Harnessing Language Models to Analyze Android App Permission Fidelity

Yunik Tamrakar*, Ritwik Banerjee†, Ethan Myers*, Lorenzo De Carli‡, Indrakshi Ray*
*Colorado State University
Fort Collins, Colorado, United States of America
Email: yunik.tamrakar@colostate.edu; etham@rams.colostate.edu; indrakshi.ray@colostate.edu
†Stony Brook University
Stony Brook, New York, United States of America
Email: rbanerjee@cs.stonybrook.edu
‡University of Calgary
Calgary, Alberta, Canada
Email: lorenzo.decarli@ucalgary.ca

***Abstract*—Android's vast app ecosystem (over 2 million apps) poses significant privacy risks, as current methods for inferring permissions from descriptions – keyword matching, traditional natural language processing (NLP), and recurrent neural networks (RNNs) – struggle with accurate inference due to imprecise, ambiguous, or incomplete natural language descriptions. This gap undermines regulatory transparency and user trust, necessitating tools that reconcile stated functionality with actual data practices. We demonstrate that large language models like GPT-4o, applied in a zero-shot inference setting, leverage contextual reasoning to infer permissions competitively, while fine-tuned encoders (BERT, BART) surpass state-of-the-art performance when trained on minimally annotated datasets augmented with paraphrases, achieving 50–70% gains in weighted and macro F_1 scores. By enabling precise permission auditing with reduced annotation costs, our work advances scalable, adaptable solutions for privacy compliance across resource-constrained and high-stakes environments.**

***Index Terms*—Mobile applications, privacy, regulation, natural language processing, machine learning, language models, classifier design and evaluation**

I. Introduction

Modern mobile applications routinely collect, process, and share vast amounts of user data – including highly sensitive personally identifiable information (PII) such as their name, location, biometric identifiers, device identifiers, and network addresses [1], [2]. The aggregation of such data poses significant privacy risks, as even ostensibly anonymized datasets can often be re-identified through linkage attacks [3], [4]. This risk escalates in domains like mobile health (mHealth), where applications frequently handle additional sensitive PII categories including medical histories, real-time physiological metrics (*e.g.*, heart rate, blood glucose), and demographic attributes [5], [6]. Under regulatory frameworks (like GDPR and, for some health-related information, HIPAA), this necessitates strict adherence to the principle of least privilege—a requirement complicated by the frequent mismatch between declared app functionalities and requested permissions [7].

Mobile app privacy documentation—including descriptions and policies—is intended to transparently disclose requested permissions, particularly those granting access to sensitive personal data (e.g., camera, microphone, or biometric data). When granted, such permissions enable direct access to sensitive PII [8], posing systemic privacy risks amplified by three unresolved challenges:

1. A **semantic gap in permission inference** arises when the permissions stated or implied in an app's description differ from those declared in its Android manifest [9]–[12]. We focus on the 'described' side of this gap: predicting permissions that are explicitly mentioned (e.g., "send SMS" $\rightarrow$ SMS) or suggested by functionality (e.g., "map your route" $\rightarrow$ LOCATION). While some cases are direct, many require deeper understanding of natural language semantics.
2. The labor-intensive nature of manual annotation for permission-description pairs creates a **dearth of representative training data**, with existing datasets being both limited in scale and skewed toward frequently downloaded categories—like messaging or entertainment—leaving critical domains like mHealth understudied. Studies show that the fraction of users who understand permission implications can be as low as 3% [13], making crowdsourced labeling unreliable. As a result, state-of-the-art models generalize poorly, particularly for domains like mHealth.
3. Since Android 6.0, permissions are requested at runtime instead of installation [14], but this shift has introduced **permission fatigue**: users are frequently prompted and tend to approve requests reflexively [13], [15]. This phenomenon—termed *consent theater* [16]—undermines informed user control by desensitizing users to privacy risks.

Users often rely on app descriptions—typically the sole source on platforms like Google Play—to assess permission requirements. While these descriptions should transparently justify requested permissions under FTC guidelines [17], their brevity and marketing-driven framing frequently obscure permission risks [18]. Ensuring description-to-permission fidelity thus demands language understanding capable of identifying functional claims (*e.g.* LOCATION for "weather alerts") while inferring potential overreach latent in ambiguous or imprecise

phrasing. Current approaches, however, remain fundamentally limited in capturing these subtle linguistic distinctions.

