University of Stuttgart
Germany

Bridging the Thresholding Gap

Ordinal Classification for Word-in-Context using Cumulative Link Models

Master's Thesis Defense

Alp Mujko

M.Sc. Artificial Intelligence and Data Science

Examiners: Prof. Dr. Sabine Schulte im Walde, Dr. Agnieszka Faleńska

Supervisor: Dr. Dominik Schlechtweg

Institut für Maschinelle Sprachverarbeitung · Universität Stuttgart

From Binary to Ordinal Semantics

The evolution of the Word-in-Context task

Binary Word-in-Context

Pilehvar et al. (2019)

TRUE / FALSE

- "She hurt her **arm**." vs. "They **arm** the soldiers."
- "She sat on the **bank**." vs. "My money is in the **bank**."

This forces **same/different dichotomy** which **oversimplifies** the **semantic continuum**.

Ordinal Word-in-Context

Schlechtweg et al. (2025)

4-point DURel scale

- "She hurt her **arm**." vs. "The statues **arm** got damaged."
- "The blood **bank** is out of blood." vs. "My money is in the **bank**."

A meaningful ordering of the **semantic continuum** is needed.

The DURel Scale

A linguistically grounded 4-point ordinal scale for semantic proximity

4	**Identical** *Identity*	"…opened a vein in her little **arm**…" vs. "…within reach of his **arm**…"
3	**Closely Related** *Context Variance*	"…the disembodied **arm** of the Statue of Liberty…" (replica; non-prototypical body part)
2	**Distantly Related** *Polysemy*	"…overlooking an **arm** of the sea…" (metaphorical extension (not always))
1	**Unrelated** *Homonymy*	"…taxed to pay for the **arms**, ammunition…" (weapons; no semantic relation)

*Target word: **arm** · examples from Schlechtweg et al. (2025)*

Sentence-BERT

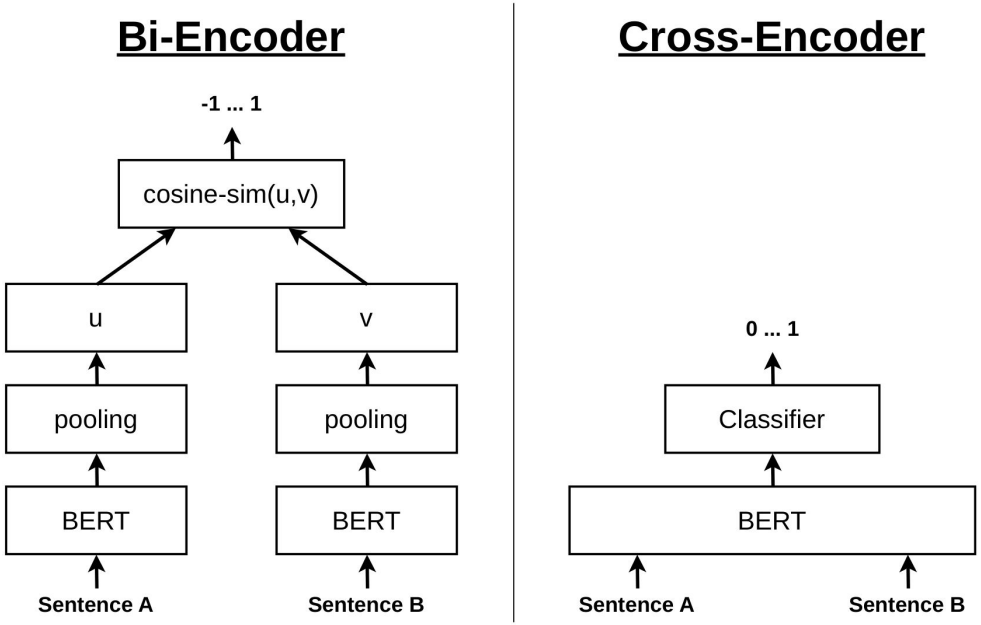
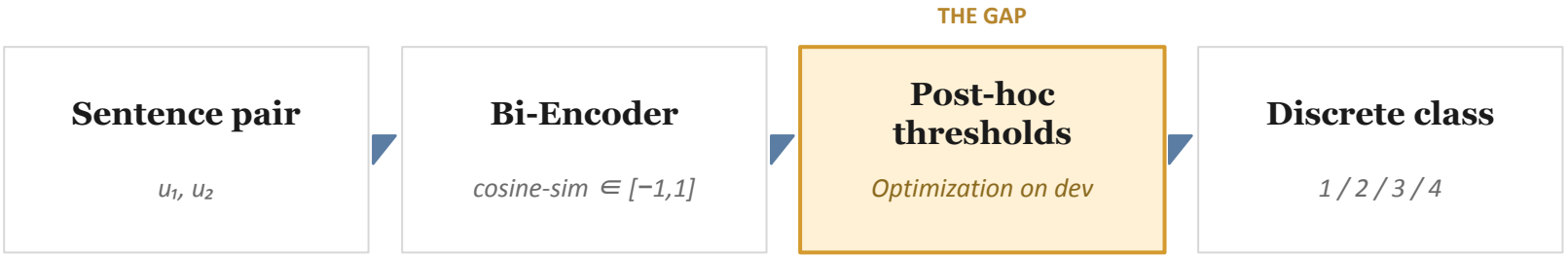


