed	Default AvgRec	Refined AvgRec	AUROC	Refined threshold
Testbed 5	85.38%	85.63%	0.9262	0.0206
Testbed 6	65.81%	73.23%	0.8815	0.0206
Testbed 7	79.65%	85.64%	0.9380	0.0206
Testbed 8	72.77%	71.60%	0.8057	0.0206

The same overall refinement pattern appears here. The refined threshold improves AvgRec strongly on Testbed 6 and Testbed 7, with Testbed 7 rising from 79.65% to 85.64%. On Testbed 5, the gain is again small, from 85.38% to 85.63%. On Testbed 8, threshold refinement causes a slight decrease, from 72.77% to 71.60%. Thus, threshold refinement is beneficial for the standard OOD settings but does not transfer cleanly to the paraphrasing-attack setting. Even so, the DeBERTa-based detector remains substantially stronger on Testbed 8 than the corresponding RoBERTa-based selected model.

4.4.5 Comparison with Longformer

In the original MAGE paper, Longformer is the best performing detector among the compared methods and outperforms other commonly used PLM backbones such as BERT, RoBERTa, and GPT-2. Since this thesis is primarily concerned with the post-adjustment evaluation setting, the comparison in this subsection focuses only on the refined results after decision-boundary adjustment. For Longformer, MAGE reports refined AvgRec values of 87.62% on Testbed 5, 81.78% on Testbed 6, 86.54% on Testbed 7, and 62.92% on Testbed 8. The corresponding AUROC values remain those reported for the same testbeds in the paper, since AUROC is threshold-invariant.

Table 8: Comparison against MAGE Longformer (Testbeds 5 & 6)

Model	**TB5 refined (AvgRec / AUROC)**	**TB6 refined (AvgRec / AUROC)**
Longformer (MAGE)	87.62% / 0.95	81.78% / 0.93
RoBERTa-base - MiniLM, CE + Triplet	84.81% / 0.9379	77.87% / 0.8492
DeBERTa-base - DistilBERT, CE + KD + Triplet	85.63% / 0.9262	73.23% / 0.8815

Table 9: Comparison against MAGE Longformer (Testbeds 7 & 8)

Model	**TB7 refined (AvgRec / AUROC)**	**TB8 refined (AvgRec / AUROC)**
Longformer (MAGE)	86.54% / 0.94	62.92% / 0.75
RoBERTa-base - MiniLM, CE + Triplet	85.81% / 0.9402	64.38% / 0.7349
DeBERTa-base - DistilBERT, CE + KD + Triplet	85.64% / 0.9380	71.60% / 0.8057

The refined comparison yields a clear and balanced picture. On the auxiliary OOD calibration testbeds TB5 and TB6, Longformer remains stronger than both selected distilled models. On TB5, Longformer reaches 87.62% AvgRec, compared with 84.81% for the best RoBERTa-based detector and 85.63% for the best DeBERTa-based detector. On TB6, the gap becomes larger, with Longformer at 81.78%, the RoBERTa-based detector at 77.87%, and the DeBERTa-based detector at 73.23%. These results indicate that, after adjustment, Longformer still provides the strongest overall calibration behavior on the two validation-oriented OOD settings.

On TB7 after refinement, the comparison is much closer. Longformer reaches 86.54% AvgRec, while RoBERTa - MiniLM, CE + Triplet reaches 85.81% and DeBERTa-base - DistilBERT, CE + KD + Triplet reaches 85.64%. Thus, Longformer remains slightly ahead on the unseen-domain/unseen-model testbed, but only by a narrow margin of 0.73 and 0.90 percentage points, respectively. This shows that the proposed distilled detectors are highly competitive with the MAGE benchmark under the main OOD evaluation condition.

The most favorable comparison appears on TB8 after refinement. On the paraphrasing-attack testbed, Longformer drops to 62.92% AvgRec, whereas the RoBERTa-based distilled detector reaches 64.38%, and the DeBERTa-based distilled detector reaches 71.60%. In particular, DeBERTa-base - DistilBERT with CE + KD + Triplet clearly surpasses the Longformer benchmark under the hardest robustness condition in MAGE. This is especially important because TB8 is the setting that most directly tests resilience to paraphrase-induced distribution shift.