Motivated by these challenges and the urgent need for scalable solutions, we conduct a systematic evaluation of language models for automated permission inference from privacy documentation, leveraging their ability to discern subtle context-sensitive nuances of natural language [19]. Specifically, we present three key advances, demonstrating (in Sec. IV):

(1) modern large language models (LLMs) achieve highly accurate permission inference in zero-shot settings—up to 18% F_1 score improvements—over fine-tuned encoder models, establishing a new baseline in permission-description fidelity assessment;
(2) upon incorporating paraphrase-based data augmentation, fine-tuned encoder models substantially surpass current state-of-the-art, achieving gains of 5 - 25% in area under the precision-recall (PR) curve; and
(3) our augmentation strategy is useful in limited-data scenarios, achieving 15% and 38% increases in area under the receiver-operating characteristic (ROC) and PR curves, respectively, and 55 - 82% in F_1 scores across permission categories when compared against fine-tuning using only 250 non-augmented samples.

Our contributions are rigorously validated through experiments across three distinct LLM architectures and two benchmark datasets, advancing mobile privacy concerns by providing (i) an instantly deployable zero-shot solution, and (ii) a data-efficient supervised learning paradigm suitable for models that require high accuracy (e.g., for regulatory compliance checks).

II. Related Work

Early work on description-to-permission consistency framed the problem with traditional NLP techniques. The WHYPER system used rule-based text matching to identify whether an app's description justifies each requested permission [10]. This framework, applied to three common permissions, achieved over 80% precision and recall, demonstrating the promise of sentence-level NLP for this task. Building on this idea, the AutoCog system learned a mapping between app descriptions and permission lists [9]. AutoCog employed hand-crafted features and learned a classifier to predict whether a description supports each permission. Its evaluation on 11 permissions demonstrated an average precision of 92.6% and recall of 92.0%. More recently, neural-network approaches have emerged. Notably, AC-Net proposed an end-to-end neural framework that assigns one or more permissions to individual description sentences [11], and reported large gains (24.5% higher accuracy) over prior methods on a dataset of 1,415 apps, showing that learned encoders can outperform hand-crafted features when sufficient labeled data is available.

While these prior studies used supervised learning on fixed datasets, our work is motivated by recent advances in large language models (LLMs) for privacy text analysis. Several recent works apply LLMs to privacy and permission problems, though not directly to permission-description fidelity of apps. For instance, Rodriguez et al. [20] used ChatGPT and Llama 2 to analyze privacy policies, achieving 93% F_1 on detecting data practice disclosures in such policies. Similarly, Tang et al. [21] devised PolicyGPT based on GPT-4 [22] to show that LLM-based zero-shot approaches can surpass traditional classifiers on identifying privacy policies in legal texts. Oishwee et al. [23] investigated ChatGPT's ability to answer Android permission questions, analyzing 1,008 StackOverflow threads and concluding that LLMs can effectively assist developers with permission-related issues. Another related line of work by Alecakir et al. [24] examines LLMs for privacy/security more broadly. However, no prior work has applied LLM prompting to the app description–permission fidelity problem. XLMR4MD, a Vietnamese-language dataset and LLM-based framework for permission-description consistency, represents the only prior work in this task using LLMs [25], albeit without exploring data augmentation for low-resource settings.

In contrast, our work leverages zero-shot LLM prompting to infer permissions from descriptions, outperforming task-specific encoders by up to 18% F_1. We further introduce a paraphrase-based augmentation strategy that boosts encoder performance by 5-25% in PR AUC, with especially significant gains in low-resource scenarios: using only 250 training examples, we observe improvements of 15-38% in ROC/PR AUC, and 55-82% in F_1. Collectively, these contributions combine modern NLP with novel training techniques to advance beyond systems like WHYPER, AutoCog, and AC-Net, while addressing previously unexplored, practical scenarios requiring zero-shot inference or limited training data.

III. Problem Formulation and Datasets

We frame permission inference as a multi-label binary classification task where, given a sentence from an app description, the model predicts a binary vector representing the presence/absence of 11 Android permissions (see Table I). These permissions, derived from Android's standardized taxonomy, correspond to the 11 categories used by Feng et al. [11], derived from empirical user-concerns [26].