Diagram from Reimers et al. (2019)

State of the Art & the Thresholding Gap

XL-DURel (Yadav & Schlechtweg, 2025) — the current best OGWiC model



Why this is a problem

- Structural disconnect: training optimizes a continuous ranking objective, inference requires discrete ordinal classes.
- Calibration depends on labeled development data — unavailable in zero-shot deployment on new domains or languages.
- Thresholds are heuristic post-hoc fixes, not learned alongside the representation.

The Cumulative Link Model

McCullagh (1980)

Structural rank consistency via a latent score and ordered thresholds

Latent variable formulation

$$y* = f(x)$$

continuous latent score (classifier output)

$$P(y \leq q|x) = g(b_q - y*)$$

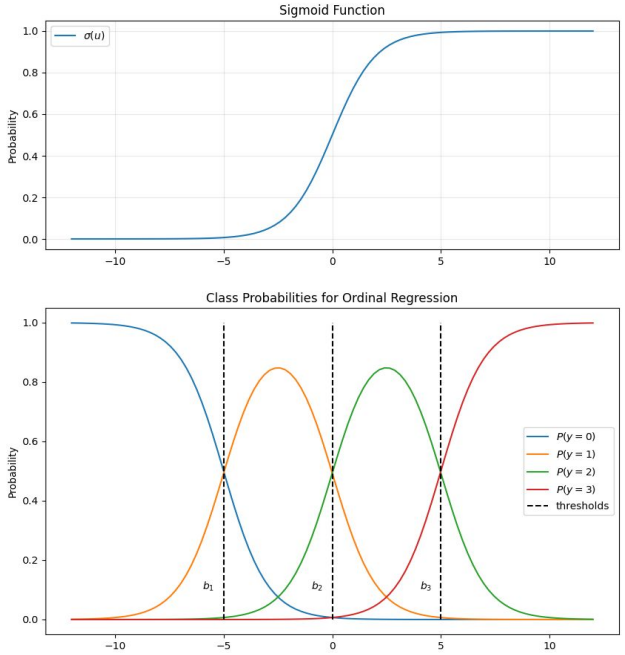
cumulative probability via link g(·)

$$g(u)=\sigma(u)=1/(1+e^{-u})$$

example g(·): Sigmoid Function

$$P(y = q|x) = P(y \leq q) - P(y \leq q - 1)$$

absolute class probability



Ordinal Loss Functions

Why cross-entropy fails on ordinal data — and what to use instead

*Both models below have **ground truth y = 4 (Identical)**. Cross-entropy penalizes them identically:*

Model A: predicts class 3

`P = [0.1, 0.1, 0.7, 0.1]`

CE loss: `−log(0.1) = 2.30`

Near miss: Context Variance ≠ Identical

Model B: predicts class 1

`P = [0.7, 0.1, 0.1, 0.1]`

CE loss: `−log(0.1) = 2.30`

Catastrophic: Homonymy ≠ Identical

Distance-aware alternatives evaluated in this thesis

Ordinal Log-Loss (OLL)

- $$\mathcal{L}_{\mathrm{OLL}-\alpha}(P, y) = -\sum_{i=1}^{N} \log(1 - p_i)\, d(y, i)^{\alpha}$$

- Penalizes **assigning probability** mass to **incorrect classes** in proportion to their **distance from the true class**.
- Hyperparameter α controls penalty steepness.

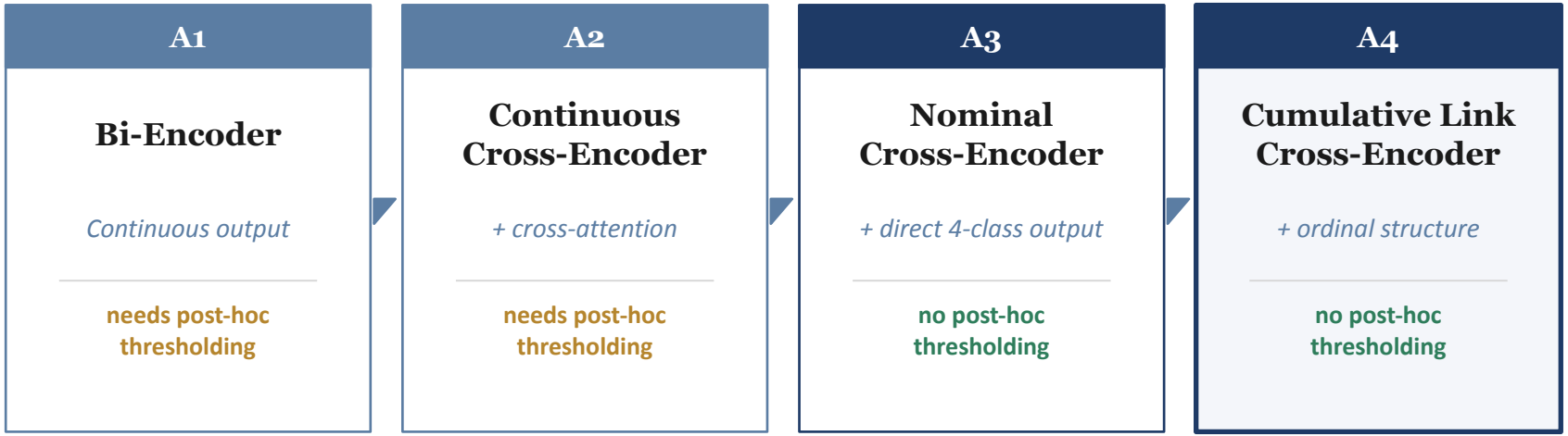
Squared Earth Mover's Distance (EMD²)

