

AAAI 2026 Supplementary Material
Anonymous Submission

Anonymous submission

Appendix A.1: Dataset Overview

Table 1 lists the sources and licenses for all datasets used in the BTZSC benchmark. All datasets are publicly available on Hugging Face Datasets¹. We use the same label verbalizers as used by (Laurer et al. 2023).

Domain	Dataset	Source	License
Emotion			
dialogue	empathetic_dialogues	(Rashkin et al. 2019)	CC BY-NC 4.0
social-media	emotiondair	(Saravia et al. 2018)	Other (research/education)
Intent			
banking	banking77	(Casanueva et al. 2020)	CC BY 4.0
social-media	biasframes_intent	(Sap et al. 2020)	CC BY 4.0
Sentiment			
apps	appreviews	(Grano et al. 2017)	Unknown
e-commerce	amazonpolarity	(Zhang, Zhao, and LeCun 2015a)	Apache-2.0
finance	financialphrasebank	(Malo et al. 2014)	CC BY-NC-SA 3.0
local-business	yelpreviews	(Zhang, Zhao, and LeCun 2015b)	Terms of Use (non-commercial, 21 Feb 2020)
movies	imdb	(Maas et al. 2011)	IMDb Non-Commercial Dataset Terms
movies	rottentomatoes	(Pang and Lee 2005)	CC0 1.0 (Public Domain)
Topic			
assistant	massive	(FitzGerald et al. 2022)	CC BY 4.0
education	trueteacher	(Gekhman et al. 2023)	CC BY-NC 4.0
news	agnews	(Zhang, Zhao, and LeCun 2015a)	Non-commercial (AG Corpus terms)
politics	capsotu	(Jones et al. 2023; Laurer et al. 2023)	CC BY-NC-SA 4.0
politics	manifesto	(Lehmann, Franzmann et al. 2024)	see Terms of Use
qa-forum	yahootopics	(Zhang, Zhao, and LeCun 2015c)	Unknown
social-media	biasframes_offensive	(Sap et al. 2020)	CC BY 4.0
social-media	biasframes_sex	(Sap et al. 2020)	CC BY 4.0
wikipedia	wikitoxic_insult	(Wulczyn, Thain, and Dixon 2017)	CC0 1.0
wikipedia	wikitoxic_obscene	(Wulczyn, Thain, and Dixon 2017)	CC0 1.0
wikipedia	wikitoxic_threat	(Wulczyn, Thain, and Dixon 2017)	CC0 1.0
wikipedia	wikitoxic_toxicaggregated	(Wulczyn, Thain, and Dixon 2017)	CC0 1.0

Table 1: Source and licenses for all BTZSC datasets.

¹<https://huggingface.co/datasets>

Appendix A.2: Model overview

Model	Yr	Arch.	Backbone	FT / train data [†]	# P	Pool / dim
Base encoders						
bert-large-uncased	2018	enc.	BERT	none	340 M	—
deberta-v3-large	2021	enc.	DeBERTa v3	none	304 M	—
ModernBERT-large	2024	enc.	ModernBERT	none	395 M	—
NLI cross-encoders						
bart-large-mnli	2020	enc-dec.	BART	SNLI, MNLI	406 M	—
nli-roberta-base	2020	enc.	RoBERTa	SNLI, MNLI	125 M	—
bert-base-uncased-nli	—	enc.	BERT	MNLI, ANLI, WANLI, FEVERNLI, LINGNLI	110 M	—
bert-large-uncased-nli	—	enc.	BERT	same as above	340 M	—
bert-large-uncased-nli-triplet	—	enc.	BERT	same as above	340 M	—
deberta-v3-base-nli	—	enc.	DeBERTa v3	same as above	184 M	—
deberta-v3-large-nli	—	enc.	DeBERTa v3	same as above	304 M	—
deberta-v3-large-nli-triplet	—	enc.	DeBERTa v3	same as above	304 M	—
modernbert-base-nli	—	enc.	ModernBERT	same as above	149 M	—
modernbert-large-nli	—	enc.	ModernBERT	same as above	395 M	—
modernbert-large-nli-triplet	—	enc.	ModernBERT	same as above	395 M	—
Rerankers						
ms-marco-MiniLM-L6-v2	2021	enc.	MiniLM	MS MARCO	22.7 M	—
gte-reranker-modernbert-base	2024	enc.	ModernBERT	large multiling. pairs	149 M	—
bge-reranker-base	2023	enc.	XLm-RoB. B	large multiling. pairs	278 M	—
bge-reranker-large	2023	enc.	XLm-RoB. L	large multiling. pairs	560 M	—
Qwen3-Reranker-0.6B	2025	dec.	Qwen3	synthetic yes/no ranking data	0.6 B	—
Qwen3-Reranker-8B	2025	dec.	Qwen3	synthetic yes/no ranking data	8 B	—
Embedding models						
all-MiniLM-L6-v2	2021	enc.	MiniLM	1 B paired sentences	22.7 M	mean / 384
e5-base-v2	2023	enc.	E5 (BERT)	270 M synthetic contrastive	110 M	mean / 768
e5-large-v2	2023	enc.	E5 (BERT)	same as above	335 M	mean / 1024
e5-mistral-7b-instruct	2024	dec.	Mistral-7B	synthetic multiling. contrastive	7 B	last / 4096
bge-base-en-v1.5	2023	enc.	BGE (RoB.)	1.5 B pair data, contrastive	137 M	CLS / 768
bge-large-en-v1.5	2023	enc.	BGE (RoB.)	same as above	434 M	CLS / 1024
gte-base-en-v1.5	2024	enc.+	GTE	MLM + contrastive pre-train	137 M	CLS / 768
gte-large-en-v1.5	2024	enc.+	GTE	same as above	434 M	CLS / 1024
gte-modernbert-base	2024	enc.	ModernBERT	same as above	149 M	CLS / 768
Qwen3-Embedding-0.6B	2025	dec.	Qwen3	synthetic multiling. contrastive	0.6 B	last / 1024
Qwen3-Embedding-8B	2025	dec.	Qwen3	synthetic multiling. contrastive	8 B	last / 4096