We employ two datasets. The first, used by Feng et al. [11] to assess the consistency of descriptions and permissions in Android apps (AC-Net), comprises 24,726 sentences across 1,415 Android apps from the Google Play Store. Unlike other datasets used to study consistency issues of app descriptions [10], [24], [27], the AC-Net corpus provides multiple permission-labels per sentence, making it the sole dataset available for multi-label classification tasks. With its 11 permission categories, it is also among the most comprehensive in terms of wide coverage. However, 80% of the corpus' sentences do not pertain to any permission category. While this provides adequate negative samples for each category, it leads to significant class imbalance and inadequate training data for several categories (which we overcome using our data augmentation strategy, described in §IV-D). We use the AC-Net dataset for zero-shot inference as well as for training and evaluating the fine-tuned models. Our test-set consists of 3,600 random samples (15%). The remaining data is split 80%/20% into training and validation sets.

TABLE I: The 11 Android permission categories (and the corresponding Android permissions) used in our multi-label classification task, shown here with exemplar input and output.

"The app will enhance your productivity based on your current location."	
Category (Permission)	
STORAGE (WRITE_EXTERNAL_STORAGE, GET_ACCOUNTS)	✗
CONTACTS (READ_CONTACTS, WRITE_CONTACTS)	✗
LOCATION (ACCESS_FINE_LOCATION, ACCESS_COARSE_LOCATION)	✓
CAMERA (CAMERA)	✗
MICROPHONE (RECORD_AUDIO)	✗
SMS (READ_SMS, SEND_SMS)	✗
CALL_LOG (READ_CALL_LOGS)	✗
PHONE (CALL_PHONE)	✗
CALENDAR (READ_CALENDAR)	✓
SETTINGS (WRITE_SETTINGS)	✗
TASKS (GET_TASKS, KILL_BACKGROUND_PROCESS)	✗

TABLE II: Permission inference across 10 runs: F_1 scores (mean μ and median $\tilde{\mu}$) for zero-shot (GPT-4o, GPT-4o-mini) and fine-tuned (BERT, BART-Large-MNLI) models.

	GPT-4o		**GPT-4o-mini**		**BERT**		**BART**	
Category	μ	$\tilde{\mu}$	μ	$\tilde{\mu}$	μ	$\tilde{\mu}$	μ	$\tilde{\mu}$
STORAGE	**0.61**	**0.62**	0.37	0.41	0.60	0.60	0.57	0.56
CONTACTS	0.62	0.61	0.49	0.49	**0.68**	**0.69**	0.66	0.65
LOCATION	**0.88**	**0.90**	0.75	0.77	0.70	0.69	0.71	0.72
CAMERA	**0.84**	**0.84**	0.56	0.57	0.70	0.71	0.72	0.72
MICROPHONE	**0.61**	**0.59**	0.34	0.33	0.48	0.47	0.47	0.48
SMS	**0.80**	**0.82**	0.65	0.66	0.73	0.74	0.74	0.72
CALL_LOG	0.57	0.58	0.45	0.47	0.45	0.41	**0.59**	**0.62**
PHONE	**0.69**	**0.66**	0.42	0.40	0.50	0.54	0.55	0.57
CALENDAR	**0.82**	**0.82**	0.44	0.45	0.76	0.76	0.77	0.78
SETTINGS	0.22	0.27	0.04	0.07	0.44	0.51	**0.49**	**0.51**
TASKS	0.37	0.39	0.05	0.09	0.35	0.35	**0.44**	**0.44**

Additionally, we introduce a small, curated dataset of 250 permission sentences (P_{250}). This corpus is independently annotated by two domain experts (one acting as an adjudicator to resolve conflicts), with inter-annotator agreement measured by Cohen's Kappa $\kappa = 0.82$.[1] We use this collection to (i) assess the general applicability of our models across datasets; (ii) conduct experiments in low-resource settings; and (iii) study the impact of data augmentation.

IV. Experiments and Findings

Our first experiments investigate modern large language models (LLMs) for inferring Android permissions based on app descriptions. We use two state-of-the-art (SOTA) generative pre-trained transformer models—GPT-4o, the flagship model from OpenAI with approx. 1.8T parameters, and its lightweight variant GPT-4o-mini, with 8B parameters—and compare them against established encoder-based baselines: BERT-base (110M parameters) and BART-Large-MNLI (406M parameters) [30]–[32]. This spectrum of model sizes (110M to 1.8T parameters) across different generations of models, allows us to showcase performance-efficiency tradeoffs critical for real-world deployment, where computational constraints often favor smaller models despite potential penalties in accuracy. Moreover, it helps to put the field's methodological evolution in context, especially how traditional fine-tuning compares to the modern prompting paradigm.