- $$\mathrm{EMD}^2(p, t) = \sum_{i=1}^{N} \left(\mathrm{CDF}_i(p) - \mathrm{CDF}_i(t)\right)^2$$

- Penalizes **differences between the predicted and target cumulative distributions**, with larger penalties when **probability mass** is shifted **farther from its correct positions**.

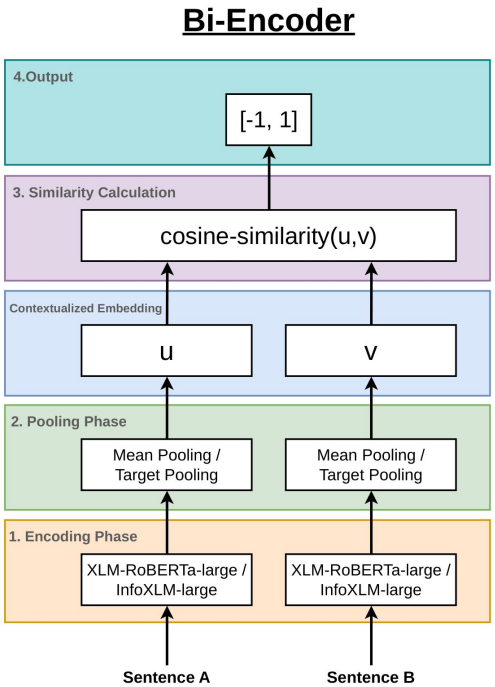
The Four-Architecture Progression

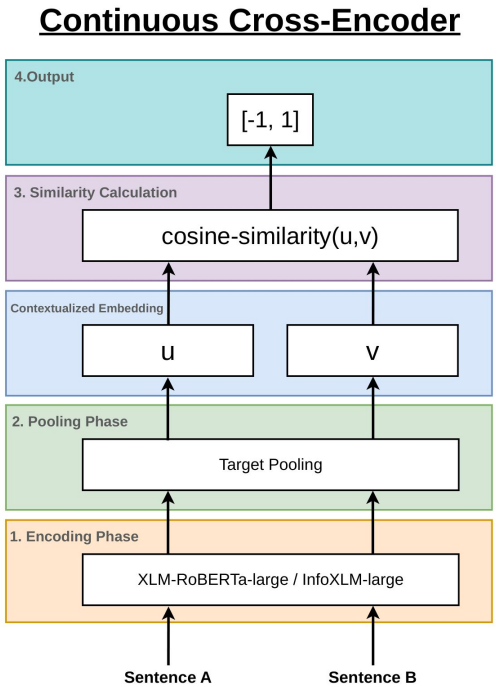
A controlled experiment: each architecture varies exactly one structural component



One-change-at-a-time principle: every transition isolates a single structural component — joint encoding (A1→A2), discrete output (A2→A3), ordinal structure (A3→A4) — so observed performance shifts can be causally attributed.

