Dual-Head Reasoning Distillation: Improving Classifier Accuracy with Train-Time-Only Reasoning

Jillian Xu[*]
University of Waterloo
j23xu@uwaterloo.ca

Dylan Zhou
Google
dylanzhou@google.com

Vinay Shukla
Google
vinayshukla@google.com

Yang Yang
Google
lizyang@google.com

Junrui Ruan
Google
junrui@google.com

Shuhuai Lin
Google
shuhuailin@google.com

Wenfei Zou
Google
wenfei@google.com

Yinxiao Liu
Google DeepMind
canoee@google.com

Karthik Lakshmanan
Google
lakshmanan@google.com

Abstract

Chain-of-Thought (CoT) prompting often improves classification accuracy but it introduces a significant throughput penalty with rationale generation (Wei et al., 2022; Cheng and Van Durme, 2024). To resolve this trade-off, we introduce *Dual-Head Reasoning Distillation* (DHRD), a simple training method for decoder-only language models (LMs) that adds (i) a pooled *classification head* used during training and inference and (ii) a *reasoning head* supervised by teacher rationales used only in training. We train with a loss function that is a weighted sum of label cross-entropy and token-level LM loss over input-plus-rationale sequences. On seven SuperGLUE tasks, DHRD yields relative gains of 0.65–5.47% over pooled baselines, with notably larger gains on entailment/causal tasks. Since we disable the reasoning head at test time, inference throughput matches pooled classifiers and exceeds CoT decoding on the same backbones by 96–142$\times$ in QPS.

1 Introduction

Decoder-only LMs are widely adapted to encoder-style classification tasks by pooling hidden states and applying a lightweight classifier (e.g. `AutoModelForSequenceClassification` from HuggingFace (Wolf et al., 2020)). While effective, this adaptation can under-utilize the model's latent reasoning abilities learned in pre-training. CoT can surface that capacity, but its token-by-token decoding can significantly reduce throughput, making it impractical for high-throughput applications (Wei et al., 2022; Cheng and Van Durme, 2024).

We propose *Dual-Head Reasoning Distillation* (DHRD), a simple method that requires just two small additions to the backbone decoder: (i) a classification head that pools token representations with last-token pooling, and (ii) a reasoning tower/LM head used only for auxiliary training loss. We optimize a weighted joint objective that balances label cross-entropy and auxiliary token-level cross-entropy; a CoT-capable teacher (we use *Gemini 2.5 Flash* (Gemini Team, Google DeepMind, 2025)) prompted with Zero-shot Chain-of-Thought ('Let's think step by step'; (Kojima et al., 2022)) provides rationales for training (for prompt refer to Appendix E.3). We evaluate on SuperGLUE tasks

[*] Work performed while working at Google

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

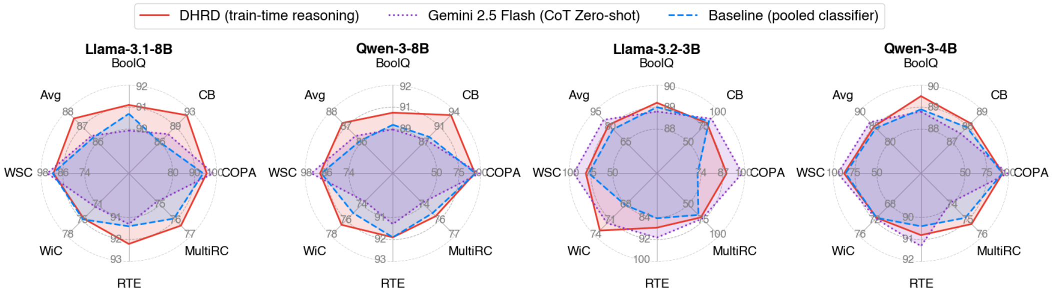

Figure 1: SuperGLUE per-task scores for four backbones. DHRD (train-time reasoning) consistently beats the pooled-classifier baseline and rivals teacher model *Gemini 2.5 Flash*, with the largest gains on CB/COPA/RTE. ‘Avg’ is the macro-average, tabulated results can be found in Table 1.

under standard model input formatting and metrics (refer to Appendix E). Ablations demonstrate that improvements are attributable to alignment of input–rationale–label triplets rather than to generic LM regularization; intentional misalignment leads to substantial drops in performance. By *moving reasoning to train time* and deploying a CoT-free classifier at test time, we show that DHRD preserves baseline latency while capturing much of CoT’s quality benefits for efficient reasoning.

2 Related Work

Adapting decoder-only LMs for encoder-style classification. Suganthan et al. introduce *Gemma Encoder*, which adapts a decoder-only LM into an encoder by enabling bidirectional attention and adding task-specific pooling plus a Multilayer Perceptron (MLP) head; their analysis finds that simple pooling methods (last-token or mean) with an MLP head perform competitively with attention pooling. DHRD augments the pooling-MLP setup with a *train-only* reasoning head so the model remains a pooled classifier at test time (no CoT generation). DHRD deliberately retains *causal* attention and *last-token* pooling to prevent future-token leakage and match autoregressive pretraining.