A. Zero-shot inference with GPT

Zero-shot inference allows models to perform specialized tasks using only natural language instructions to guide their reasoning (e.g., *"From the given app description, determine the set of Android permissions that the app is likely to require.'*) without task-specific training. We evaluate GPT-4o and GPT-4o-mini with structured text prompts that comprise task guidelines and input sentences from the AC-Net dataset. For model hyperparameters that control response randomness, we retain the default values—temperature = 0.5, and nucleus sampling (`top_p`) = 1—to (i) reflect how practitioners often deploy "out of the box" models in security-related workflows (e.g., Carlini et al. [33]), and (ii) ensure reproducible results through fixed configurations associated with the model version [22].

B. Fine-tuning baselines

For comparison, we fine-tune BERT-base and BART-Large-MNLI on the AC-Net dataset using their default tokenizers, a learning rate of 2×10^{-5}, batch size 16, and 5 epochs. All experiments ran on NVIDIA A100 GPUs (40 GB memory), and repeated 10 times under identical conditions to assess the reliability and stability of results. We observe narrow agreement between mean and median F_1 scores, suggesting minimal variability. Results—of zero-shot inference and fine-tuning with smaller encoder models—are shown in Table II.

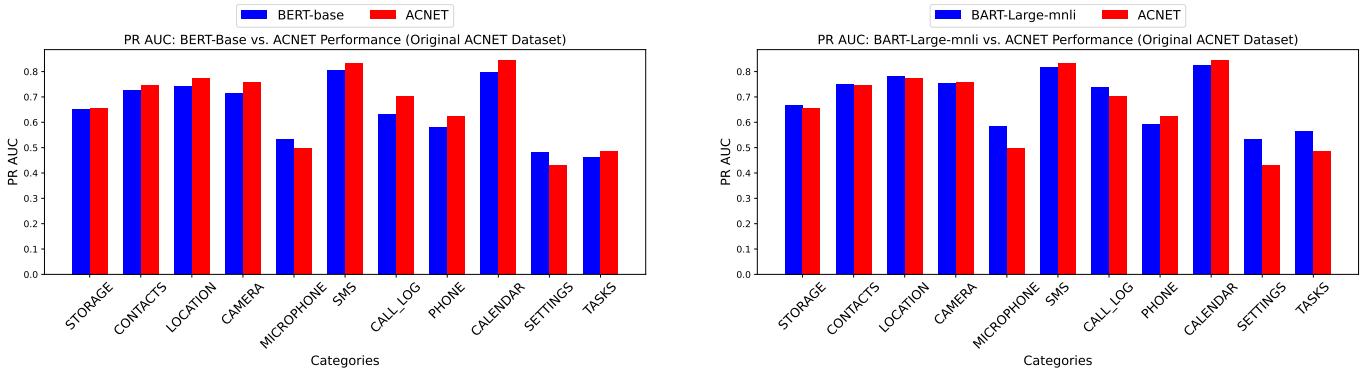
C. Key findings and implications

GPT-4o outperforms both fine-tuned baselines on 7 of 11 permission labels, with the largest gain observed for LOCATION (18% higher F_1 score, see Table II). GPT-4o-mini, however, lags significantly, achieving only 63% of GPT-4o's F_1 score. This significant difference underscores the tradeoff between model sizes and their zero-shot inference capability.

Both GPT variants struggle with SETTINGS and TASKS, with GPT-4o lagging behind the fine-tuned smaller BART model by 27% and 7%, respectively. A manual error analysis suggests that app descriptions tend to lack clear, explicit mentions of these permissions, often requiring implicit and/or more complex reasoning—challenges that even human annotators faced, and that the latest LLMs fail to consistently capture.

Clearly, the latest LLMs are largely successful in permission inference, thereby justifying the high compute cost of a model like GPT-4o. This is unsurprising, as these models are trained on a large number of publicly available natural language inference datasets. The challenges faced in SETTINGS and TASKS, coupled with the relative success of the fine-tuned smaller models, indicate that there are specific permission indicators in the language patterns of Android app descriptions, which are not found in general-purpose language inference. Thus, if there is limited data and/or computational resources, or the application is latency-sensitive, smaller models remain viable,

[1]This inter-annotator agreement reflects excellent reliability and near-perfect consensus, aligning with established benchmarks [28], [29].