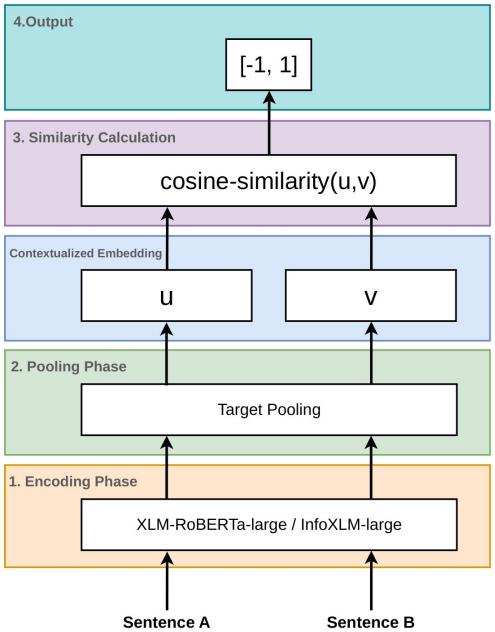
From Architecture 1 to Architecture 2



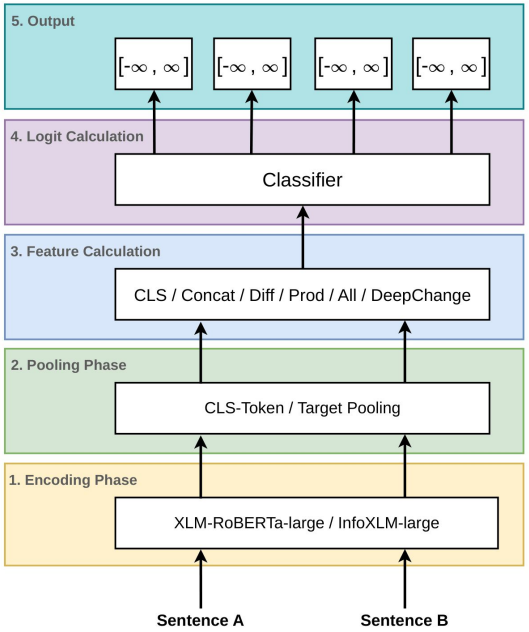


From Architecture 2 to Architecture 3

Continuous Cross-Encoder

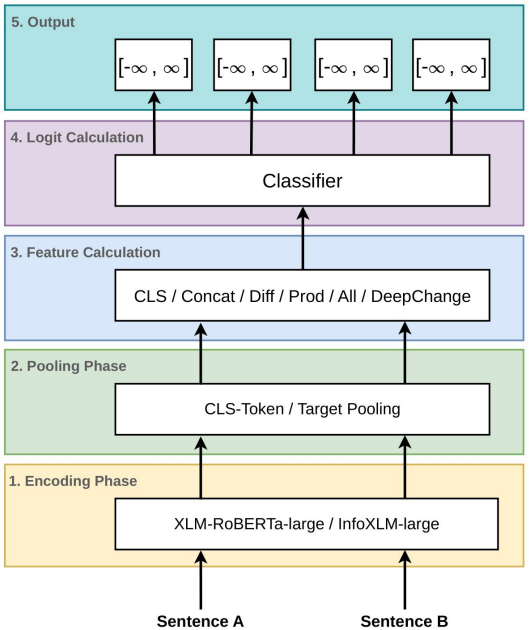


Nominal Cross-Encoder

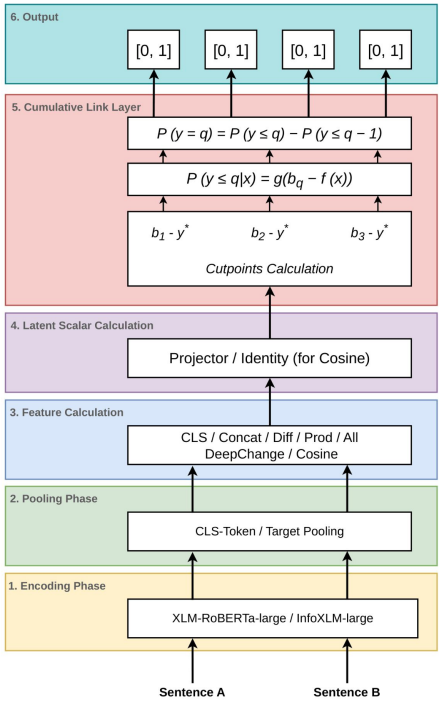


From Architecture 3 to Architecture 4

Nominal Cross-Encoder



Cumulative Link Cross-Encoder



Research Questions & Contributions

Three RQs answered through a controlled four-architecture progression

Research Questions

RQ1 **Does joint encoding (Cross-Encoder) outperform independent encoding (Bi-Encoder)?**

RQ2 **Does direct nominal classification improve the Cross-Encoder?**

RQ3 **Does explicit ordinal modeling via a Cumulative Link Model outperform the other architectures?**

Data

Three role-specific dataset categories

Primary Data
Ordinal supervision

- CoMeDi (DURel 1–4)
- 7 languages: DE, EN, ES, NO, RU, SV, ZH
- 47k train / 8k dev / 15k test

Auxiliary Training Data
Binary supervision (label-mapped)

- Original WiC, MCL-WiC, XL-LEXEME (whole & subset), SPCD , SweWiC, DanWiC

Zero-Shot Evaluation
Held-out ordinal benchmarks

- EN (DWUG), ES (DiaWUG), NL (DWUG-NL), SL (Slovenian), JA (Japanese LSCD)

Preprocessing

- Target words got marked with **<t>** & **</t>**.
- Label mapping: Binary ⟶ Ordinal
  - Binary 0 (Different Meaning) ⟶ Ordinal 2 (Polysemy)
  - Binary 1 (Similar Meaning) ⟶ Ordinal 4 (Identical)