Overall, the refined comparison shows that the proposed distilled detectors do not uniformly replace Longformer, since Longformer remains stronger on Testbed 5, Testbed 6, and slightly on Testbed 7. However, the proposed approach becomes particularly attractive on the most difficult downstream setting, namely TB8, where the best DeBERTa-based distilled detector clearly outperforms the Longformer benchmark. The most accurate conclusion is therefore that the proposed detectors are competitive with Longformer overall and clearly stronger under paraphrase robustness when the DeBERTa-based distillation setting is used.

4.5 Qualitative Analysis: Proposed Distilled Detector vs Longformer Baseline

Quantitative metrics such as AvgRec and AUROC summarize performance over entire test sets, but they do not reveal which kinds of inputs are responsible for failures under distribution shift. To complement the aggregate evaluation, this subchapter presents representative qualitative examples in which the proposed

distilled detector is correct while the MAGE Longformer baseline is incorrect, and vice versa.

The qualitative analysis is centered on the best-performing distilled configuration identified in this thesis, namely DeBERTa-base - DistilBERT trained with cross-entropy, logit knowledge distillation, and teacher-guided triplet distillation. In the remainder of this subchapter, this model is referred to as the distilled student. Each detector outputs a human-class score $p(\text{human} \mid x)$, defined as the softmax probability assigned to label 1. Final decisions are made using the calibrated decision boundaries employed in the final evaluation protocol. Accordingly, the examples below should be interpreted under the final evaluation setting rather than under the default threshold of 0.5 so both the distilled student and Longformer baseline are evaluated with their respective refined thresholds. Labels follow the dataset convention Human = 1 and Machine = 0 and the examples shown are exclusive wins, meaning that exactly one of the two detectors predicts the correct label.

Table 10: Qualitative examples on Testbed 5

Correct model	**Fold**	**Ground truth**	**Example (first sentence)**
Longformer	unseen_model__7B	Machine	*Crowd counting from unconstrained scene images is a crucial task in many real-world applications like urban surveillance and management, but it is greatly challenged by the camera's perspective that causes severe scale variance among different individuals in the scene and complicated occlusion.*
Longformer	unseen_model_flan_t5_small	Machine	*Jimmy went to the local diner to get some breakfast.*
Distilled Student	unseen_model__7B	Human	*Andorra has become a creative hub filled with social media influencers from across Europe.*

Correct model	Fold	Ground truth	Example (first sentence)
Distilled Student	unseen_model_gpt-3.5-turbo	Human	*Media playback is not supported on this device Britain beat the USA and Australia at the venue last year, on their way to winning the title for the first time in 79 years.*

Table 11: Qualitative examples on Testbed 6

Correct model	Fold	Ground truth	Example (first sentence)
Longformer	unseen_domain_wp	Human	*March 9th, 1859.*
Longformer	unseen_domain_hswag	Machine	*is spotting her to ensure proper form and safety.*
Distilled Student	unseen_domain_xsum	Human	*Mark Mason, 48, of Rhyl, Denbighshire, was stabbed to death in the car park of the town's Home Bargains on 27 October.*
Distilled Student	unseen_domain_squad	Machine	*The pound-force has a metric counterpart, less commonly used than the newton: the kilogram-force (kgf) (sometimes kilopond), is the force exerted by standard gravity on one kilogram of mass.*

Table 12: Qualitative examples on Testbed 7

Correct model	**Source**	**Ground truth**	**Example (first sentence)**
Longformer	dialog-like human sample	Human	*Person1: Can I be of any service to you?*
Longformer	imdb-like machine sample	Machine	*This soap is worse than bad: it's poisonous.*
Distilled Student	cnn_human	Human	*The BBC faced angry criticism for giving an Election platform to 'mini Russell Brand' Gareth Shoulder (above) who mocked David Cameron over his disabled son.*
Distilled Student	imdb_human	Human	*I have to admit, that out of the many many thriller movies i have seen, this has to be one of the worst.*

Table 13: Qualitative examples on Testbed 8

Correct model	**Source**	**Ground truth**	**Example (first sentence)**
Longformer	dialogsum_human	Human	*Person1: Do you know that I'm checking out in about 30 minutes?*
Longformer	pubmed_gpt4_para	Machine	*As per a research paper released in the Journal of Pediatric Urology during 2011, it was observed that instructing medical practitioners about the Society for Fetal Urology (SFU) rating methodology for pediatric hydronephrosis by utilizing Computer Enhanced Visual Learning (CEVL) brought about significant improvement in their comprehension and knowledge about the grading system.*

