Context-Aware Detection of Specialized Language

Valerie Lopez, Adam Bawatneh, Dr. Santu Karmaker, Dr, Song Wang, Dr, Mubarak Shah
University of Central Florida, Orlando, FL

Overview

People often encounter unfamiliar language that is not well covered by dictionaries, search engines, or standard translation tools. Examples include newly coined slang, undefined acronyms, domain-specific jargon, technical terms, coded phrases, dog whistles, and multilingual expressions. These terms can be difficult to interpret because they may appear only a few times, depend on a narrow context, change meaning across communities, or have meanings that are not obvious from the surrounding text alone.

We propose a context-aware **detection pipeline, CADS,** to identify specialized language. Rather than relying only on keyword spotting or direct large-language-model prompting, this project explores how multiple weak signals can be combined to detect unusual terms.

Methodology

CADS is composed of four detection methods that evaluate single and multi-word spans before fusing predictions into a single score. Thresholds are determined by slang dataset, Genz-A's, performance.

Machine Translation

- Round-trip translation is applied four separate times, with Spanish, Arabic, Chinese, Japanese
- Mistranslations and omitted spans are compiled to the candidate set that evaluation methods inherit
- Consecutive missing words form phrase candidates, rather than relying on permutations

Evaluation Methods

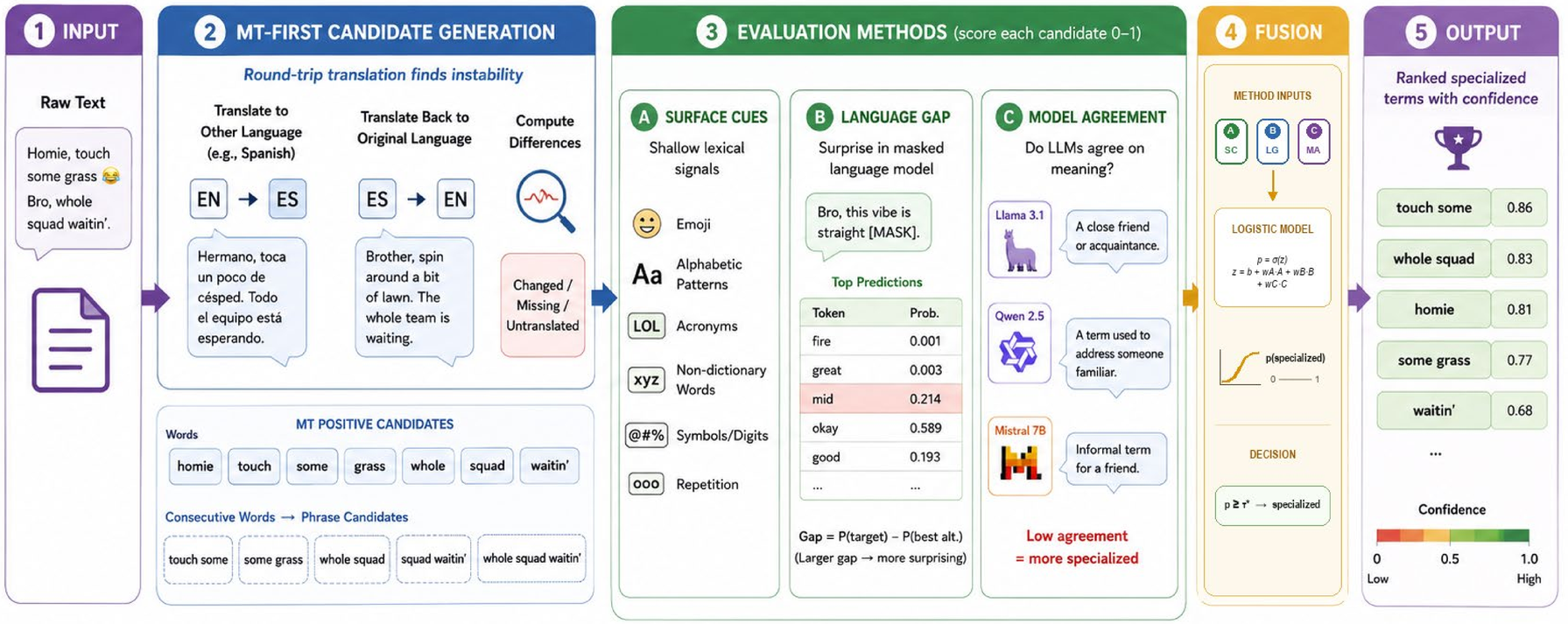
- Surface Clues: Initial observations based on punctuation, capitalization, repetition
- Likelihood Gap: Candidates are masked to evaluate the unexpectedness of their appearance
- Model Agreement: Semantic similarity between three model definitions