Relation to Rationale Supervision and Knowledge Distillation (KD). KD traditionally transfers knowledge from a teacher to a student model via soft-label or logit matching (Hinton et al., 2015), and sequence-level KD is widely used in sequence-to-sequence settings (Kim and Rush, 2016). For LLMs, recent work explores distilling *reasoning traces* or explanations (e.g., e-SNLI Camburu et al. 2018). DHRD is a form of *reasoning-aware distillation*: we supervise the student with teacher-provided rationales, but *only at train time*, and we do *not* rely on answer logit matching or inference-time rationale generation. At test time, DHRD uses a pooled classification head (no CoT decoding).

Auxiliary Language-Modeling (LM) Loss. Adding auxiliary LM loss during supervised fine-tuning can improve generalization and accelerate convergence Radford et al. (2018). For our training, we use an auxiliary LM loss comprised of the input classification target plus rationale tokens produced by a CoT-capable teacher model. Our ablations indicate that the lift stems from *aligned* input–rationale–label triplets rather than generic LM regularization.

3 Method

3.1 Dual-Head Architecture

We attach two lightweight heads to a shared decoder-only transformer with causal attention, enabling a single forward pass that supports both classification and train-time reasoning.

Training-time input and slicing For a sample i, we form a single sequence by concatenating the original classification input x_i with a teacher rationale r_i (Gemini 2.5 Flash) and the gold label y_i:

$$s_i = [\, x_i,\ \texttt{<REASON>},\ r_i,\ \texttt{<ANS>},\ y_i \,], \qquad L_i = |s_i|, \qquad L_i^{(x)} = |x_i|. \tag{1}$$

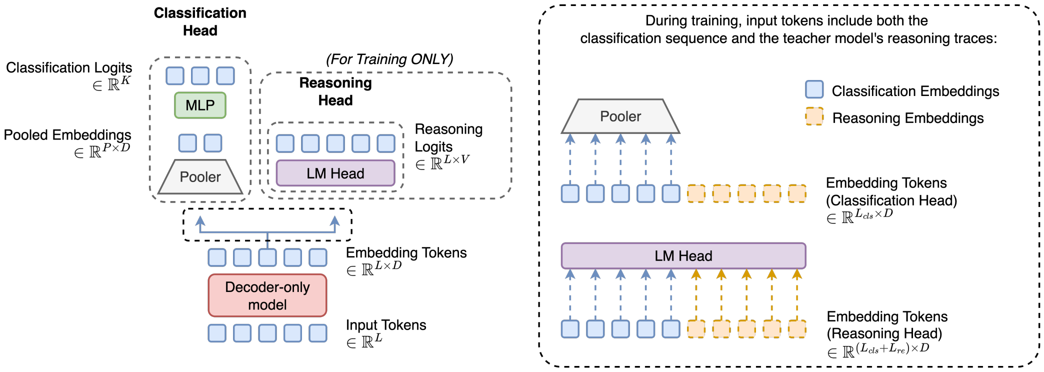
Classification Head
Classification Logits $\in \mathbb{R}^K$
MLP
Pooled Embeddings $\in \mathbb{R}^{P\times D}$
Pooler
(For Training ONLY)
Reasoning Head
Reasoning Logits $\in \mathbb{R}^{L\times V}$
LM Head
Embedding Tokens $\in \mathbb{R}^{L\times D}$
Decoder-only model
Input Tokens $\in \mathbb{R}^{L}$
During training, input tokens include both the classification sequence and the teacher model's reasoning traces:
Pooler
Classification Embeddings
Reasoning Embeddings
Embedding Tokens (Classification Head) $\in \mathbb{R}^{L_{cls}\times D}$
LM Head
Embedding Tokens (Reasoning Head) $\in \mathbb{R}^{(L_{cls}+L_{re})\times D}$

Figure 2: Dual-head fine-tuning on a shared decoder. The *classification head* pools hidden states over the input span (blue) to produce K class logits. The train-only *reasoning head* applies a causal LM loss over the full sequence, covering both classification input tokens (blue) and teacher rationale tokens (orange). During training, inputs concatenate task text with teacher rationales. At inference, only the classification head is used.

Classification head. We pool the hidden state of the last real input token (before `<REASON>`) and map to K logits (for K classes) via a 2-layer Multi-layer Perceptron (MLP) following the adaptation proposed in Suganthan et al. (2025). This is the only head used for prediction at test time.

Reasoning (LM) head, train-time only. We reuse the base model's LM head to obtain next-token distributions; the standard causal LM loss is computed over the entire sequence s_i.

