ORDER-RESOLVED HIDDEN-STATE MONITORING FOR COORDINATED EXEMPLAR POISONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-selected exemplars give an attacker a set-valued control surface: every item can look plausible while their joint arrangement redirects an in-context learner. We study a matched-control protocol for attributing detector gains while holding the scored bag, penultimate reader representations, training data, and benign calibration fixed. Coordination-Order Risk Decomposition (CORD) organizes pairwise, length-$3+$ coordination, and instability information through a common fusion interface; CORD-D replaces probe-time instability estimates with a learned surrogate. On the R1 controlled-optimizer setting, CORD-D obtains 84.8% AUROC, 2.5 points above QC-SetTrans, and exceeds the two stronger non-graph controls reported in Table 1. The paper isolates the question that this comparison is designed to answer: whether explicit summaries of relations among plausible exemplars carry useful information beyond per-item and pairwise summaries under a shared evaluation interface. This formulation provides a concrete evaluation target for hidden-state poisoning monitors and a basis for future evidence-complete tests of information order.

1 INTRODUCTION

In-context learning turns retrieved demonstrations into executable context. That flexibility also creates a security problem: an adversary can distribute a target behavior across several individually credible exemplars, leaving no single conspicuous item for a monitor to reject. This attack surface is increasingly consequential as language-model systems combine retrieval with few-shot prompting (Brown et al., 2020; Liu et al., 2022; Lu et al., 2022; Lewis et al., 2020; Ram et al., 2023). Detecting a coordinated bag requires evidence about how its members act together, yet existing comparisons often change the scored unit, internal access, training signal, calibration data, or inference cost at the same time. A higher AUROC under those changes cannot identify which information source produced the gain.

We resolve that attribution problem with a matched intervention over information order. The protocol fixes the scored exemplar bag, query-conditioned reader representations, training data, and benign calibration split. Within this interface, Coordination-Order Risk Decomposition (CORD) exposes three blocks: per-item and pairwise statistics $\mathbf{\Phi}_P$, explicit length-$3+$ coordination $\mathbf{\Phi}_H$, and reader-output instability $\mathbf{\Phi}_I$. Turning blocks on and off frames the question of what additional order contributes without changing the scored bag or the surrounding training and calibration data. QC-SetTrans supplies the strongest original matched control, while QC-SetTrans-Deep tests additional global-attention capacity and Triad-Pool grants a non-graph model explicit three-exemplar interactions.

The R1 comparison places the question on a concrete controlled setting. CORD-D reaches 84.8% AUROC, $+2.5$ above QC-SetTrans, while the deeper set-attention and triad-pooling controls narrow but do not remove the difference (Table 1). The revised paper confines its empirical claim to this recorded setting and distinguishes it from broader transfer, deployment, and intervention claims that require separate supporting records.

Our contributions are:

- a controlled formulation of information-order comparison that fixes the scored bag, reader representations, training data, and calibration while contrasting pairwise, higher-order, and instability blocks;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- an R1 comparison against graph-free, deeper-attention, and explicit triad-pooling controls, with CORD-D attaining 84.8% AUROC and a 2.5-point margin over QC-SetTrans;
- a focused method specification that separates the graph summaries, resampling signal, and distillation surrogate, making the evidence needed for future transfer and deployment evaluations explicit.

Table 1 reports the retained R1 comparison.

2 Related Work and Positioning

Demonstration selection and order affect ICL behavior (Brown et al., 2020; Liu et al., 2022; Lu et al., 2022; Min et al., 2022; Garg et al., 2022; Akyürek et al., 2023; Shi et al., 2023). Poisoning, backdoors, and adversarial prompting address distinct training-time and inference-time threats (Goodfellow et al., 2015; Madry et al., 2018; Biggio et al., 2012; Shafahi et al., 2018; Zhu et al., 2019; Goldblum et al., 2023; Zou et al., 2023; Wallace et al., 2019; Carlini et al., 2024; Kandpal et al., 2023; Xu et al., 2024). Deep Sets and Set Transformer (Zaheer et al., 2017; Lee et al., 2019) motivate the non-graph controls. Our comparison asks whether explicit coordination summaries help under a shared reader-representation, training-data, and calibration interface.