Krippendorff's α for Evaluation

Ordinal-aware, chance-corrected agreement

Formula

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o = *observed disagreement*

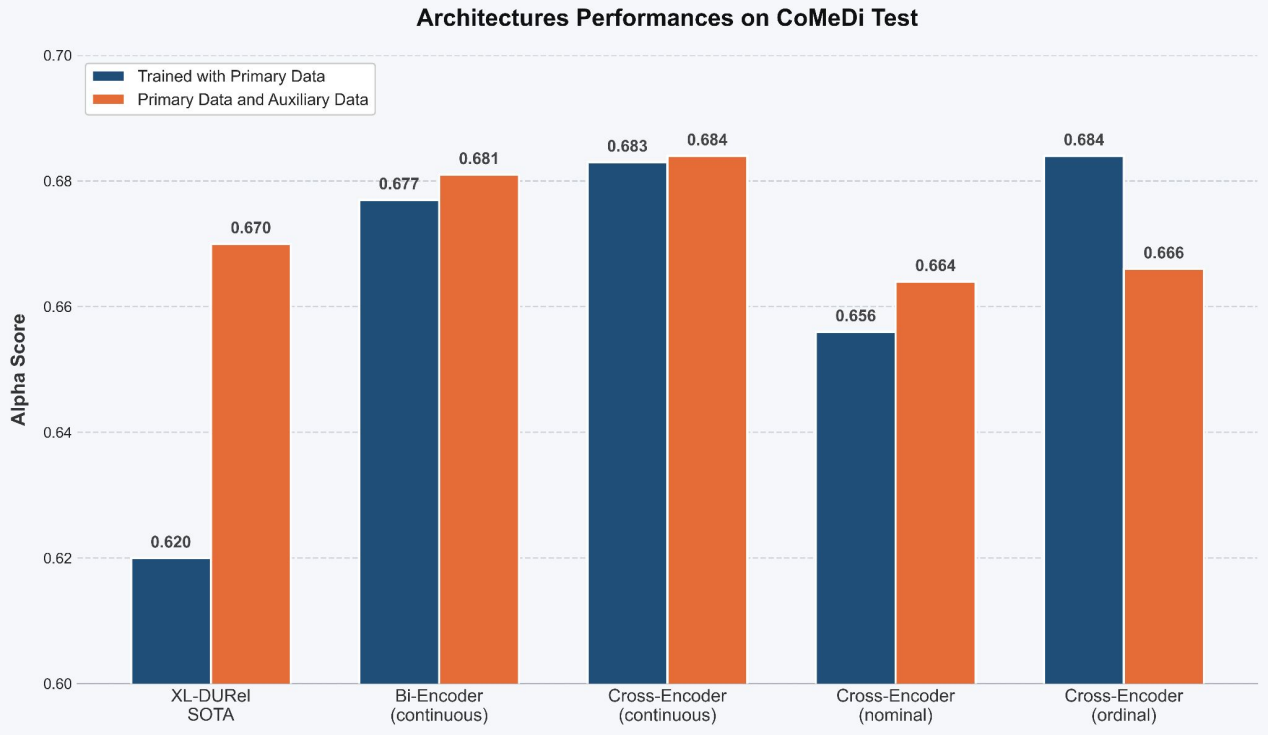
D_e = *expected disagreement under chance*

- Penalizes severity: 1→4 worse than 3→4.
- Chance-corrected: no reward for always predicting majority class.
- Robust to extreme class imbalance (63% Class 4 in CoMeDi).