3.2 Weighted Objective

Let batch size be B, $\mathbf{z}_i \in \mathbb{R}^K$ be class logits for example i, and $y_i \in \{1, \dots, K\}$ the class label. The classification loss (explicit log-softmax form) is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\Big(\mathbf{z}_i[y_i] - \log\sum_{k=1}^{K} e^{\mathbf{z}_i[k]}\Big). \tag{2}$$

For reasoning (causal LM), let the vocabulary size be V, batch size B, sequence lengths L_i, token logits $\ell_{i,t} \in \mathbb{R}^V$ at position t, and targets $v_{i,t} \in \{1, \dots, V\}$ and a binary mask $m_{i,t} \in \{0, 1\}$ indicating whether position t is valid (so the loss uses $m_{i,t+1}$ to mask the next-token target). The reasoning loss is standard causal-LM (next-token) cross-entropy over the entire sequence:

$$\mathcal{L}_{\text{reason}} = -\frac{1}{N}\sum_{i=1}^{B}\sum_{t=1}^{L_i-1} m_{i,t+1}\,\log\left(\frac{\exp\{\ell_{i,t}[\,v_{i,t+1}\,]\}}{\sum_{w=1}^{V}\exp\{\ell_{i,t}[w]\}}\right), \qquad N = \sum_{i=1}^{B}\sum_{t=1}^{L_i-1} m_{i,t+1}. \tag{3}$$

The joint objective is

$$\mathcal{L}_{\text{total}} = \beta\,\mathcal{L}_{\text{cls}} + \alpha\,\mathcal{L}_{\text{reason}}, \quad \alpha, \beta \ge 0. \tag{4}$$

At inference, we input only x, ignore the reasoning head, and produce $\mathbf{z}$ from pooled encoder states.

4 Evaluation

4.1 Datasets

We evaluate on seven SuperGLUE tasks (Wang et al., 2019): BoolQ (Clark et al., 2019), CB (de Marneffe et al., 2019), COPA (Roemmele et al., 2011), MultiRC (Khashabi et al., 2018), RTE (Dagan et al., 2006), WiC (Pilehvar and Camacho-Collados, 2019), and WSC (Levesque et al., 2012). ReCoRD is excluded as an extractive QA task. Following common practice, CB is reported as F1/Accuracy and MultiRC as F1a/EM; Avg is a macro-average across the seven tasks after first averaging the two metrics for CB and MultiRC. A brief summary of task formats and dataset sizes appears in Appendix C.

Table 1: SuperGLUE results (higher is better). *Rel.* Δ *(%)* is the relative percentage change versus the pooled baseline for the same backbone (α=0, β=1). All DHRD rows use the optimal weights selected on the validation split: α=β=1 for Llama-3.1-8B, Llama-3.2-3B, and Qwen-3-8B; α=0.5, β=1 for Qwen-3-4B. Full α/β ablations can be found in Appendix B.

Model	Setting	BoolQ	CB	COPA	MultiRC	RTE	WiC	WSC	Avg	Rel. Δ (%)
Gemini 2.5 Flash	CoT Zero-shot	88.8	83.1/92.0	98.4	86.2/59.1	91.3	71.6	94.5	86.40	
Llama-3.1-8B	Baseline (pooled classifier)	90.7	81.9/90	93.8	88.9/62.9	91.4	**74.8**	**91.5**	86.29	
	DHRD (train-time reasoning)	**91.1**	**89.0/94.8**	**95.4**	**89.0/63.7**	**92.2**	74.6	91.1	**87.52**	**+1.43%**
Qwen-3-8B	Baseline (pooled classifier)	89.4	83.5/93.2	**93.8**	88.2/62.0	91.9	75.1	89.0	86.09	
	DHRD (train-time reasoning)	**90.5**	**90.8/95.6**	93.2	**88.5/62.5**	**91.9**	**76.6**	**89.7**	**87.23**	**+1.32%**
Llama-3.2-3B	Baseline (pooled classifier)	89.0	**78.6/88.8**	72.2	83.5/49.8	84.3	68.1	77.4	77.34	
	DHRD (train-time reasoning)	**89.2**	73.8/87.2	**89.2**	**85.9/55.3**	**87.7**	**73.0**	**80.8**	**81.57**	**+5.47%**
Qwen-3-4B	Baseline (pooled classifier)	88.9	84.3/91.6	92.4	87.7/62.0	90.4	71.4	85.6	84.50	
	DHRD (train-time reasoning)	**89.5**	**84.4/92.0**	**92.8**	**88.1/62.4**	**90.8**	**71.4**	**87.4**	**85.05**	**+0.65%**

SuperGLUE Results. Table 1 shows that the 8B models perform best with (β=α=1); Llama-3.2-3B sees the largest relative lift (+5.47%), while Qwen3-4B prefers a lower α=0.5. Both 8B DHRD settings exceed the teacher model's (Gemini 2.5 Flash) average SuperGLUE score.