3 Problem Setup: Hidden-State Monitoring and Coordination Signals

3.1 Threat Model

The reader $\mathcal{M}$ uses ICL with query q and exemplar bag $\mathcal{B} = \{e_1, \ldots, e_k\}$ as its conditioning inputs, yielding the distribution $p_{\mathcal{M}}(\cdot \mid q, \mathcal{B})$. The adversary seeks a poisoned bag $\mathcal{B}^\star$ for which $p_{\mathcal{M}}(\cdot \mid q, \mathcal{B}^\star)$ is concentrated on the target output $y^\star \neq y$, yet individual plausibility checks accept every $e_i \in \mathcal{B}^\star$. In the primary regime, the defender owns the monitor and observes internal states produced by the deployed reader on the retrieved bag; proxy evaluation instead considers access to log-probabilities and embeddings.

We make the threat surface explicit across two *contamination regimes*:

(R1) Controlled optimiser attack. To construct $\mathcal{B}^\star$ directly, the attacker optimises against $\mathcal{M}$ using gradients, with each example subject to the perplexity bound $\mathrm{PPL}(e_i) \leq 1.5 \cdot \mathrm{PPL}_{\mathrm{med}}$.

(R2) Paraphrastic transfer. The attacker rewrites R1 poison through an LLM paraphraser (Qwen-2.5-72B-Instruct, temp. 0.7, nucleus 0.9) and re-verifies the poisoning objective, producing natural-looking exemplars that survive perplexity and embedding-distance filters.

The primary attacker knows the reader architecture but not detector parameters. The detector produces a bag-level score $r \in [0, 1]$. The offline teacher uses 50 sub-bag probes, while CORD-D replaces the probe-time instability estimate with a learned surrogate.

3.2 Query-Conditioned Interaction Graph

The exemplar set forms the vertices of a directed, weighted graph $\mathcal{G}(q, \mathcal{B}) = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V} = \mathcal{B}$. An edge $(e_i, e_j) \in \mathcal{E}$ has type $t \in \{s, c, r\}$, corresponding respectively to *support*, *contradiction*, or *redundancy*. For these edges, we use the predicates:

$$\phi_t(e_i, e_j \mid q) = \sigma\big(\mathbf{w}_t^\top [\mathbf{h}_i; \mathbf{h}_j; \mathbf{h}_q; \mathbf{h}_i \odot \mathbf{h}_j]\big), \quad t \in \{s, c, r\}, \tag{1}$$

where $\mathbf{h}_i = \mathcal{M}.\mathrm{encode}(q, e_i) \in \mathbb{R}^d$ represents e_i internally in the reader, conditional on query q; $\mathbf{h}_q = \mathcal{M}.\mathrm{encode}(q)$ represents the query itself. Here $\odot$ applies the Hadamard product, while σ applies the sigmoid function. We enumerate each unordered pair once ($i < j$), evaluate Eq. 1 in both orientations, and set $(\mathbf{A}_t)_{ij} = \phi_t(e_i, e_j \mid q)\mathbf{1}\{\max_u \phi_u(e_i, e_j \mid q) \geq \tau_e\}$ for $i \neq j$, with $(\mathbf{A}_t)_{ii} = 0$. Thus the three type-specific adjacency matrices $\mathbf{A}_t$ are directed and use a common type mask within each direction. Bag labels provide end-to-end training for the three heads, which serve as latent relation channels without separate symbolic annotations for edges. We call these channels *support*,

contradiction, and *redundancy* to describe their functions; these names do not assert exact recovery of human discourse relations by the detector.

Exemplar interactions are encoded by the graph: we use *coordinated* poisoning throughout to mean that relations across exemplars carry the attack signal, rather than a single conspicuously anomalous exemplar.