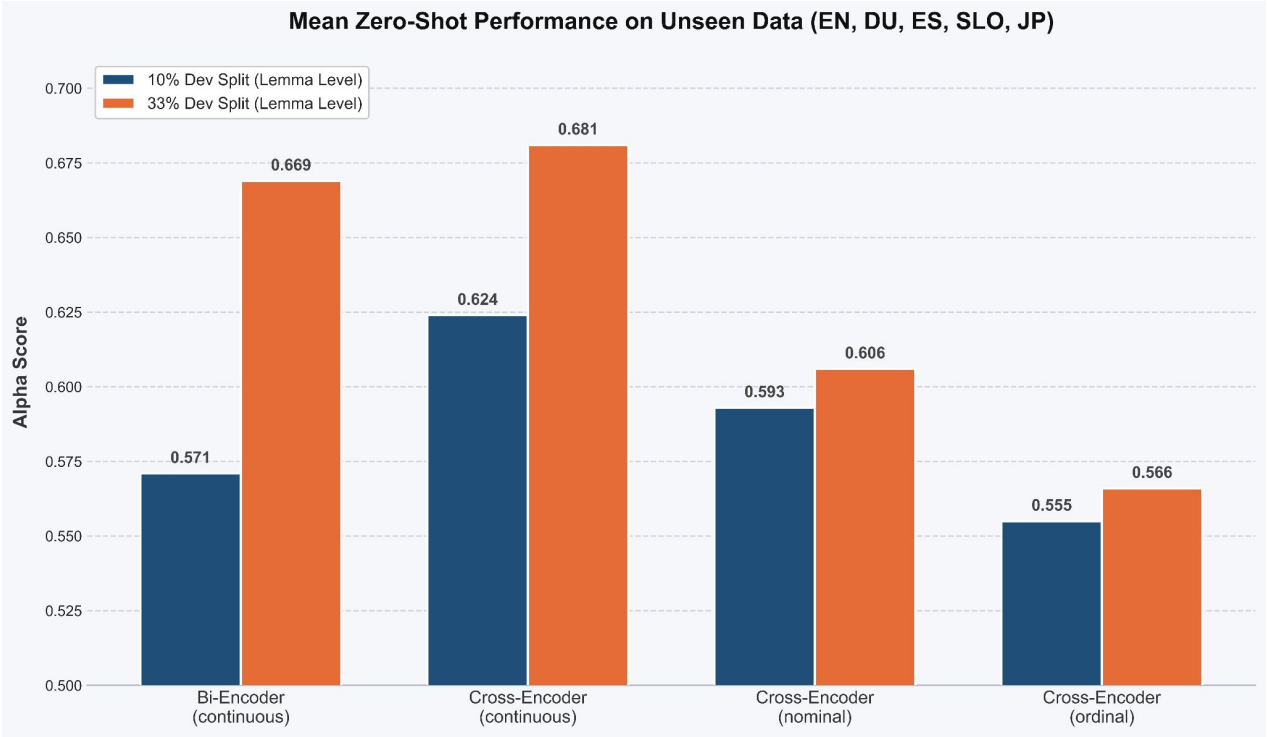
Interpretation

- α = 1 → perfect agreement.
- α = 0 → chance level.
- α = 0.66 → roughly ⅓ as many and ⅓ as severe errors as random.

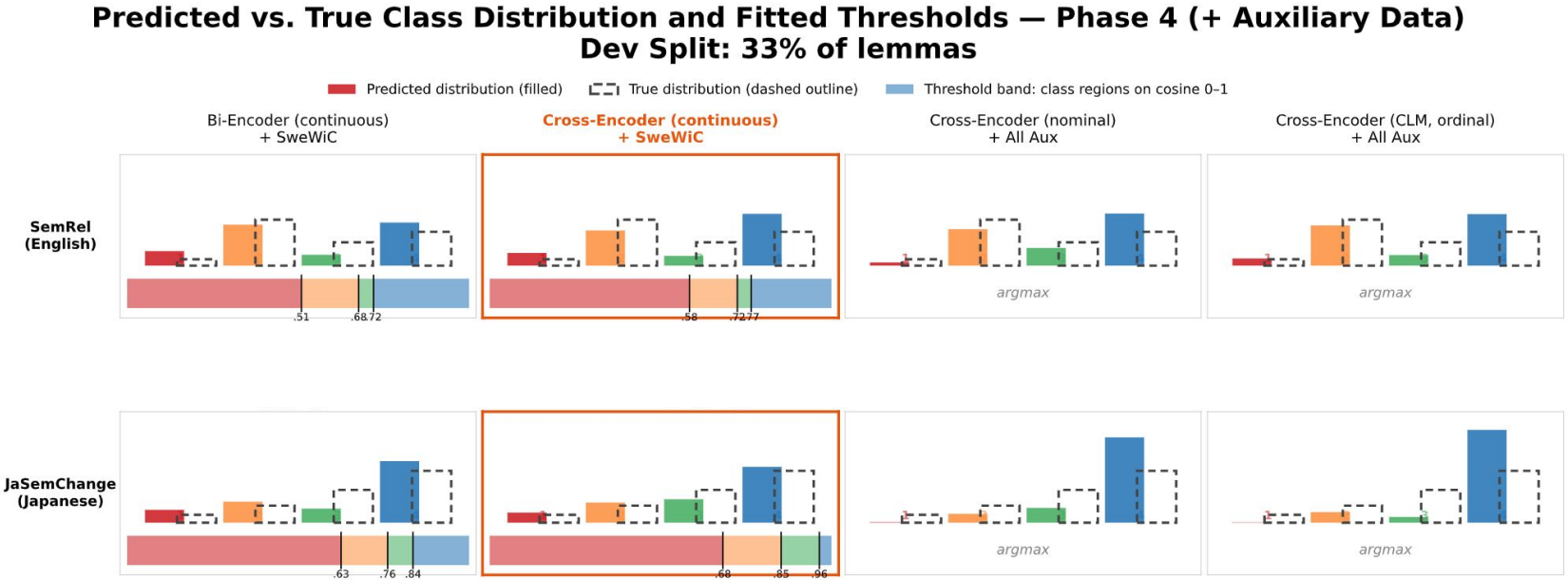
Results on CoMeDi



Results in Zero-Shot Scenario



Understanding Zero-Shot Results



Discussing the Results

On CoMeDi (withing distribution)

- **Cumulative Link Model:** Achieves the highest CoMeDi performance **(α = 0.684)** and derives semantically meaningful thresholds.
- **Bi-Encoder & Continuous Cross-Encoder:** Maintain strong, comparable performance driven by their training ranking objectives.
- **Nominal Cross-Encoder:** Yields the weakest results; treating labels as independent classes prevents the architecture from learning necessary ordinal relationships.