Correct model	Source	Ground truth	Example (first sentence)
Distilled Student	dialogsum_para	Machine	*Person1: Peter, would you like to come and have tea now?*
Distilled Student	cnn_human	Human	*The BBC faced angry criticism for giving an Election platform to 'mini Russell Brand' Gareth Shoulder (above) who mocked David Cameron over his disabled son.*

Across the four testbeds, qualitative evidence points to complementary failure modes rather than a uniformly dominant detector. Several distilled-student wins occur on human texts from shifted domains or settings where the baseline appears overly conservative and rejects unfamiliar human writing. Conversely, Longformer retains isolated advantages on highly fluent machine generations and on some paraphrased technical passages that closely resemble authentic human text. Overall, the examples are fully consistent with the quantitative findings reported earlier in this chapter: the proposed distilled student benefits substantially from threshold refinement and achieves a strong balance between robustness and computational efficiency, while Longformer remains competitive on a subset of especially human-like machine outputs.

It is also worth noting that the best RoBERTa-based distilled model showed a different profile. The strongest RoBERTa-based student was RoBERTa - MiniLM with student CE + teacher-triplet distillation, which achieved 79.50% AvgRec on Testbed 7 under the default threshold and 85.81% after refinement, indicating that it was also highly competitive in the standard OOD setting. However, its Testbed 8 performance was clearly lower, dropping from 66.34% AvgRec before refinement to 64.38% after refinement. By contrast, the DeBERTa-based distilled student achieved 72.77% on Testbed 8 before refinement and 71.60% after refinement. For this reason, the qualitative discussion in the main text is centered on the DeBERTa-based student, which is the strongest final distilled model and the most relevant one for the thesis conclusions.

4.6 Runtime Comparison

OOD robustness is important, but practical deployment also requires efficient inference. For this reason, inference latency was measured for the strongest distilled detector and compared with the MAGE Longformer baseline on Testbeds 5-8. Runtime is reported in milliseconds per example, using the same timing configuration as in the experiments, including GPU warm-up followed by timed inference.

In our benchmarks, the student model is evaluated with inputs truncated to a maximum length of 512 tokens, while the Longformer baseline is evaluated under a long-context configuration with a maximum length of 4096 tokens. Because inference cost grows with sequence length (and depends on batch size), we report these runtimes as measured and interpret the resulting speedups considering the long context setting used for Longformer.

Table 14: Inference latency on OOD Testbeds

Testbed	**Distilled student runtime (ms/example)**	**Longformer runtime (ms/example)**	**Speedup (Longformer / Distilled Student)**
Testbed 5	3.655	27.000	7.39
Testbed 6	3.605	25.992	7.21
Testbed 7	2.023	17.244	8.52
Testbed 8	2.104	17.485	8.31

The distilled student consistently provides substantially lower latency across all testbeds, supporting higher-throughput screening and lower-cost deployment. The observed speedup ranges from 7.21x on Testbed 6 to 8.52x on Testbed 7, showing that the proposed detector is substantially more efficient than Longformer under the evaluation conditions used in this thesis.

Part of this speed advantage is due to the fact that Longformer is evaluated under a long-context configuration, which increases the computational burden per example. This should be interpreted as part of the practical trade-off between the two approaches. Longformer offers long-context modeling capacity at

substantially higher computational cost, whereas the distilled student provides much lower latency and higher throughput under the standard setting adopted here.

From a deployment perspective, this result is particularly important because the final detector is only the student model. The larger DeBERTa-base teacher is required during training, but not during inference. The proposed approach therefore combines two practical advantages: strong robustness under difficult OOD conditions and substantially lower inference cost than the Longformer baseline.

These results show that the selected distilled student is not only effective, but also efficient. This makes it a strong candidate for large-scale screening and other latency-sensitive applications where both robustness and computational cost matter.

5 Conclusions

5.1 Summary of the Research

This thesis investigated Out-of-Distribution Detection of machine generated text under the MAGE benchmark, with emphasis on the most practically relevant OOD settings such as Unseen Models, Unseen Domains, combined distribution shifts, and Paraphrasing attacks. The main objective was to examine whether compact student detectors can remain robust under these conditions and whether their performance can be improved through knowledge distillation and representation-shaping objectives such as triplet loss and supervised contrastive loss. To keep the evaluation computationally feasible, the experimental workflow was staged: Testbed 7 was used for the initial selection of the best distilled configuration, Testbed 6 for threshold refinement, Testbed 5 for transfer evaluation before and after calibration and Testbed 8 for paraphrasing robustness before and after applying the refined threshold.