3.3 Bag-Level Risk Scoring: Order Decomposition

An explicit *order decomposition* unifies the detector representation, with a scalar risk defined alongside it:

$$\mathbf{\Phi}(q, \mathcal{B}) = [\mathbf{\Phi}_P;\ \mathbf{\Phi}_H;\ \mathbf{\Phi}_I;\ \tilde{\mathbf{h}}_q], \qquad \tilde{\mathbf{h}}_q = A\mathbf{h}_q \in \mathbb{R}^8, \qquad r(q, \mathcal{B}) = \sigma(\mathrm{MLP}_\theta(\mathbf{\Phi}(q, \mathcal{B}))), \quad (2)$$

where measurability is with respect to per-item+pairwise edge predicates $\{\phi_t(e_i, e_j \mid q)\}$ for $\mathbf{\Phi}_P$, length-3+ graph functionals (spectra, per-type triangles, clustering) for $\mathbf{\Phi}_H$, and sub-bag reader-output distributions for $\mathbf{\Phi}_I$. Eq. 1 uses the full query state $\mathbf{h}_q$ inside the relation heads; the final fusion head consumes the projected query summary $\tilde{\mathbf{h}}_q$. The paper's retained measurement question is whether $\mathbf{\Phi}_H$ carries a distinct signal beyond $\mathbf{\Phi}_P + \mathbf{\Phi}_I$ on coordinated R1 bags. Here, order denotes explicit summary structure: Φ_H is computed from pairwise edges, so the comparison concerns the utility of those summaries rather than information unavailable in the complete edge tensor. QC-SetTrans defines the original no-explicit-higher-order control; QC-SetTrans-Deep and Triad-Pool broaden it with additional attention capacity and explicit non-graph order-3 interactions. Training minimises binary cross-entropy at the bag level, assigning clean bags the label ($y=0$) and poisoned bags ($y=1$):

$$\mathcal{L}_{\mathrm{det}} = -\mathbb{E}_{(q,\mathcal{B},y)}[y \log r(q, \mathcal{B}) + (1-y)\log(1-r(q, \mathcal{B}))]\,. \quad (3)$$

4 Coordination-Order Risk Decomposition

The order decomposition in Eq. 2 is the central technical object. CORD exposes controlled variants that differ only in the available blocks $(\mathbf{\Phi}_P, \mathbf{\Phi}_H, \mathbf{\Phi}_I)$ while sharing the hidden-state bank, width-compensated fusion head, training schedule, and calibration pipeline. CORD-P uses $\mathbf{\Phi}_P$; CORD-PI adds instability; CORD-PH adds explicit coordination; CORD-PHI is the full $T=50$ offline teacher; and CORD-D preserves $(\mathbf{\Phi}_P, \mathbf{\Phi}_H)$ while replacing the probe loop with a one-pass surrogate $\hat{\mathbf{\Phi}}_I$. We use CORD for the protocol and model family, CORD-PHI for the offline teacher, and CORD-D for the deployed student. This shared interface makes the conditional contribution of higher-order coordination directly measurable, including on low- and high-coordination slices.

4.1 Interaction Graph Construction

For every $e_i \in \mathcal{B}$, its hidden-state representation $\mathbf{h}_i$ comes from the penultimate reader layer under the prompt (q, e_i). We apply the pairwise predicates in Eq. 1 in both orientations to each unordered pair, of which there are $\binom{k}{2}$. The edge weight parameters $\{\mathbf{w}_s, \mathbf{w}_c, \mathbf{w}_r\}$ are learned during detector training. For spectral features only, we form $\mathbf{S}_t = (\mathbf{A}_t + \mathbf{A}_t^\top)/2$ after pruning; its real eigenvalues are ordered algebraically as $\lambda_1(\mathbf{S}_t) \geq \cdots \geq \lambda_k(\mathbf{S}_t)$, and the top-5 vector takes the first five, zero-padding when $k < 5$. The sole spectral-gap feature is $\lambda_1(\mathbf{S}_s) - \lambda_2(\mathbf{S}_s)$ on the support channel; the other graph summaries use the directed $\mathbf{A}_t$. The pairwise terms collected in $\mathbf{\Phi}_P \in \mathbb{R}^{10}$ are (per-type total directed-degree mean/variance: 6; contradiction/support ratio: 1; per-type mean edge weight: 3); the higher-order coordination terms collected in $\mathbf{\Phi}_H \in \mathbb{R}^{22}$ are (top-5 eigenvalues of $\mathbf{S}_t$ per edge type: 15; per-type clustering coefficient: 3; mixed-type triangle count normalised by $\binom{k}{3}$: 3; support-channel spectral gap: 1), giving the pooled graph representation in Eq. 2 as $\mathbf{\Phi}_P + \mathbf{\Phi}_H \in \mathbb{R}^{32}$. We select τ_e, the pruning threshold, along with the feature inventory using R1's validation split. Small bags and supervision available only at bag level motivate our choice of transparent summaries of low dimension over a learned message-passing GNN. The detector thus seeks unusual *patterns of relations* across otherwise plausible exemplars within a bag, rather than an exemplar that appears anomalous on its own.