On Unseen Data (out of distribution)

- **Top OOD Performer:** Architectures 1 and 2 achieve the highest zero-shot performances when dev data is sufficient **(up to α = 0.681)**.
- **Poor Ordinal Generalization:** The **Cumulative Link Model fails to generalize** to unseen data, dropping to the **lowest performance** among all architectures.
- **Calibration Dependency:** Continuous models **rely heavily on dev set size**, showing dramatic gains when the split increases from 10% to 33%.

Answering the Research Questions

Each RQ admits a more nuanced answer than initially anticipated

RQ1 **Does joint encoding (Cross-Encoder) outperform independent encoding (Bi-Encoder)?** *Tentatively Yes*

- Cross-Encoder shows a small but consistent advantage,
- However, the gain is modest and the Cross-Encoder is more vulnerable when calibration data is scarce.

RQ2 **Does direct nominal classification improve the Cross-Encoder?** *Not really*

- Without auxiliary training data, Architecture 3 is the weakest model ($\Delta\alpha \approx -0.030$ vs. A2).
- With it, it largely catches up, but does not surpass Architecture 2.

RQ3 **Does explicit ordinal modeling via a Cumulative Link Model outperform the other architectures?** *Conditionally Yes*

- Within same data distribution, the CLM achieves the highest best-run score (0.684).
- Outside that regime, its learned thresholds become a liability: It ranks last on zero-shot generalization.

Conclusion & Future Work

The thresholding gap is reframed, not closed

The principal takeaway

- **Continuous Models:** Require calibration data at deployment, which can be beneficial if available.
- **Discrete Models:** Do not require calibration, but are constrained by training data.
- **The Takeaway:** Choose based on the data you have available in your target scenario.

Contributions

- Comparison of four different AI architectures
- First application of CLM to contextual semantic tasks
- Characterization of the thresholding gap
- Improvement of the SOTA model for OGWiC task

Future Work

- Class-balanced training
- Post-hoc threshold recalibration for CLM
- Instead of set thresholds, use calculated ones per inference
- Hybrid architectures: Bi-Encoder + ordinal heads
- Comparison to continuous Architectures with global set thresholds

References

- Peter McCullagh. Regression models for ordinal data. Journal of the Royal Statistical Society: Series B (Methodological), 42(2):109–127, 1980. doi: https://doi.org/ 10.1111/j.2517-6161.1980.tb01109.x. URL https://rss.onlinelibrary.wiley. com/doi/abs/10.1111/j.2517-6161.1980.tb01109.x.

- Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

- Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In Michael Roth and Dominik Schlechtweg, editors, Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation, pages 33–47, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics. URL https://aclanthology.org/2025. comedi-1.4/.

- Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: 133 Human Language Technologies, Volume 1 (Long and Short Papers), pages 1267– 1273, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1128. URL https://aclanthology.org/N19-1128/

- Sachin Yadav and Dominik Schlechtweg. XL-DURel: Finetuning sentence transformers for ordinal word-in-context classification. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh, editors, Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 338–351, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. URL https://aclanthology.org/ 2025.findings-ijcnlp.19/.

Thank you

Questions and discussion

Alp Mujko · Institut für Maschinelle Sprachverarbeitung · Universität Stuttgart

Detailed Results on CoMeDi

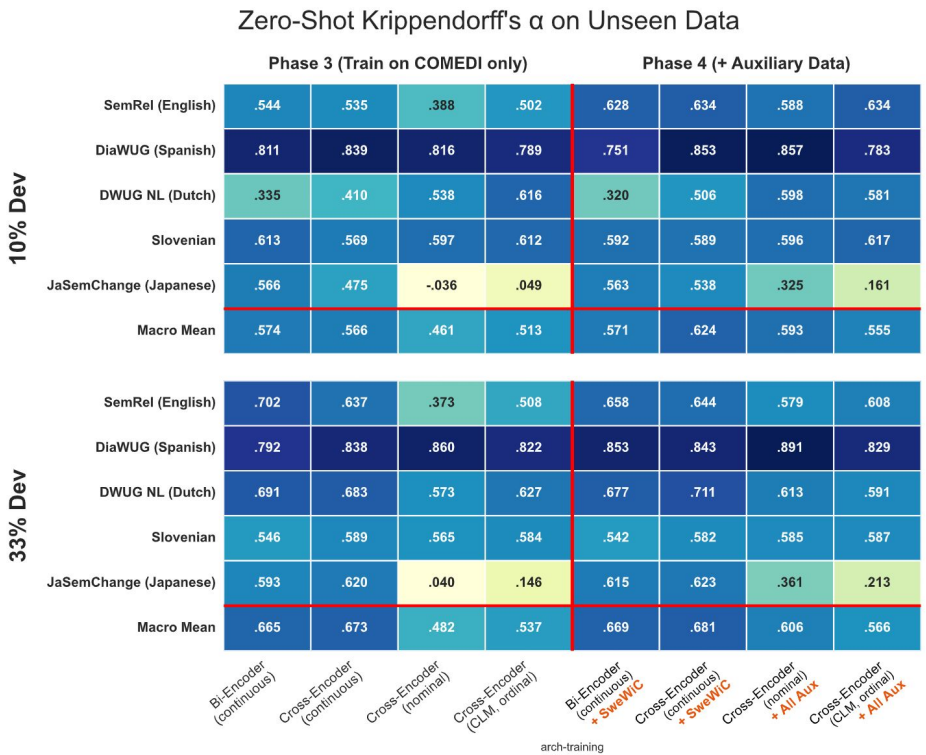
Cross-Architecture Synthesis — Mean α across seeds

