University of Toronto

The Optimal Language Is All You Need?

CSCC11 Group Project Proposal

Divy, Jason, Kevin, Owen
@mail.utoronto.ca

October 30, 2025

1 Overview

Large Language Models (LLMs) have rapidly advanced natural language processing. Most state-of-the-art models rely on the transformer architecture and focus heavily on using attention to extract the semantic meaning behind the words.

Different languages show vastly different ambiguity profiles. English exhibits dense lexical ambiguity, whereas Turkish, with its agglutinative morphology, provides richer disambiguation cues. We will quantify how quickly a transformer recovers the intended sense under these contrasting typological regimes.

2 Motivation

Recent transformer work prioritizes aggregate accuracy, perplexity, or broader modal coverage, yet few studies rigorously quantify lexical sense resolution itself.

Without precise measurements of how swiftly and with what structural economy a model disambiguates polysemous tokens across languages, we lack the evidence needed to steer fine-tuning policies or pruning strategies.

By mapping the layer-wise emergence of disambiguation and profiling the linguistic cues that catalyze it, this research supplies actionable diagnostics for adapting models to typologically diverse corpora while preserving efficiency.

3 Methodology

We evaluate multilingual disambiguation by pairing the transformer models in Section 3.1, the curated cross-lingual corpora in Section 3.2, and the difficulty-aware metric suite in Section 3.3, enabling a consistent comparison of how contextual representations evolve across languages.

3.1 Models

3.1.1 Model Families

To ensure a coverage over most common language model families, we plan to test the languages on multiple model families namely

1. Architecture: Encoder-Only vs Decoder-Only vs Encoder-Decoder
2. Tokenizer: WordPiece vs Unigram (SentencePiece) vs BPE (SentencePiece)
3. Objective: MaskedLM vs Seq2Seq Denoising vs Causal LM vs Contrastive

3.1.2 Models

To ensure coverage across the 3 family axes, we plan to benchmark on the following models

1. mBERT (Encoder-only, WordPiece, MaskedLM) [3]
2. XLM-R base (Encoder-only, Unigram, MaskedLM) [2]
3. mT5 small (Encoder–Decoder, Unigram, seq2seq denoising) [20]
4. Llama 3.x 8B (Decoder-only, SentencePiece–BPE, Causal LM) [4]
5. LaBSE (Encoder-only, WordPiece, Contrastive) [6]
6. MiniLM-L12-v2 (Encoder-only, WordPiece, Contrastive) [19]

According to our hypothesis, if our context burden truly ties to topology, the effects should be consistent across all these family of models.

3.2 Datasets

We ground the study in three multilingual sense-evaluation corpora:

- **XL-WSD** [12]: token-level, cross-lingual sense annotations aligned across inventories, enabling consistent disambiguation depth measurements.
- **XL-WiC** [13]: sentence-pair contrasts driven by contextual variation, supporting layer-wise probes of sense separability.
- **MCL-WiC** [10]: typologically balanced coverage, permitting controlled comparisons among languages with different ambiguity profiles.

3.3 Metrics

We group our evaluations into three layers:

- *Core metrics*: primary cross-language disambiguation measures.
- *Supporting metrics*: stress tests and controls for the core suite.
- *Explanatory metrics*: linguistic signals that contextualize the results.

We additionally introduce the Context Burden Score (CBS), a composite statistic that quantifies per-language disambiguation difficulty for the model.

3.3.1 Statistical Aggregaion

To normalize cross-lingual and lemma-specific difficulty, every metric is aggregated with a hierarchical Bayesian partial-pooling model. For each instance we compute a raw score $y_{m,\ell}$ for metric m and lemma ℓ, then fit

$$y_{m,\ell} \sim \mathcal{N}(\mu + \alpha_m + \beta_\ell, \sigma^2), \qquad \alpha_m \sim \mathcal{N}(0, \sigma_\alpha^2), \qquad \beta_\ell \sim \mathcal{N}(0, \sigma_\beta^2),$$

and report the posterior means as difficulty-adjusted values. This shrinkage keeps the metrics comparable even when some language-lemma combinations are data-sparse.

3.3.2 Context Burden Score (CBS)

We condense the core metrics into a language-level Context Burden Score that reflects how challenging disambiguation is for each language.