Train-time reasoning improves accuracy most on *entailment and cause–effect* style tasks (CB, RTE, COPA), both Llama and Qwen 8B improved by +8.7% accuracy on CB, and Llama 3B improved by +23.5% accuracy on COPA; consistent improvements are observed in RTE, MultiRC, and BoolQ. WiC/WSC are relatively stable, likely because word-sense and coreference tasks are less reliant on multi-step verbalization.

4.2 Inference Efficiency (no CoT vs. CoT)

Since DHRD disables rationale generation at test time, the relevant throughput is the pooled *classification* head's forward pass (no decoding). Appendix D shows that DHRD preserves pooled-classifier throughput and *avoids* CoT decoding costs. For context, we also report the throughput of CoT-style decoding on the same backbones (for settings refer to Appendix E.6), which shows that the pooled path is (96–142×) faster than CoT-style inference on the same decoder-only backbones.

4.3 Ablations: Reasoning

Table 2: Ablations on rationale/label alignment (SuperGLUE). ConsistentReasoningLabel (aligned `<REASON>` and `<ANS>`), OnlyLabel (aligned `<ANS>`), ShuffleReasoning (misaligned `<REASON>`, aligned `<ANS>`), ShuffleReasoningLabel (misaligned `<REASON>` and `<ANS>`). In all settings, the LM loss is applied to the input tokens and present `<REASON>`/`<ANS>` segments. All rows use α=β=1.

Model	DHRD Setting	BoolQ	CB	COPA	MultiRC	RTE	WiC	WSC	Avg	Rel. Δ (%)
Llama-3.1-8B	ConsistentReasoningLabel	**91.1**	**89.0/94.8**	**95.4**	**89.0/63.7**	**92.2**	74.6	**91.1**	**87.52**	
	OnlyLabel	90.7	87.6/93.6	93.8	84.6/53.2	91.1	**76.4**	90.4	85.99	-1.75%
	ShuffleReasoning	90.4	59.5/76.4	95.2	84.9/53.1	90.9	75.9	88.4	82.54	-5.70%
	ShuffleReasoningLabel	62.2	35.4/51.6	44.6	0.0/0.5	50.0	50.0	65.1	45.09	-48.48%
Llama-3.2-3B	ConsistentReasoningLabel	**89.2**	**73.8/87.2**	89.2	**85.9/55.3**	**87.7**	73.0	**80.8**	**81.57**	
	OnlyLabel	88.0	54.5/79.2	**91.4**	84.5/51.3	84.3	**75.3**	75.34	78.44	-3.84%
	ShuffleReasoning	81.4	57.8/76.0	90.2	11.2/0.6	87.3	68.6	55.12	65.06	-20.24%
	ShuffleReasoningLabel	62.3	21.7/48.4	50.0	0.0/0.5	50.1	49.7	65.1	44.64	-45.27%

Ablation Results. Table 2 compares our best setting (ConsistentReasoningLabel) to three controls that remove or corrupt the explanation signal. The results show removing rationales hurts (OnlyLabel), while misalignment degrades performance, especially on entailment/causal tasks (CB, RTE, COPA).

5 Conclusion

Dual-Head Reasoning Distillation (DHRD) is a reasoning-classification joint-training technique that moves the computational cost of reasoning to training time, improving accuracy without inference overhead. Across seven SuperGLUE tasks and 3–8B decoder-only backbones, DHRD improves pooled baselines by **0.65–5.47%** (macro-avg) with the largest gains on entailment/causal tasks (CB, RTE, COPA). With no rationale generation at inference time, DHRD preserves the pooled classifier's latency and achieves **96–142**× higher throughput than CoT-style decoding on the same backbones. Ablations show the lift comes from *aligned* input–rationale–label triplets rather than generic LM regularization: removing or misaligning rationales degrades accuracy, especially on entailment. In practice, we find $\alpha=1$ works well for 8B models while Qwen 4B model prefer a milder LM weight ($\alpha\approx0.5$). Limitations include dependence on the quality of rationale and sensitivity to the LM-loss weight; we also report single runs per setting due to computational constraints (refer to Appendix A).

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. Promptsource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, 2022. doi: 10.18653/v1/2022.acl-demo.9. URL https://aclanthology.org/2022.acl-demo.9/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *arXiv preprint*, 2018. URL https://arxiv.org/abs/1812.01193.

Samuel Carton, Surya Kanoria, and Chenhao Tan. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.86.pdf.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. Mcc-kd: Multi-cot consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.454. URL https://aclanthology.org/2023.findings-emnlp.454/.