Overall, the results show that compact detectors can achieve strong OOD performance and in the strongest configuration, remain highly competitive with the MAGE Longformer benchmark. At the same time, neither distillation nor any single auxiliary loss is universally best in every setting. Instead, performance depends on the teacher-student pairing, the loss design, and the specific type of distribution shift.

5.2 Effect of Knowledge Distillation

Knowledge distillation proved beneficial in the most challenging settings of this study, although its effect was not uniform across all experiments. In the RoBERTa-based setting, the strongest default results on Testbeds 7 and 8 were achieved by student-only models, while the distilled variants remained competitive without clearly surpassing them. On Testbed 7, the best RoBERTa-based student-only result reached 80.59% AvgRec, while the strongest distilled RoBERTa-based model reached 79.50%. On Testbed 8, the best RoBERTa-based student-only result reached 71.59% AvgRec, compared with 70.37% for the strongest distilled RoBERTa-based variant. A slightly different picture emerges

in AUROC, where the strongest RoBERTa student-only model on Testbed 7 achieved 0.9086, while the selected distilled RoBERTa model reached 0.9402, indicating that the distilled model preserved stronger ranking quality even when its default-threshold AvgRec was slightly lower.

However, in the DeBERTa-based setting, the hybrid distilled model DeBERTa-base - DistilBERT trained with cross-entropy, knowledge distillation, and teacher-guided triplet loss emerged as the strongest distilled detector and produced the best Testbed 8 result, making it the most successful final model in the study. On Testbed 8, this model reached 72.77% AvgRec, compared with 69.52% for the corresponding student-only DistilBERT model and 65.42% for the teacher-only cross-entropy baseline. Its AUROC on Testbed 8 was 0.8057, compared with 0.7788 for the student-only DistilBERT triplet model and 0.7082 for the teacher-only cross-entropy model, which strengthens the conclusion that distillation was most beneficial in the paraphrasing setting.

The teacher-only, student-only, and distilled comparisons on Testbeds 7 and 8 show that the effect of distillation depends on the setting. Some student models outperformed their teachers, especially on Testbed 7, and the best DeBERTa-based distilled student outperformed both teacher-only and student-only baselines on Testbed 8. This suggests that a larger teacher is not automatically better calibrated or more robust under distribution shift. On Testbed 7 in the same DeBERTa-based setting, the best teacher-only result reached 76.08% AvgRec, the student-only DistilBERT model with triplet loss reached 80.15%, and the best distilled model reached 79.65%, indicating that distillation helped selectively rather than uniformly.

5.3 Effect of the Training Objectives

The analysis of the training objectives revealed a clear and practically useful pattern, even though no single loss function was best in every setting. Cross-entropy remained the essential backbone objective in all strong configurations. Beyond that, triplet-based representation shaping was the most consistently useful auxiliary mechanism, especially for student-only and hybrid-distilled models. The strongest distilled configuration in the thesis was the hybrid model trained with cross-entropy, knowledge distillation, and teacher-guided triplet

loss, while strong student-only results were also achieved with cross-entropy and triplet loss and, in one RoBERTa-based Testbed 7 setting, with cross-entropy and supervised contrastive loss.

Teacher-only, student-only and distilled comparisons also clarify the role of triplet loss. For teacher-only models, triplet loss was not consistently beneficial across both Testbed 7 and Testbed 8. For student-only models, its contribution was more clearly positive, especially in the DeBERTa-based DistilBERT setting, where performance on Testbed 7 increased from 72.80% AvgRec with cross-entropy alone to 80.15% when triplet loss was added. AUROC also improved from 0.8918 to 0.9470 in that comparison, showing that the gain was not only a threshold effect but also reflected better separation between human written and machine generated text. For distilled models, the strongest evidence again comes from the DeBERTa-based hybrid detector, where KD and teacher-guided triplet produced the best paraphrasing robustness in the study.

5.4 Differences across the OOD Testbeds

The results varied substantially across the MAGE Testbeds, confirming that OOD detection performance depends strongly on the type of distribution shift. Testbed 5 and Testbed 7 were the most favorable settings for the selected final models, especially after calibration. Testbed 6 was more challenging at the default operating point but improved strongly after threshold refinement, while Testbed 8 remained the most distinctive and demanding setting because paraphrasing weakens many of the surface-level cues used by standard detectors.