For any metric $x \in \{\mathrm{DDI}, \mathcal{S}, H_{L^\star}, \mathrm{HB}_{L^\star}\}$ we form a unit-interval normalization

$$\hat{x}_\ell = \frac{x_\ell - \min_{L \in \mathcal{L}} x_L}{\max_{L \in \mathcal{L}} x_L - \min_{L \in \mathcal{L}} x_L},$$

where $\mathcal{L}$ indexes languages and the inputs x are the posterior means from the partial-pooling model. Higher sense-separability indicates easier disambiguation, so we use its complement $\tilde{\mathcal{S}}_\ell = 1 - \hat{\mathcal{S}}_\ell$ when measuring burden.

The resulting CBS is the linear combination

$$\mathrm{CBS}_\ell = \lambda_1 \hat{\mathrm{DDI}}_\ell + \lambda_2 \tilde{\mathcal{S}}_\ell + \lambda_3 \hat{H}_{L^\star,\ell} + \lambda_4 \hat{\mathrm{HB}}_{L^\star,\ell},$$

with non-negative weights λ that sum to one and are selected to best align with held-out disambiguation error.

Larger CBS values therefore correspond to languages that require deeper layers, less separable senses, diffuse attention, or larger head budgets, signalling greater contextual burden.

3.3.3 Core Metrics

We estimate disambiguation efficiency with the following primary metrics:

Disambiguation Depth Index (DDI). Let $\mathcal{S}(L)$ denote the sense-separability score at layer L. We define

$$\mathrm{DDI} = \min\{L \mid \mathcal{S}(L) \geq \tau\},$$

where τ is a language-specific reliability threshold derived from held-out validation accuracy. This captures the earliest layer that stably encodes the correct sense and builds on prior layer-wise probing analyses [1, 15].

Sense Separability. For each layer we evaluate:

1. Linear probes (logistic regression, MLP, k-NN) trained on gold sense labels to obtain $\mathcal{S}_{\text{probe}}(L)$ [1, 17].
2. Cluster purity and adjusted Rand index (ARI) after k-means, yielding $\mathcal{S}_{\text{cluster}}(L) = \frac{1}{N}\sum_k \max_j n_{kj}$ and $\text{ARI}(L)$, respectively [15].
3. Fisher Discriminant Ratio (FDR) with $\text{FDR}(L) = \sum_{c \neq c'} \frac{(\mu_c - \mu_{c'})^2}{\sigma_c^2 + \sigma_{c'}^2}$ over sense centroids μ and variances σ^2 [7].

Attention Entropy at DDI. For the DDI layer $L^\star$, we compute the normalized attention distribution $a^{(L^\star)}$ over context tokens and report

$$H_{L^\star} = -\sum_i a_i^{(L^\star)} \log a_i^{(L^\star)},$$

which reflects how concentrated the model's focus is when disambiguation emerges, following entropy-based attention diagnostics [18].

Head Budget at DDI. We randomly mask attention heads at $L^\star$ and sweep retention ratios $p \in \{0.25, 0.5, 0.75, 1.0\}$. The head budget is

$$\text{HB}_{L^\star} = \min\{p \mid \Delta\text{Acc}_{L^\star}(p) \leq \epsilon\},$$

the smallest fraction of heads that preserves probe performance within tolerance ϵ, akin to structured head-pruning studies [11].

3.3.4 Supporting Metrics

These metrics interrogate modelling assumptions and surface corroborating evidence for the core suite.

Linear–MLP–kNN Agreement. For each layer L we compare the predicted labels produced by the linear probe $\hat{y}_{i,L}^{\text{lin}}$, the MLP probe $\hat{y}_{i,L}^{\text{mlp}}$, and the kNN classifier $\hat{y}_{i,L}^{\text{kNN}}$ over N evaluation instances. The agreement score

$$\text{Agree}(L) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{y}_{i,L}^{\text{lin}} = \hat{y}_{i,L}^{\text{mlp}} = \hat{y}_{i,L}^{\text{kNN}}\right]$$

highlights layers where linear and non-linear decision boundaries align, a proxy for probe stability advocated in recent probing diagnostics [8, 14]. We pool $\text{Agree}(L)$ across instances with the same hierarchical model used for the core metrics to obtain language-level difficulty-adjusted agreement.

Cross-Layer Separability Gain (CLSG). Let $L^{(0)}$ denote the token embedding layer and $L^{(x)}$ a target layer of interest. Re-using the Fisher Discriminant Ratio $\mathrm{FDR}(\cdot)$ from the core metrics, the gain is

$$\mathrm{CLSG}(L^{(x)}) = \mathrm{FDR}(L^{(x)}) - \mathrm{FDR}(L^{(0)}),$$

with positive values indicating improved sense separability relative to the embedding baseline, mirroring cross-layer transfer analyses in prior work [9, 15]. We report the posterior mean of $\mathrm{CLSG}(L^{(x)})$ per language.