(a) PR-AUC: BERT vs AC-Net

(b) PR-AUC: BART-Large-MNLI vs AC-Net

Fig. 1: Comparison vs. SOTA on the original (non-augmented) AC-Net dataset

provided the performance gaps can be mitigated. Next, we demonstrate paraphrase-based augmentation alleviating these concerns, surpassing SOTA results on the AC-Net dataset.

D. Data Augmentation

Drawing from the above analysis, we aim to increase the training data for the permission categories (collectively comprising only 20% of the AC-Net corpus). To preserve semantics, we augment data via paraphrasing with GPT-4o. For each input sentence in category i, GPT-4o is prompted to generate paraphrases so that each class reaches equal size.[2] The augmentation yields a dataset with a 55:45 split between permission and non-permission sentences. We then fine-tune models and observe the results.

Fine-tuning BERT on the augmented AC-NET dataset achieves SOTA performance, surpassing the prior AC-Net baseline in ROC-AUC across all permission categories. Notably, PR-AUC—a critical metric for imbalanced datasets—improves by 10-30% (Table IV). However, PR-AUC scores for SETTINGS and TASKS remain lower than other categories even after augmentation. Similar improvements over SOTA are replicated with other BERT variants and the BART architecture, confirming that the gains can be generalized.

Impact: Data augmentation yields statistically significant improvements ($p < 0.05$) in recall and F_1 scores across all models, while precision improvements are either insignificant or negligible. This aligns with security priorities, as higher recall minimizes missed permission misuse.

With paraphrases of existing data, the models learn robust linguistic representations, improving generalization to unseen inputs. To quantify this, we measure prediction entropy for the base BERT model trained on augmented versus non-augmented data, showing (Table III) that augmentation leads to lower entropy, indicating greater prediction stability.

Resource-efficient training: To evaluate the value of our augmentation strategy in limited-data settings, we fine-tune models on the small manually annotated dataset of 250 samples, dubbed P_{250} (Sec. III). After augmenting this dataset with paraphrases (50 per sample), training on the augmented corpus leads to substantial improvements when evaluated on the AC-Net dataset, as shown in Fig. 2. In particular, we observe: +15% ROC-AUC (0.91), +38% PR-AUC (0.71), +67% weighted F_1 (0.67), and +69% macro-F_1 (0.69). Notably, SMS and TASKS show no PR-AUC improvement, likely because baseline performance for these labels was already high prior to augmentation, thus suggesting diminishing returns for categories where models approach ceiling performance.

V. Discussion

Our experiments demonstrate that zero-shot inference with modern LLMs like GPT-4o establish a new baseline in permission-description fidelity assessment, achieving **up to 18% F_1 score improvements over fine-tuned smaller models** (like BERT and BART-Large-MNLI) across critical

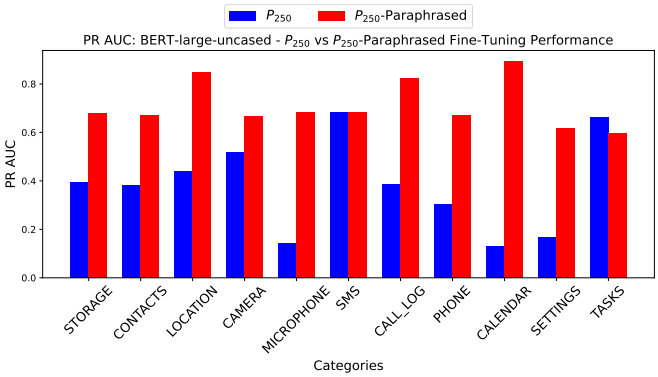


Fig. 2: PR-AUC: BERT fine-tuned on the annotated P_{250} dataset and its augmentation.