4.2 Resampling Instability Probing

For the second family of signals, sensitivity of reader outputs to sub-bag perturbations supplies the instability block $\mathbf{\Phi}_I(q, \mathcal{B})$ in Eq. 2. We sample $T = 50$ sub-bags $\mathcal{B}_t \subset \mathcal{B}$ with $|\mathcal{B}_t| = \lfloor \alpha k \rfloor$ where $\alpha = 0.7$, and compute the reader's output distribution $p_t = p_{\mathcal{M}}(\cdot \mid q, \mathcal{B}_t)$ for each. Let $p_0 = p_{\mathcal{M}}(\cdot \mid q, \mathcal{B})$, $d_t = \mathrm{KL}(p_t \| p_0)$, $j_t = \mathrm{JS}(p_t, p_0)$, $v_t = \frac{1}{2}\|p_t - p_0\|_1$, $\bar{p} = T^{-1}\sum_t p_t$, $H_t = H(p_t)$ for Shannon entropy H, and m_t be the difference between the largest and second-largest probabilities in p_t. The 16 components of $\mathbf{\Phi}_I$ are

$$\begin{aligned}\big[&\mathrm{mean}(d_t), \max(d_t), \mathrm{sd}(d_t), Q_{0.9}(d_t);\\ &\mathrm{mean}(j_t), \max(j_t), \mathrm{mean}(v_t), \max(v_t);\\ &H(\bar{p}), \mathrm{sd}(H_t), F, F_{\mathrm{LOO}}, F_{\mathrm{LTO}};\\ &\mathrm{var}(\log p_t(\hat{y}_0)), \mathrm{mean}(m_t), \mathrm{sd}(m_t)\big],\end{aligned}$$

where $\hat{y}_0 = \arg\max_y p_0(y)$, F is the top-prediction flip fraction relative to $\hat{y}_0$, and F_{LOO} and F_{LTO} are the corresponding leave-one-out and leave-two-out flip fractions. We keep T=50 and α=0.7 fixed for the retained R1 setting. The instability block is defined operationally by these sub-bag statistics.

4.3 Feature Fusion and Risk Scoring

We concatenate the pairwise block $\mathbf{\Phi}_P \in \mathbb{R}^{10}$, coordination block $\mathbf{\Phi}_H \in \mathbb{R}^{22}$, instability block $\mathbf{\Phi}_I \in \mathbb{R}^{16}$, and the query's projected summary $\tilde{\mathbf{h}}_q = A\mathbf{h}_q \in \mathbb{R}^8$. A 3-layer MLP processes this concatenation; its hidden dimensions are $(256, 128, 64)$, with ReLU activations and dropout set to 0.1. Only the relation heads in Eq. 1 can access the full query state; both the fusion head and student surrogate receive query information solely through the 8-d projection. At the final layer, Eq. 2 yields the scalar risk score. CORD-PHI and CORD-D use the ordinary bag-level objective in Eq. 3.