Embedding-Only Sense Score. To quantify how much sense information is recoverable without contextualization, we score the layer-0 representations using a linear probe and compute the macro F_1:

$$\mathrm{F1}_{L^{(0)}} = \frac{2 \cdot \mathrm{Prec}_{L^{(0)}} \cdot \mathrm{Rec}_{L^{(0)}}}{\mathrm{Prec}_{L^{(0)}} + \mathrm{Rec}_{L^{(0)}}},$$

where the precision and recall are aggregated across senses. High $\mathrm{F1}_{L^{(0)}}$ suggests the model encodes sense-specific information before contextual layers are applied, providing a lower bound for the core evaluations and echoing observations about contextualization gains in [5, 9].

3.3.5 Explanatory Metrics

These diagnostics surface linguistic pressure points that may support the core and supporting metrics.

Tokenization Fragmentation Factor (TFF). For each occurrence i of lemma w in language ℓ we let $c_{i,\ell}$ denote the number of wordpieces produced by the tokenizer. The fragmentation factor averages this count:

$$\mathrm{TFF}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} c_{i,\ell},$$

where N_ℓ is the total number of lemma occurrences examined in language ℓ. Large TFF_ℓ values indicate heavier tokenization splitting, which can raise contextual burden for the disambiguation probes and mirrors tokenizer-focused analyses in multilingual models [16].

Polysemy Load Index (PLI). Let $f_\ell(w)$ be the corpus frequency of lemma w in language ℓ, and $s(w)$ the number of senses listed for w in the WSD inventory. The polysemy load is the frequency-weighted average

$$\mathrm{PLI}_\ell = \frac{\sum_{w \in \mathcal{V}_\ell} f_\ell(w)\, s(w)}{\sum_{w \in \mathcal{V}_\ell} f_\ell(w)},$$

with vocabulary $\mathcal{V}_\ell$ restricted to lemmas observed in the evaluation set. Higher PLI_ℓ suggests that a language exposes the model to denser sense inventories, echoing findings that contextual models struggle with multi-sense distributions [5].

4 Timeline

Milestone	Activities and Deliverables
End of Oct	Finalize literature review.
Nov (Week 1)	Start the codebase; integrate models and data pipeline for benchmarking.
Nov (Week 2)	Complete core metric pipeline for mBERT.
Nov (Week 3)	Extend core metrics to XLM-R, mT5, Llama, LaBSE, MiniLM-L12; implement CBS.
Nov (Week 4)	Visualize metrics per language, organizing tables by model.
December	Develop the supporting metrics suite.
January	Build explanatory metrics and finalize the report draft.
February	Buffer month for contingency work.
March	Additional buffer month.
April	Finalize submission package and submit the paper.

Table 1: Planned research schedule with weekly checkpoints for November and monthly milestones thereafter.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 2019.

[4] Abhimanyu Dubey, Sébastien Bubeck, Thomas Darcet, Michael Denneulin, Ronen Eldan, Suriya Gunasekar, Kuang Hu, Sham Kakade, Yin Tat Lee, Hongseok Namkoong, et al. The Llama 3 herd of models. Technical report, Meta AI, 2024.

[5] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP*, 2019.

[6] Fuli Feng, Yinbo Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of ACL*, 2020.

[7] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[8] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*, 2019.

[9] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, 2019.

[10] Federico Martelli, Alessandro Miaschi, Maria Costanza, and Bernardo Magnini. MCL-WiC: A multilingual evaluation benchmark for word-in-context tasks. In *Proceedings of EACL*, 2021.

[11] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Proceedings of NeurIPS*, 2019.

[12] Tommaso Pasini and Roberto Navigli. XL-WSD: Evaluating cross-lingual word sense disambiguation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

[13] Tommaso Pasini, Alessandro Raganato, José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of EMNLP*, 2020.

[14] Tiago Pimentel, Ryan Cotterell, Sebastian Ruder, Hila Gonen, Matthew E. Peters, and Phil Blunsom. Information-theoretic probing for linguistic structure. In *Proceedings of EMNLP*, 2020.

[15] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of BERT. *arXiv preprint arXiv:1906.02715*, 2019.

[16] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, Goran Glavaš, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL*, 2021.

[17] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, 2019.

[18] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 2019.

[19] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of NeurIPS*, 2020.

[20] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, 2021.