Jeffrey Cheng and Benjamin Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*, 2024. URL https://arxiv.org/abs/2412.13171.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, 2019. URL https://aclanthology.org/N19-1300.pdf.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, ECML 2005, LNCS 3944*, pages 177–190. Springer, 2006. URL https://link.springer.com/chapter/10.1007/11736790_9.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*, 2019. URL https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/601. Volume 2, pp. 107–124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Gemini Team, Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Hugo Touvron, Laurens van der Maaten, et al. The llama 3 herd of models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Avelina Asada Hadji-Kyriacou and Ognjen Arandjelovic. Would i lie to you? inference time alignment of language models using direct preference heads. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2405.20053.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint*, 2016. doi: 10.48550/arXiv.1606.08415. URL https://arxiv.org/abs/1606.08415.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/pdf?id=nZeVKeeFYf9.

Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. C3ot: Generating shorter chain-of-thought without compromising effectiveness. *arXiv preprint arXiv:2412.11664*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint*, 2020. doi: 10.48550/arXiv.2001.08361. URL https://arxiv.org/abs/2001.08361.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of NAACL-HLT*, 2018. URL https://aclanthology.org/N18-1023.pdf.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Øyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, 2020. doi: 10.18653/v1/2020.findings-emnlp.171. URL https://aclanthology.org/2020.findings-emnlp.171/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of EMNLP*, 2016. URL https://aclanthology.org/D16-1139/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.771. URL https://aclanthology.org/2020.acl-main.771/.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the 13th International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, 2012. ISBN 978-1-57735-560-1. URL https://cdn.aaai.org/ocs/4492/4492-21843-1-PB.pdf.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Xinglin Wang, Bin Sun, Heda Wang, and Kan Li. Escape sky-high cost: Early-stopping self-consistency for multi-step reasoning. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.10480.

Tianqiao Liu, Zui Chen, Zitao Liu, Mi Tian, and Weiqi Luo. Expediting and elevating large language model reasoning via hidden chain-of-thought decoding. *arXiv preprint arXiv:2409.08561*, 2024.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

Mohammad Taher Pilehvar and Jose Camacho-Collados. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of NAACL-HLT*, 2019. URL https://aclanthology.org/N19-1128/.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of AAAI*, 2011. URL https://ict.usc.edu/pubs/Choice%20of%20Plausible%20Alternatives-%20An%20Evaluation%20of%20Commonsense%20Causal%20Reasoning.pdf.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Teven Le Scao and Alexander M. Rush. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, 2021. doi: 10.18653/v1/2021.naacl-main.208. URL https://aclanthology.org/2021.naacl-main.208/.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.441. URL https://aclanthology.org/2023.findings-acl.441/.

Vinay Shukla, Yang Yang, Siddarth Malreddy, Jinoo Baek, Dale Johnson, Wenfei Zou, Karthik Lakshmanan, Mark Williams, and Minh Pham. Embedding user-generated content using structural supervision and generative models. In *Proceedings of the Third Workshop on Efficient Natural Language and Speech Processing (ENLSP-III) at NeurIPS 2023*, 2023. URL https://neurips2023-enlsp.github.io/papers/paper_9.pdf.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2408.03314. To appear at ICLR 2025.

Paul Suganthan, Fedor Moiseev, Le Yan, Junru Wu, Jianmo Ni, Jay Han, Imed Zitouni, Enrique Alfonseca, Xuanhui Wang, and Zhe Dong. Adapting decoder-based language models for diverse encoder downstream tasks. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2503.02656. URL `https://arxiv.org/abs/2503.02656`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL `https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf`.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL `https://papers.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf`.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. URL `https://aclanthology.org/2023.acl-long.147.pdf`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2203.11171. URL `https://arxiv.org/abs/2203.11171`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2201.11903. URL `https://arxiv.org/abs/2201.11903`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. URL `https://aclanthology.org/2020.emnlp-demos.6/`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2505.09388. URL `https://arxiv.org/abs/2505.09388`.

Penghui Yang, Chen-Chen Zong, Sheng-Jun Huang, Lei Feng, and Bo An. Dual-head knowledge distillation: Enhancing logits utilization with an auxiliary head. *arXiv preprint*, 2024. URL `https://arxiv.org/abs/2411.08937`.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2305.10601. URL `https://arxiv.org/abs/2305.10601`.

Zijian Zhang, Koustav Rudra, and Avishek Anand. Explain and predict, and then predict again. *arXiv preprint*, 2021. URL `https://arxiv.org/abs/2101.04109`.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2205.10625. URL `https://arxiv.org/abs/2205.10625`.

Chiwei Zhu, Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, and Zhendong Mao. Rationales are not silver bullets: Measuring the impact of rationales on model performance and reliability. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5808–5835, Vienna, Austria, 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.findings-acl.302.pdf`.