Table 2: Architectural and training overview of the 31 models evaluated. Columns list publication year (Yr), encoder/decoder architecture (Arch.), backbone, principal fine-tuning or pre-training data, parameter count (#P), and pooling strategy with embedding dimensionality.

Appendix A.3: Experimental Setup

Cross-Encoder Architectures for NLI

Let a paired input sequence (premise || hypothesis) be tokenised as $\mathbf{x} = (x_0 = [\text{CLS}], x_1, \dots, [\text{SEP}], \dots, x_{S-1})$ and encoded by a pre-trained Transformer backbone $f_\theta : \mathbb{N}^S \rightarrow \mathbb{R}^{S \times E}$ with hidden size E :

$$H = f_\theta(\mathbf{x}) \in \mathbb{R}^{S \times E}, \quad h = H_0 \in \mathbb{R}^E \quad (\text{CLS row}).$$

A two-layer classification head with dropout $p = 0.1$ transforms h :

$$\tilde{h} = \text{Dropout}_{0.1}(h), \quad (1)$$

$$u = \text{GELU}(W_1 \tilde{h} + b_1), \quad W_1 \in \mathbb{R}^{E \times E}, \quad b_1 \in \mathbb{R}^E, \quad (2)$$

$$z = \text{LayerNorm}(u), \quad (3)$$

$$\ell = W_2 z + b_2, \quad W_2 \in \mathbb{R}^{E \times C}, \quad b_2 \in \mathbb{R}^C, \quad (4)$$

where C is the number of label logits returned by the head. This setup mimicks the standard classification head of (Warner et al. 2024)

Binary variant. Here $C = 1$ and $\ell \in \mathbb{R}$ is an *entailment logit*. The probability of entailment is $\sigma(\ell) = (1 + e^{-\ell})^{-1}$ and the model is optimised with binary-cross-entropy:

$$\mathcal{L}_{\text{BCE}}(y, \ell) = -y \log \sigma(\ell) - (1 - y) \log (1 - \sigma(\ell)), \quad y \in \{0, 1\}.$$

Three-way variant (triplet). Now $C = 3$ with logits $\ell = (\ell_{\text{ent}}, \ell_{\text{neut}}, \ell_{\text{contra}})$. During *training* the standard multi-class cross-entropy is used:

$$\mathcal{L}_{\text{CE}}(y, \ell) = -\log \frac{\exp(\ell_y)}{\sum_{c=1}^3 \exp(\ell_c)}, \quad y \in \{1, 2, 3\}.$$

During *evaluation* the scalar entailment score is

$$s = \ell_{\text{ent}} - \log(e^{\ell_{\text{neut}}} + e^{\ell_{\text{contra}}}),$$

which is the log-odds of the ENTAILMENT class versus the union of the other two classes. The corresponding probability is $\sigma(s)$.

Dimensions.

- B : batch size (omitted above for clarity),
- S : sequence length,
- E : hidden size of the backbone,
- $C \in \{1, 3\}$: number of logits.

Training Procedure

Validation signal. Early stopping is triggered by the dev-set loss computed on an *equal-sized, balanced* union

$$\mathcal{D}_{\text{dev}} = \text{MNLI}_{\text{m}} \cup \text{MNLI}_{\text{mm}} \cup \text{ANLI}_{r1} \cup \text{ANLI}_{r2} \cup \text{ANLI}_{r3} \cup \text{WANLI} \cup \text{FEVERNLI} \cup \text{LINGNLI}.$$

At every evaluation step the loss is measured, and training stops when this loss fails to decrease for 10 consecutive evaluations, or 3 epochs, whichever comes first. Evaluation is performed every 1% of total steps.