	Phase 3 (Train on COMEDI)				Phase 4 (+ Auxiliary Data)			
	Bi-Encoder (continuous)	**Cross-Encoder (continuous)**	Cross-Encoder (nominal)	Cross-Encoder (CLM, ordinal)	Bi-Encoder (continuous) + SweWiC	Cross-Encoder (continuous) + SweWiC	Cross-Encoder (nominal) + All Aux	Cross-Encoder (CLM, ordinal) + All Aux
COMEDI_german_alpha	.780	.776	.730	.751	.771	.779	.736	.760
COMEDI_english_alpha	.739	.741	.727	.747	.742	.765	.687	.699
COMEDI_spanish_alpha	.702	.705	.736	.739	.717	.722	.736	.736
COMEDI_norwegian_alpha	.620	.624	.644	.632	.640	.617	.651	.647
COMEDI_russian_alpha	.641	.665	.598	.645	.656	.654	.640	.639
COMEDI_swedish_alpha	.729	.737	.688	.748	.709	.729	.697	.693
COMEDI_chinese_alpha	.466	.501	.412	.450	.465	.468	.474	.462
MCLWIC_ar-ar_alpha	.720	.735	.518	.568	.717	.711	.706	.696
MCLWIC_en-en_alpha	.820	.814	.590	.661	.817	.834	.821	.828
MCLWIC_fr-fr_alpha	.726	.733	.646	.624	.724	.740	.744	.760
MCLWIC_ru-ru_alpha	.699	.700	.648	.664	.705	.717	.743	.724
MCLWIC_zh-zh_alpha	.643	.641	.223	.276	.664	.646	.657	.613
WIC_en_alpha	.418	.416	.315	.387	.429	.407	.511	.523
COMEDI_mean_alpha	.668	.678	.648	.673	.671	.676	.660	.663
MCLWIC_mean_alpha	.722	.725	.525	.559	.725	.730	.734	.724

Cross-Architecture Synthesis — Best run α (selected by mean α)

	Reference (SOTA)	Phase 3 (Train on COMEDI)				Phase 4 (+ Auxiliary Data)			
	XL-DURel SOTA	Bi-Encoder (continuous)	Cross-Encoder (continuous)	Cross-Encoder (nominal)	**Cross-Encoder (CLM, ordinal)**	Bi-Encoder (continuous) + SweWiC	Cross-Encoder (continuous) + SweWiC	Cross-Encoder (nominal) + All Aux	Cross-Encoder (CLM, ordinal) + All Aux
COMEDI_german_alpha	.740	.782	.771	.728	.744	.781	.780	.753	.763
COMEDI_english_alpha	.730	.741	.742	.741	.730	.765	.757	.690	.712
COMEDI_spanish_alpha	.760	.724	.701	.736	.757	.725	.710	.746	.735
COMEDI_norwegian_alpha	.650	.635	.651	.671	.610	.620	.627	.645	.656
COMEDI_russian_alpha	.660	.644	.670	.607	.653	.643	.663	.637	.643
COMEDI_swedish_alpha	.690	.751	.755	.697	.780	.758	.760	.691	.701
COMEDI_chinese_alpha	.440	.459	.494	.414	.513	.474	.488	.485	.451
MCLWIC_ar-ar_alpha	.680	.710	.722	.503	.574	.713	.717	.694	.683
MCLWIC_en-en_alpha	.840	.828	.818	.560	.680	.818	.826	.845	.818
MCLWIC_fr-fr_alpha	.690	.732	.732	.642	.621	.732	.744	.760	.766
MCLWIC_ru-ru_alpha	.720	.706	.691	.638	.689	.710	.714	.753	.737
MCLWIC_zh-zh_alpha	.690	.647	.622	.226	.401	.688	.660	.653	.606
WIC_en_alpha	.430	.406	.414	.273	.390	.441	.410	.503	.515
COMEDI_mean_alpha	.670	.677	.683	.656	.684	.681	.684	.664	.666
MCLWIC_mean_alpha	.720	.724	.717	.514	.593	.732	.732	.741	.722

Detailed Results in Zero-Shot Scenario



Understanding Zero-Shot Results