A Broader Impact and Limitations

Broader impact. *Dual-Head Reasoning Distillation* (DHRD) retains the accuracy benefits of chain-of-thought supervision while restoring pooled-classifier throughput by removing inference-time rationales, lowering latency and compute. We observe the largest gains on entailment and cause–and–effect tasks, which can strengthen knowledge-intensive QA by verifying whether evidence *supports* or *contradicts* an answer and improve safety moderation by checking semantic *entailment* of policy violations rather than keywords.

Risks include reduced transparency and contestability when rationales are not emitted and amplification of biases under distribution shift. There is also a possibility of false positives/negatives in moderation or QA, potential misuse for high-throughput surveillance or censorship, and rebound effects from scaling automated decisions. To mitigate these, we commit to detailed model cards with per-group performance and calibration with human-in-the-loop review and appeal mechanisms in high-stakes settings. We avoid training on sensitive personal information and consider staged/gated release with usage policies.

Limitations. Observed gains depend on high-quality, label-consistent teacher rationales; removing or misaligning rationales substantially harms accuracy, particularly on entailment/causal tasks. Performance is sensitive to the auxiliary LM-loss weight α: on 4B backbones, large α can regress accuracy (moderate values, e.g., $\alpha{\approx}0.5$, are safer), whereas 8B models tolerated $\alpha{=}1$; more systematic schedules are left to future work. Our evaluation covers seven SuperGLUE tasks (excluding ReCoRD) and 3–8B decoder-only backbones with last-token pooling; generalization to other languages, modalities, domains, and architectures remains untested.

We have not yet conducted a direct, budget-controlled comparison to alternative *CoT compression* methods such as HiddenCoT, C3oT, and CCoT, which target reduced rationale tokens at training and/or inference (Liu et al., 2024; Kang et al., 2024; Cheng and Van Durme, 2024). As a result, it remains unclear how DHRD trades off accuracy, calibration, and latency relative to these approaches.

Due to compute constraints, we report single runs without confidence intervals; we partially mitigate this by reporting per-task metrics and macro-averages, showing trends across models and α values, and specifying exact training/eval configurations to aid replication (App. E.5).

B Loss Weighting Ablations

Table 3: SuperGLUE results (higher is better). Rel. Δ (%) is the relative percentage change compared to each model's pooled baseline ($\alpha{=}0, \beta{=}1$). Reasoning / CoT fine-tuned models follows ($\alpha{=}1, \beta{=}0$) with CoT at inference using the Reasoning Head (refer to Appendix E.6).

Model	Setting	BoolQ	CB	COPA	MultiRC	RTE	WiC	WSC	Avg	Rel. Δ (%)
Gemini 2.5 Flash	CoT Zero-shot	88.8	83.1/92.0	98.4	86.2/59.1	91.3	71.6	94.5	86.40	
Llama-3.1-8B	Reasoning / CoT	80.6	53.5/77.6	91.8	71.5/27.4	71.7	66.3	76.12	71.65	
Qwen3-8B	Reasoning / CoT	81.8	63.5/70.4	92.6	81.4/44.7	90.9	74.3	77.4	78.14	
Llama-3.1-8B	Baseline	90.7	81.9/90	93.8	88.9/62.9	91.4	74.8	91.5	86.29	
	DHRD ($\beta{=}1, \alpha{=}0.5$)	91.0	**89.1/94.8**	94.8	88.5/63.3	**92.3**	72.5	**92.5**	87.28	+1.15%
	DHRD ($\beta{=}1, \alpha{=}1$)	**91.1**	89.0/94.8	**95.4**	89.0/63.7	92.2	74.6	91.1	**87.52**	**+1.43%**
	DHRD ($\beta{=}0.5, \alpha{=}1$)	90.7	85.6/92.8	94.4	**89.1/63.8**	91.1	**74.8**	88.4	86.44	+0.17%
Qwen3-8B	Baseline	89.4	83.5/93.2	93.8	88.2/62.0	91.9	75.1	89.0	86.09	
	DHRD ($\beta{=}1, \alpha{=}0.5$)	89.7	**91.8/95.6**	92.8	88.2/62.0	91.8	76.4	88.4	86.84	+0.87%
	DHRD ($\beta{=}1, \alpha{=}1$)	90.5	90.8/95.6	93.2	**88.5/62.5**	**91.9**	76.6	**89.7**	**87.23**	**+1.32%**
	DHRD ($\beta{=}0.5, \alpha{=}1$)	**90.6**	89.3/95.2	**94.0**	87.5/60.1	91.7	**77.9**	87.7	86.85	+0.88%
Llama-3.2-3B	Baseline	89.0	**78.6/88.8**	72.2	83.5/49.8	84.3	68.1	77.4	77.34	
	DHRD ($\beta{=}1, \alpha{=}0.5$)	89.1	52.4/76.4	89.6	86.4/57.7	**87.8**	72.5	**84.2**	79.95	+3.37%
	DHRD ($\beta{=}1, \alpha{=}1$)	**89.2**	73.8/87.2	89.2	85.9/55.3	87.7	**73.0**	80.8	**81.57**	**+5.47%**
	DHRD ($\beta{=}0.5, \alpha{=}1$)	88.3	62.6/78.4	**90.8**	**87.2/59.1**	86.8	56.6	77.4	77.65	+0.40%
Qwen3-4B	Baseline	88.9	84.3/91.6	92.4	87.7/62.0	90.4	71.4	85.6	84.50	
	DHRD ($\beta{=}1, \alpha{=}0.5$)	89.5	**84.4/92.0**	92.8	**88.1/62.4**	90.8	71.4	**87.4**	**85.05**	**+0.65%**
	DHRD ($\beta{=}1, \alpha{=}1$)	**89.6**	73.0/82.4	**93.2**	87.9/60.5	**91.2**	71.8	87.6	83.61	-1.05%
	DHRD ($\beta{=}0.5, \alpha{=}1$)	88.5	68.2/88.0	91.4	87.5/60.1	90.1	**72.6**	84.2	82.67	-2.17%