[2]The number of paraphrases generated for the i^{th} category is $L/n_i - 1$, where $L = \text{LCM}(n_1, n_2, \ldots, n_K)$, and n_j is the number of samples in the i_{th} category in the original corpus. GPT-4o is prompted with: "Please generate n paraphrases for the following sentence. Make sure each paraphrase is different from the input sentence and from each other. Try to make them as diverse as possible while retaining the original semantics. Return the output as a JSON object with the key "results", which maps to an array of paraphrases"

TABLE III: BERT – Average prediction entropy: Augmented vs. Non-augmented training

Label	Augmented	Non-Augmented
STORAGE	0.0659	0.1881
CONTACTS	0.0572	0.1713
LOCATION	0.0268	0.1181
CAMERA	0.0238	0.1088
MICROPHONE	0.0398	0.1231
SMS	0.0449	0.1321
CALL_LOG	0.0306	0.1388
PHONE	0.0378	0.1732
CALENDAR	0.0267	0.1221
SETTINGS	0.0344	0.1574
TASKS	0.0217	0.1484
Sum of Averages	0.4096	1.5814

TABLE IV: ROC-AUC and PR-AUC: BERT vs AC-Net (augmented)

Class	ROC-AUC			PR-AUC		
	BERT (aug)	BERT (unaug)	AC-Net	BERT (aug)	BERT (unaug)	AC-Net
STORAGE	**0.966**	0.942	0.942	**0.736**	0.652	0.655
CONTACTS	**0.990**	0.961	0.971	**0.839**	0.727	0.746
LOCATION	**0.989**	0.968	0.983	**0.826**	0.742	0.774
CAMERA	**0.995**	0.710	0.982	**0.827**	0.713	0.758
MICROPHONE	**0.990**	0.956	0.962	**0.740**	0.535	0.496
SMS	**0.997**	0.984	0.989	**0.902**	0.804	0.834
CALL_LOG	**0.999**	0.992	0.993	**0.943**	0.632	0.705
PHONE	**0.997**	0.977	0.991	**0.877**	0.581	0.624
CALENDAR	**0.998**	0.983	0.994	**0.908**	0.796	0.844
SETTINGS	**0.973**	0.945	0.951	**0.650**	0.481	0.432
TASKS	**0.986**	0.903	0.949	**0.717**	0.460	0.486
Mean	**0.990**	0.938	0.973	**0.815**	0.647	0.668

permission categories such as LOCATION and CAMERA. This underscores the viability of generative AI for permission inference *without task-specific training*.

While GPT-4o matches or exceeds fine-tuned baselines in most categories, its lightweight variant, GPT-4o-mini, lags in F_1 by 13% (LOCATION, CONTACTS) to 38% (CALENDAR). This gap reflects the inherent tradeoff between model scale and capability: GPT-4o's 1.8T parameters (compared to 8B of GPT-4o-mini) generally enable more nuanced semantic reasoning (e.g., inferring location access from phrases like "nearby weather alerts"). Crucially, GPT-4o's zero-shot performance rivals dedicated fine-tuned models trained on domain-specific data (AC-Net), validating its generalizability as a universal permission inference tool. This delivers our *first contribution*: modern LLMs obviate resource-intensive fine-tuning while setting new benchmarks in accuracy.

Our experiments demonstrate that paraphrase-based data augmentation enables fine-tuned LLMs to substantially surpass SOTA baselines, achieving **5–25% gains in area under the precision-recall (PR) curve** across permission categories. These improvements are also pronounced for high-risk labels like LOCATION (PR-AUC: +8%) and CAMERA (PR-AUC: +11%), where ambiguous language has been a challenge for traditional models. Crucially, without augmentation, fine-tuned LLMs only *match* SOTA performance, with no statistically significant advancements. This substantiates our *second contribution*: even the smaller models, when paired with strategic augmentation, push permission inference benchmarks. With paraphrases resolving class imbalance and data scarcity, models surpass the robustness achieved by tine-tuning.

We decisively answer a critical question in security-focused ML: *can language models achieve robust permission inference with minimal annotated data?* While training on only 250 manually annotated samples yields lackluster performance, augmenting this limited dataset with paraphrases drives dramatic gains in **ROC-AUC (+15%)**, **PR-AUC (+38%)**, and F_1 **(+55–82% across permission categories)**. These results highlight our *third contribution*: paraphrase-based augmentation enables language models to perform effectively in low-resource settings, vastly reducing reliance on costly manual annotation. For instance, the MICROPHONE label exhibits a 67% F_1 gain, as paraphrases like "record audio notes" and "capture voice memos" help the model generalize beyond limited lexical signals. By synthetically expanding small, high-confidence datasets, this approach mitigates the persistent cold start problem faced by users and developers alike.