Optimiser and schedules. Fine-tuning uses the PyTorch AdamW optimiser with default settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight-decay = 0.01). The learning rate employs a *linear warm-up* for the first 10% of steps followed by *cosine decay*:

$$\eta_t = \begin{cases} \eta_0 \frac{t}{0.1T}, & 0 \leq t < 0.1T, \\ \frac{1}{2} \eta_0 \left(1 + \cos \frac{\pi(t - 0.1T)}{0.9T} \right), & 0.1T \leq t \leq T, \end{cases}$$

with separate initial rates for the backbone (η_{enc}) and classification head (η_{head}).

- **Large backbones:** $\eta_{\text{enc}} = 8 \times 10^{-6}$, $\eta_{\text{head}} = 4 \times 10^{-5}$.
- **Base backbones:** $\eta_{\text{enc}} = 2 \times 10^{-5}$, $\eta_{\text{head}} = 1 \times 10^{-4}$.

All models train for $E = 3$ epochs with mini-batch size $B = 32$ and no layer freezing.

Qwen3 Reranker

Prompt Template For every *query–document* pair we build a single decoder-only prompt of the form

$$P = \text{prefix} + \langle \text{Instruct} \rangle: I + \langle \text{Query} \rangle: q + \langle \text{Document} \rangle: d + \text{suffix}.$$

Fixed strings.

- **Prefix**

```
<|im_start|>system
Judge whether the Document meets the requirements based on the Query
and the Instruct provided. Note that the answer can only be "yes" or "no".
<|im_end|>
<|im_start|>user
```

- **Suffix**

```
<|im_end|>
<|im_start|>assistant
<think>

</think>
```

Instructions I.

- **NLI retrieval**

Given a piece of text, retrieve the passage that entails the text the best

- **Label retrieval**

Given a piece of text, retrieve relevant label descriptions that best match the text

Binary decision via “yes/no” tokens Let τ_{yes} and τ_{no} be the token IDs that realise the strings “yes” and “no”. Denote the final-step logit vector by $v = L_{S-1} \in \mathbb{R}^V$. We extract

$$v_{\tau_{\text{yes}}}, v_{\tau_{\text{no}}}$$

and compute the entailment probability as

$$p_{\text{yes}} = \frac{e^{v_{\tau_{\text{yes}}}}}{e^{v_{\tau_{\text{yes}}}} + e^{v_{\tau_{\text{no}}}}}$$

References

- Casanueva, I.; Temcinas, T.; Gerz, D.; Henderson, M.; and Vulić, I. 2020. Efficient Intent Detection with Dual Sentence Encoders. In *Proceedings of the 28th International Conference on Computational Linguistics*.
- FitzGerald, J.; et al. 2022. MASSIVE: A 1M-Example Multilingual NLU Dataset with 51 Typologically-Diverse Languages. In *EMNLP*.
- Gekhman, Z.; et al. 2023. TrueTeacher: Learning Factual Consistency Evaluation with Large Language Models. In *EMNLP*.
- Grano, G.; Sorbo, A. D.; Mercaldo, F.; Visaggio, C. A.; Canfora, G.; and Panichella, S. 2017. Android Apps and User Feedback: A Dataset for Software Evolution and Quality Improvement. In *Proceedings of the 2nd ACM SIGSOFT International Workshop on App Market Analytics (WAMA)*, 8–11.
- Jones, B. D.; Baumgartner, F. R.; Theriault, S. M.; Epp, D. A.; Eissler, R.; Lee, C.; and Sullivan, M. E. 2023. Policy Agendas Project: State of the Union Speeches. <https://www.comparativeagendas.net/>. Comparative Agendas Project dataset.
- Laurer, M.; van Atteveldt, W.; Casas, A.; and Welbers, K. 2023. Building Efficient Universal Classifiers with Natural Language Inference. ArXiv:2312.17543 [cs].
- Lehmann, P.; Franzmann, S.; et al. 2024. The Manifesto Data Collection, Version 2024a. In *WZB Data Release*.
- Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *ACL*.
- Malo, P.; et al. 2014. Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of Emerging Technologies in Accounting*.
- Pang, B.; and Lee, L. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *ACL*.
- Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*.

- Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N. A.; and Choi, Y. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *ACL*.
- Saravia, E.; Liu, H.-C. T.; Huang, Y.-H.; Wu, J.; and Chen, Y.-S. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of EMNLP 2018*, 3687–3697. Brussels, Belgium.
- Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; Cooper, N.; Adams, G.; Howard, J.; and Poli, I. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv:2412.13663*.
- Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex Machina: Personal Attacks Seen at Scale. In *WWW*.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015a. Character-level Convolutional Networks for Text Classification. In *NIPS*.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015b. Character-level Convolutional Networks for Text Classification. In (Zhang, Zhao, and LeCun 2015a). Yelp Reviews Polarity subset.
- Zhang, X.; Zhao, J.; and LeCun, Y. 2015c. Character-level Convolutional Networks for Text Classification. In (Zhang, Zhao, and LeCun 2015a). Yahoo Answers Topics subset.