C SuperGLUE Dataset Overview

Table 4: SuperGLUE benchmark overview with task type, train, validation, and test dataset sizes

Benchmark (SuperGLUE)	Task	Metric	Train	Validation	Test
BoolQ	QA	Accuracy	9427	3270	3245
CB	NLI	F1/Accuracy	250	56	250
COPA	QA	Accuracy	400	100	500
MultiRC	QA	F1/EM	27243	4848	9693
RTE	NLI	Accuracy	2490	277	3000
WiC	WSD	Accuracy	5428	638	1400
WSC	Coref.	Accuracy	554	104	146

D QPS Results on same decoder backbones (no CoT vs. CoT)

Table 5: Throughput (queries per second, higher is better). Classification uses a pooled head at inference (no decoding). Reasoning uses CoT-style decoding (train-time only in DHRD). The rightmost column shows the speedup of our deployed path over CoT decoding on the same backbone.

Model	QPS (Pooled Classifer)	QPS (Reasoning / CoT)	Speedup (no-CoT / CoT)
Llama-3.1-8B	414.91	4.14	$\sim$100×
Llama-3.2-3B	631.16	4.44	$\sim$142×
Qwen3-8B	341.15	3.56	$\sim$96×
Qwen3-4B	440.83	3.76	$\sim$117×

E Fine-tuning Details

E.1 Hardware

All experiments were run with distributed data parallel training on 8× H100 (80GB) GPUs using NCCL. Unless noted otherwise, per-device train/eval batch size was 1 with gradient accumulation to reach effective batch sizes (see §E.5). Mixed precision used `bf16`.

E.2 Prompt Templates (UnifiedQA-style)

We adopt a single *UnifiedQA-style* prompt across tasks (Khashabi et al., 2020), phrased to normalize inputs to a QA format. The student's **classification input** is:

```
Passage: {paragraph}
Question: {question}
Answer: {answer}
Is the answer correct Yes or No?
```

During training, we append a train-time only **reasoning suffix** (teacher rationale r and gold label y) so the reasoning head can compute token-level likelihood:

```
Reasoning: {explanation}
Final Answer: {label}
```

At inference, only the classification head is used; no rationale is generated.

Task mapping. We convert each SuperGLUE task to the same schema:

- **MultiRC**: `paragraph`→ passage, `question`→ question, candidate `answer`→ answer; label is Yes/No for correctness.
- **BoolQ**: passage is the Wikipedia paragraph; the Boolean `answer` becomes the candidate; label is whether “Yes” is correct.

- **CB/RTE**: cast as “Is the hypothesis entailed by the premise? `Answer`: "Yes/No”.
- **COPA**: premise as passage; candidate cause/effect as `answer`; label indicates the correct alternative (Yes if the shown choice is correct).
- **WiC**: sentence pair as `Passage` with the target word highlighted; `Question`: “Does the word have the same meaning?”; `Answer`: “Yes/No”.
- **WSC**: passage contains the pronoun and candidate antecedent; `Question`: “Does the pronoun refer to {candidate}?”; `Answer`: “Yes/No”.

This unified phrasing stabilizes the pooled classifier and makes the rationale generator consistent across tasks, following the spirit of UnifiedQA’s “single prompt for many formats.”

E.3 Teacher Prompt for Rationale Generation

Rationales r are produced by a CoT-capable teacher (Gemini 2.5 Flash v3; Gemini Team, Google DeepMind, 2025) using a guarded prompt. The gold label is provided to the teacher but *the rationale is not allowed to restate the label.* We enforce a sentinel to separate explanation and answer.