A second important observation is that the two selected detector families showed different robustness profiles. The RoBERTa-based distilled detector became highly competitive on the standard OOD setting after refinement, but its Testbed 8 performance remained clearly lower. By contrast, the DeBERTa-based distilled detector combined near Longformer performance on Testbed 7 with clearly stronger robustness on Testbed 8, making it the strongest final detector in the thesis. After refinement, the DeBERTa-based detector reached 85.64% AvgRec on Testbed 7 and 71.60% on Testbed 8. The corresponding AUROC values remained 0.9380 on Testbed 7 and 0.8057 on Testbed 8, confirming that

paraphrasing was not only harder at the chosen threshold but also harder in terms of overall class separability.

5.5 Effect of Threshold Refinement

Threshold refinement was one of the most useful findings of the study. For both selected distilled models, calibration on Testbed 6 produced substantial gains on Testbed 7 and clear improvements on Testbed 6, while the effect on Testbed 5 was small but non-negative. This shows that decision boundary adjustment is highly effective for the standard OOD settings represented by unseen domains and combined unseen domain/unseen model transfer. For the selected RoBERTa-based distilled detector, AvgRec on Testbed 7 increased from 79.50% to 85.81% after refinement. For the selected DeBERTa-based distilled detector, the corresponding improvement was from 79.65% to 85.64%. Importantly, AUROC did not change in these comparisons for Testbed 7 and remained 0.9402 for the selected RoBERTa-based model and 0.9380 for the selected DeBERTa-based model. This confirms that threshold refinement improved the operating point rather than the underlying ranking quality of the detector.

At the same time, the refined threshold did not transfer equally well to Testbed 8. In both selected models, paraphrasing robustness decreased slightly after threshold adjustment, which suggests that Testbed 8 has a different optimal operating point from the standard OOD testbeds. Even so, the DeBERTa-based distilled detector remained strong on Testbed 8 after refinement and still outperformed the Longformer benchmark there. On Testbed 8, the DeBERTa-based model decreased only slightly, from 72.77% to 71.60% AvgRec, while the RoBERTa-based model dropped from 66.34% to 64.38%. Again, AUROC stayed unchanged at 0.8057 for the DeBERTa-based model and 0.7349 for the selected RoBERTa-based model, reinforcing the interpretation that the drop came from threshold transfer rather than a change in discriminative ability.

5.6 Competitiveness against the Longformer Benchmark

The comparison with the MAGE Longformer baseline shows a mixed but encouraging pattern. Longformer remained stronger on Testbeds 5 and 6 after

refinement, and it also kept a small advantage on Testbed 7. However, the gap on Testbed 7 was very small: the proposed distilled detectors reached 85.81% and 85.64% AvgRec, compared with 86.54% for Longformer. This shows that compact distilled models can approach the benchmark very closely under the main OOD evaluation condition.

The strongest result appears on Testbed 8. On the paraphrasing-attack testbed, the RoBERTa-based distilled detector slightly exceeded Longformer, while the DeBERTa-based distilled detector reached 71.60% AvgRec, clearly above the 62.92% reported for Longformer. This advantage is also reflected in AUROC, where the DeBERTa-based distilled model reached 0.8057, compared with 0.75 for Longformer, showing that its superiority on TB8 is not limited to threshold selection. This matters because Testbed 8 is the most challenging robustness setting in MAGE and is closely related to adversarial masking of machine generated text.

5.7 Efficiency

An important practical contribution of the thesis is efficiency. In the runtime measurements, the strongest distilled student required 3.655 ms per example on Testbed 5, 3.605ms on Testbed 6, 2.023ms on Testbed 7, and 2.104ms on Testbed 8, whereas the Longformer baseline required 27.000 ms, 25.992ms, 17.244ms, and 17.485ms, respectively. This corresponds to speedups ranging from 7.21x to 8.52x, depending on the testbed.

This advantage should be interpreted together with the input length setting. In this thesis, the proposed student models were evaluated with a maximum input length of 512 tokens, while the Longformer baseline used a 4096 token context window. Longformer therefore benefits from much longer context, but at substantially higher computational cost.

5.8 Main Conclusions

The findings can be summarized in four main conclusions. First, compact student detectors can remain highly competitive under difficult OOD conditions. Second, knowledge distillation helps selectively rather than universally, with its strongest contribution appearing in the DeBERTa-based hybrid model. Third, triplet-based

representation shaping is the most consistently useful auxiliary mechanism across the strongest student and distilled settings. Fourth, the final distilled detector is not only robust but also efficient, approaching Longformer on Testbeds 7, surpassing it on Testbed 8, and doing so with much lower inference cost.