Limitations and Mitigation Strategies

We acknowledge that our experiments focus on Android permissions, and other ecosystems may exhibit different patterns. Moreover, the applicability of our method may not extend to low-resource languages. We present a short self-assessment regarding other threats to the validity of our approach:

1) *Annotation quality:* Even when high-agreement (Cohen's $\kappa = 0.82$), annotations may introduce noise due to subjective interpretations (e.g., of vague language like "enhance user experience"). We mitigate this by releasing our dataset for community scrutiny.
2) *Fixed decision threshold:* Using a threshold of 0.5 may bias predictions if class-specific optimal thresholds differ. We address this by reporting threshold-agnostic metrics (ROC-AUC, PR-AUC) alongside traditional F_1 scores.
3) *Class imbalance:* Despite augmentation, rare permissions (e.g., SETTINGS) may still skew macro-F_1 scores. We account for this by reporting both macro and weighted F_1, with detailed per-category results discussed in Sec. IV.

VI. Conclusion and Future Work

This work addresses the challenge of automated permission inference from app descriptions, a critical task for privacy auditing, by rigorously evaluating zero-shot approaches using LLMs as well as fine-tuned encoder models. Our findings are threefold: (1) By augmenting imbalanced training data through paraphrasing, fine-tuned models like BERT and BART achieve state-of-the-art performance (25% PR-AUC gains), demonstrating that linguistic diversity, not just dataset size, determines success. (2) GPT-4o matches or exceeds fine-tuned

baselines (18% F_1 improvements for sensitive permission categories like LOCATION) without task-specific training, validating the utility of zero-shot inference as a low-effort, high-accuracy tool for practitioners. (3) Synthetic expansion of a small annotated dataset (250 instances) yields improvements of over 60% in F_1 scores, proving that strategic augmentation—even with relatively smaller language models—mitigates annotation costs without compromising quality. Our work shows that models, combined with robust data augmentation strategies, can reliably infer permissions even under constraints like limited model capacity or small annotated datasets.

Extensions of this research can span multiple avenues, including real-world auditing, where models can be deployed to detect discrepancies between inferred permissions and actual app manifests for proactive privacy monitoring. Our work can also be extended to conduct longitudinal evaluation of models, as apps and their descriptions evolve. Another promising direction is adversarial testing of model robustness, to handle intentionally obfuscated app descriptions (e.g., "enhance user experience" to mask data collection).

ACKNOWLEDGEMENTS

This work was supported in part by the Cybersecurity Center of the State of Colorado (SB 18-086), by the U.S. National Science Foundation (Award Nos. CNS-2335686, CNS 2335687, DMS 2123761), by the U.S. National Institute of Standards and Technologies (Award No. 60NANB23D152).

REFERENCES

[1] S. Liu, B. Zhao, R. Guo, G. Meng, F. Zhang, and M. Zhang, "Have you been properly notified? Automatic compliance analysis of privacy policy text with GDPR Article 13," in *Proc. of WWW, 2021*. ACM, 2021, pp. 2154–2164.

[2] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh, "MAPS: Scaling privacy compliance analysis to a million apps," *in Proc. of PETS, 2019*, vol. 2019, pp. 66–86, 2019.

[3] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PloS One*, vol. 6, no. 12, p. e28071, 2011.

[4] Y.-A. De Montjoye, S. Gambs, V. Blondel, G. Canright, N. De Cordes, S. Deletaille *et al.*, "On the privacy-conscientious use of mobile phone data," *Scientific Data*, vol. 5, no. 1, p. 180286, 2018.

[5] K. Huckvale, J. Prieto, M. Tilney, P.-J. Benghozi, and J. Car, "Unaddressed privacy risks in accredited health and wellness apps: a cross-sectional systematic assessment," *BMC Medicine*, vol. 13, p. 214, 2015.

[6] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and privacy analysis of mobile health applications: The alarming state of practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018.

[7] J. Benjumea, J. Ropero, O. Rivera-Romero, E. Dorronzoro-Zubiete, and A. Carrasco, "Privacy assessment in mobile health apps: Scoping review," *JMIR Mhealth and Uhealth*, vol. 8, no. 7, p. e18868, 2020.

[8] P. Faruki, A. Bharmal, V. Laxmi, V. Ganmoor, M. S. Gaur, M. Conti, and M. Rajarajan, "Android security: a survey of issues, malware penetration, and defenses," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 998–1022, 2014.