> **Instruction (to teacher).** Given the passage, question, and candidate answer, write a short explanation (2–5 sentences) that justifies the *gold label*. Let’s think step by step.
> **Rules:** (1) Do *not* include “Yes/No”, “True/False”, or synonyms in the explanation; (2) End the explanation with `[END OF REASONING]`; (3) The explanation should be a short paragraph. (3) On a new line, output `Answer: Yes` or `Answer: No`.
>
> **I/O Format**

```
Passage: ...
Question: ...
Answer: ...
Reasoning: <concise explanation not revealing the label> [END OF
REASONING]
Answer: <Yes|No>
```

E.4 LoRA Configuration

We adapt the backbone with PEFT-LoRA applied to attention and MLP projections:

Target modules	`[q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj]`
Rank r / α / dropout	$r = 16$, $\alpha = 32$, dropout $= 0.1$
Bias	`none`
Task type	`CAUSAL_LM`

Only the LoRA adapters and the small classification head are trainable.

E.5 Training Hyperparameters

We keep the base decoder architecture unchanged and optimize a weighted loss $\mathcal{L}_{\text{total}} = \beta\mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{reason}}$.

- **Optimizer**: AdamW (`optim=adamw_torch`), LR 2×10^{-4}, weight decay 0.01.
- **Schedule**: cosine with `warmup_steps=5`.
- **Precision**: `bf16`.
- **Batching**: per-device batch size $= 1$; `gradient_accumulation_steps=32`.
- **Epochs**: 3; `save_strategy=eval_strategy=epoch`.
- **Loss weights**: unless otherwise specified in tables, we report `Reasoning1` with β=1, α=1 (and ablate $\alpha \in \{0.5, 1.0\}$).
- **Collator**: dual-sequence collator that pads classification and explanation streams independently and masks explanation padding to `-100`.
- **DDP**: `ddp_backend=NCCL`, `ddp_find_unused_parameters=True`.

E.6 Generation Settings for Reasoning Head CoT

When we generate rationales (or analyze decoding throughput), we use:

- `do_sample=True`, `temperature=0.1`, `top_p=0.7`, `max_new_tokens=500`.
- EOS/pad from the tokenizer; sequences are decoded with special tokens skipped.

E.7 Throughput (QPS) Measurement Protocol

Classification-head QPS is computed as *number of samples / wall-clock* for a pure forward pass (no text generation). Reasoning-head QPS measures decoding with the settings above. For multi-GPU, we report per-rank stats and an aggregated global QPS using a max-reduce of wall-clock time across ranks to avoid optimistic scaling.

F Future Work

We plan to explore larger backbones and grid search for optimal α, β values and introduce task-aware auxiliary objectives (e.g., span supervision for WiC/WSC). We also plan to investigate curriculum schedules that gradually anneal the reasoning loss and explore knowledge distillation with soft labels for the train-time reasoning head (e.g., temperature-scaled KL divergence between teacher and student token distributions over the input+rationale span) to test whether soft targets improve stability and transfer without changing inference cost. We also should consider running head-to-head evaluations against recent *chain-of-thought compression* approaches (e.g., HiddenCoT, C3oT, and CCoT) under matched token budgets and latency constraints, reporting accuracy-per-token and robustness to teacher-rationale quality (Liu et al., 2024; Kang et al., 2024; Cheng and Van Durme, 2024). We will also explore hybrid variants that combine DHRD with lightweight rationale selection or summary-style sketches to further reduce reasoning tokens while preserving accuracy.

G Assets & Licenses

We use third-party datasets and models under their original terms; all use is for non-commercial research. Table 6 lists the assets and licenses.

Table 6: Third-party assets and licenses.

Asset	Type	License	Notes
BoolQ (Clark et al., 2019)	Dataset	CC BY-SA 3.0	Used via SuperGLUE splits for research.
CB (de Marneffe et al., 2019)	Dataset	Original terms	Research use per SuperGLUE inclusion; cite original source.
COPA (Roemmele et al., 2011)	Dataset	Original terms	Research use per SuperGLUE inclusion.
MultiRC (Khashabi et al., 2018)	Dataset	Original terms	Research use per SuperGLUE inclusion.
RTE (Dagan et al., 2006)	Dataset	Original terms	Research use per SuperGLUE inclusion.
WiC (Pilehvar and Camacho-Collados, 2019)	Dataset	CC BY-NC 4.0	Non-commercial.
WSC (Levesque et al., 2012)	Dataset	CC BY 4.0	—
Llama 3.x (Grattafiori et al., 2024)	Model	Meta Llama Community License	Used for research fine-tuning.
Qwen 3 (Yang et al., 2025)	Model	Apache-2.0	—
Gemini 2.5 Flash (Gemini Team, Google DeepMind, 2025)	API outputs	Gemini API Additional Terms	Used to generate training rationales.

Compliance. We follow the license obligations for each asset (attribution, non-commercial restrictions, and terms of service) and use these assets solely for research. We do not redistribute datasets or model weights.