5.9 Limitations

The conclusions of this study are shaped by several practical constraints. First, the fixed maximum input length of 512 tokens may limit performance in settings where longer range context cues are important. Second, the threshold refinement strategy relied on a single boundary adjustment derived from unseen domain calibration and although this improved performance on Testbed 7, it did not transfer equally well to the paraphrasing setting. Third, the most computationally expensive testbeds could not be explored with the same level of experimental exhaustiveness as the lighter settings, which made a staged workflow necessary for model selection and calibration.

5.10 Future work

A logical next step is to explore stronger teacher models and alternative student architectures beyond the combinations evaluated here. Promising teacher candidates include ModernBERT, RoBERTa-large, and larger DeBERTa variants, which may provide stronger supervisory signals during distillation. On the student side, it would be useful to evaluate other compact encoders such as TinyBERT, MobileBERT and ELECTRA-small. It would also be worth examining moderately longer student context windows, since the current 512 token setting may limit performance in cases where longer range cues are important. In addition, future work should investigate testbed-specific threshold calibration more systematically, since the current results suggest that a threshold tuned for standard OOD transfer is not optimal for paraphrasing attacks.

References

Bao, G., Zhao, Y., Teng, Z., Yang, L., & Zhang, Y. (2023). Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. *arXiv preprint arXiv:2310.05130*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J. M., Xu, H., Ippolito, D., & Callison-Burch, C. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12463-12492). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.674

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111-116). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3019

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... & Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., ... & Goldstein, T. (2024). Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. *arXiv preprint arXiv:2401.12070*.

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa:* Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). *MGTBench: Benchmarking machine-generated text detection* (arXiv:2303.14822). arXiv.

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*(pp. 1808-1822). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.164

Iyer, H., Seo, S., Diduch, L., Peterson, K., Awad, G., & Lee, Y. (2025). *2024 NIST GenAI (Pilot Study): Text-to-text evaluation overview and results* (NIST AI 700-1). National Institute of Standards and Technology.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4163-4174). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.372

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 17061-17084). PMLR.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, *33*, 18661-18673.

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems, 36, 27469-27500*

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, Y., Awad, G., Butt, A., Diduch, L., Peterson, K., Seo, S., Soboroff, I., & Iyer, H. (2024). *2024 NIST Generative AI (GenAI): Evaluation plan for text-to-text (T2T) discriminators*. National Institute of Standards and Technology.

Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L., Yang, L., Shi, S., & Zhang, Y. (2024). MAGE: Machine-generated text detection in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 36-53). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.3

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., Srba, I., Le, T., Lee, D., Simko, J., & Bielikova, M. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9960-9987). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.616

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 24950-24962). PMLR.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected?. *arXiv preprint arXiv:2303.11156*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv.

Sarvazyan, A. M., González, J. Á., Franco-Salvador, M., Rangel, F., Chulvi, B., & Rosso, P. (2023). Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815-823).

Schuster, T., Schuster, S., Shah, D., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. In *Proceedings of the 2nd Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda* (pp. 75-80). Association for Computational Linguistics.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology, 60*(3), 538-556. https://doi.org/10.1002/asi.21001

Su, J., Zhuo, T., Wang, D., & Nakov, P. (2023). DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 12395-12412).

Su, Z., Wu, X., Zhou, W., Ma, G., & Hu, S. (2023). HC3 Plus: A semantic-invariant human ChatGPT comparison corpus. *arXiv preprint arXiv:2309.02731*.

Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2001-2016). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.172

UNESCO. (2023). *Guidance for generative AI in education and research*. United Nations Educational, Scientific and Cultural Organization.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MinLLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, *33*, 5776-5788.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Mohammed Afzal, O., Mahmoud, T., Puccetti, G., & Arnold, T. (2024). SemEval-2024 Task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)* (pp. 2057-2079). Association for Computational Linguistics.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., ... & Ronneberger, O. (2020). Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. F., & Chao, L. S. (2023). *A survey on LLM-generated text detection: Necessity, methods, and future directions* (arXiv:2310.14724). arXiv.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32-35. https://doi.org/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3

Yuan, L., Tay, F. E., Li, G., Wang, T., & Feng, J. (2020). Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3903-3911).

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *Advances in neural information processing systems*, *32*.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2021). *Domain generalization: A survey* (arXiv:2103.02503). arXiv.