[9] Z. Qu, V. Rastogi, X. Zhang, Y. Chen, T. Zhu, and Z. Chen, "Autocog: Measuring the description-to-permission fidelity in android applications," in *Proc. of CCS, 2014*, 2014, pp. 1354–1365.

[10] R. Pandita, X. Xiao, W. Yang, W. Enck, and T. Xie, "{WHYPER}: Towards automating risk assessment of mobile applications," in *Proc. of USENIX Security, 2013*, 2013, pp. 527–542.

[11] Y. Feng, L. Chen, A. Zheng, C. Gao, and Z. Zheng, "AC-Net: Assessing the consistency of description and permission in Android apps," *IEEE Access*, vol. 7, pp. 57 829–57 842, 2019.

[12] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov, "Android permissions remystified: A field study on contextual integrity," in *Proc. of USENIX Security, 2015*. Washington, D.C.: USENIX Association, 2015, pp. 499–514.

[13] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner, "Android permissions: User attention, comprehension, and behavior," in *Proc. of SOUPS, 2012*, 2012, pp. 1–14.

[14] "Android 6.0 changes," https://developer.android.com/about/versions/marshmallow/android-6.0-changes, Accessed: April 2, 2025.

[15] B. Bonné, S. T. Peddinti, I. Bilogrevic, and N. Taft, "Exploring decision making with Android's runtime permission dialogs using in-context surveys," in *Proc. of SOUPS, 2017*. USENIX Association, 2017, pp. 195–210.

[16] M. Fassl, L. T. Gröber, and K. Krombholz, "Stop the consent theater," in *Proc. of CHI, 2021*, ser. CHI EA '21. ACM, 2021.

[17] Federal Trade Commission, "Truth in advertising," http://www.ftc.gov/news-events/topics/truth-advertising, Accessed: April 2, 2025.

[18] J. Bhatia, T. D. Breaux, J. R. Reidenberg, and T. B. Norton, "A theory of vagueness and privacy risk perception," in *Proc. of RE, 2016*. IEEE, 2016, pp. 26–35.

[19] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu *et al.*, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, 2024.

[20] D. Rodriguez, I. Yang, J. M. Del Alamo, and N. Sadeh, "Large language models: A new approach for privacy policy analysis at scale," *Computing*, vol. 106, no. 12, pp. 3879–3903, 2024.

[21] C. Tang, Z. Liu, C. Ma, Z. Wu, Y. Li, W. Liu, D. Zhu, Q. Li, X. Li, T. Liu *et al.*, "Policygpt: Automated analysis of privacy policies with large language models," *arXiv preprint arXiv:2309.10238*, 2023.

[22] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," OpenAI, Tech. Rep., 2023.

[23] S. J. Oishwee, N. Stakhanova, and Z. Codabux, "Large language model vs. stack overflow in addressing android permission related challenges," in *Proc. of MSR, 2024*, 2024, pp. 373–383.

[24] H. Alecakir, B. Can, and S. Sen, "Attention: There is an inconsistency between android permissions and application metadata!" *International Journal of Information Security*, pp. 1–19, 2021.

[25] Q. N. Nguyen, N. T. Cam, and K. Van Nguyen, "XLMR4MD: New Vietnamese dataset and framework for detecting the consistency of description and permission in android applications using large language models," *Computers & Security*, vol. 140, p. 103814, 2024.

[26] A. P. Felt, S. Egelman, and D. Wagner, "I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns," in *Proc. of SPSM, 2012*, 2012, pp. 33–44.

[27] T. Watanabe, M. Akiyama, T. Sakai, H. Washizaki, and T. Mori, "Understanding the inconsistency between behaviors and descriptions of mobile apps," *IEICE Transactions on Information and Systems*, vol. 101, no. 11, pp. 2584–2599, 2018.

[28] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tutor Quant Methods Psychol*, vol. 8, no. 1, pp. 23–34, 2012.

[29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT, 2019*, J. Burstein, C. Doran, and T. Solorio, Eds. ACL, 2019, pp. 4171–4186.

[31] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. of ACL, 2020*. ACL, 2020, pp. 7871–7880.

[32] W. Yin, J. Hay, and D. Roth, "Benchmarking zero-shot text classification: Datasets, evaluation and entailment Approach," in *Proc. of EMNLP-IJCNLP, 2019*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. ACL, 2019, pp. 3914–3923.

[33] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. of USENIX Security, 2021*. USENIX Association, 2021, pp. 2633–2650.