Fusion Model

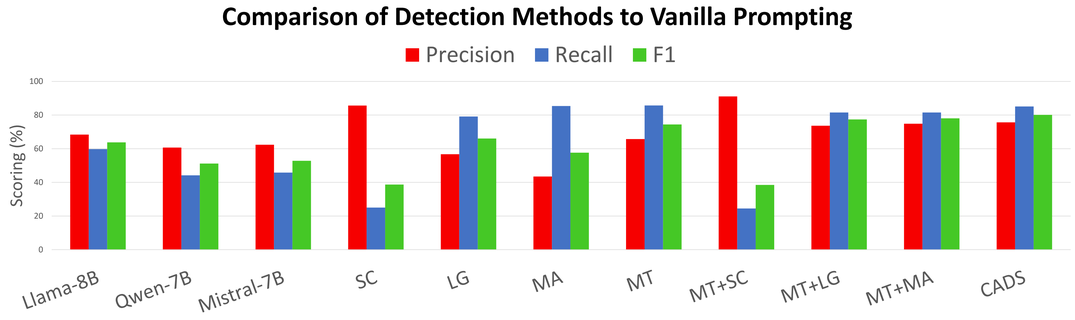
- Logistic regression model that learned the weights of evaluation methods combines predictions into a single score

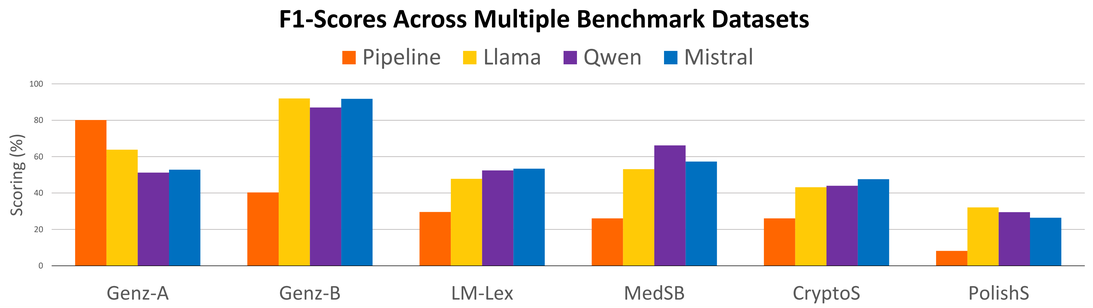
CADS Architecture

Machine Translation finds candidates → other methods score them → fusion produces specialized terms



Analysis





Key Takeaways

- While MT rivals CADS, the need for evaluation methods is displayed by increased precision and F1
- CADS outperforms 3 LLMs on Genz-A English slang dataset but fails on new domains and languages due to poor threshold transfer
- CADS is still in its infancy and requires further tuning and calibration across additional languages and domains
- Future studies will train a model on the parameters to better calibrate the pipeline for various domains

References

[1] Grattafiori, A., et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783. Retrieved from https://arxiv.org/abs/2407.21783
[2] Yang, A., et al. (2024). Qwen2.5 Technical Report. arXiv:2412.15115. Retrieved from https://arxiv.org/abs/2412.15115
[3] Jiang, A. Q., et al. (2023). Mistral 7B. arXiv:2310.06825. Retrieved from https://arxiv.org/abs/2310.06825
[4] Liu, Y., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692
[5] Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084. Retrieved from https://arxiv.org/abs/1908.10084