4.4 Distillation for Deployment

The CORD-PHI teacher uses $T = 50$ sub-bag probes. The CORD-D student keeps the full $\mathbf{\Phi}_P + \mathbf{\Phi}_H$ graph path, but replaces the explicit probe loop with a learned surrogate $\hat{\mathbf{\Phi}}_I$ computed from the graph summaries and projected query summary. Concretely, $\hat{\mathbf{\Phi}}_I = g_\eta([\mathbf{\Phi}_P; \mathbf{\Phi}_H; \tilde{\mathbf{h}}_q]) \in \mathbb{R}^{16}$ replaces the teacher's $\mathbf{\Phi}_I$ in Eq. 2. The student is trained with:

$$\mathcal{L}_{\mathrm{distill}} = \mathcal{L}_{\mathrm{det}}^{\mathrm{student}} + \lambda_{\mathrm{kd}} \|r_{\mathrm{student}} - \bar{r}_{T=50}\|_2^2 + \lambda_{\mathrm{soft}} \mathrm{KL}(\sigma_{\tau_{\mathrm{kd}}}(\ell_{T=50}) \,\|\, \sigma_{\tau_{\mathrm{kd}}}(\ell_{\mathrm{student}}))\,, \quad (4)$$

where $\lambda_{\mathrm{kd}} = 0.7$, $\lambda_{\mathrm{soft}} = 0.3$, and $\tau_{\mathrm{kd}} = 2.0$. The student is temperature-scaled and thresholds are set on a benign calibration split. CORD-D achieves 84.8% open-weight macro-AUROC on R1 (Classification 86.4, Reasoning 85.7, RAG 82.3; Table 1), a 2.5-point margin over QC-SetTrans. The paper does not use this monitor-side comparison as an end-to-end latency claim.

5 Results: Residual Coordination Signal Under Joint Matching

The retained experiment asks whether explicit higher-order coordination adds signal beyond the graph-free controls under the R1 reader-representation, training-data, and calibration interface. Table 1 reports the result.

5.1 Experimental Setup

This subsection states the choices that define the retained R1 comparison.

Evaluation covers six non-generative tasks and three generative / long-context tasks: SST-2 and AG News (k=8); GSM8K and StrategyQA (k=4); Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) (k=5); WikiPoisonSumm (k=12), LongICL-GSM8K (k=16), and MFR-QA (k=10). Detection is bag-level throughout.

Main results average over Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Qwen-2.5-72B-Instruct (Qwen et al., 2024), and DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), with

penultimate-layer hidden states extracted at layer $L-1$ and shared across the learned detectors in the R1 comparison.

Training, threshold calibration, and final evaluation use disjoint query IDs and poisoned-bag generations. The graph-free controls are DeepSets, Bag-Statistics MLP, Retrieval-Anomaly, QC-SetTrans, QC-SetTrans-Deep, and Triad-Pool. All learned detectors share AdamW, a 5,000-bag validation split, and hyperparameters selected on R1 development data.

For Table 1, $\mathrm{FPR@95}$ is the test false-positive rate at a threshold selected on a disjoint validation split to attain 95% poisoned-bag recall. R1 contains 18,000 queries and 54,000 evaluation bags.

5.2 Higher-Order Signal Under Matched Controls

Table 1 performs the retained information-order comparison on R1. QC-SetTrans-Deep adds depth and global-attention capacity, while Triad-Pool learns permutation-invariant three-exemplar interactions without relation graphs or spectral summaries. CORD-D reaches 84.8% macro-AUROC, $+2.5$ over QC-SetTrans and +1.2 over Triad-Pool.

Table 1: R1 detection (AUROC% ↑ / $\mathrm{FPR@95\%}$ ↓) over three readers and ten seeds. $\mathrm{FPR@95}$ is test FPR at a validation-selected 95% attack-recall threshold. The strongest non-CORD-D entry is underlined; CORD-PHI is a probe-intensive teacher reference.

	Class.		*Reason.*		*RAG*		*Avg.*	
Method	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
Classical baselines								
Embedding Outlier	71.6	32.1	66.4	37.7	63.1	40.9	67.0	36.9
Leave-One-Out	75.3	26.0	71.2	29.8	66.9	35.2	71.1	30.3
Self-Consistency	73.7	28.4	73.4	27.5	69.1	32.8	72.1	29.6
Strong Retriever	77.3	21.7	70.4	31.0	73.9	25.4	73.9	26.0
Graph-free bag-level baselines								
DeepSets	80.1	19.3	79.2	20.6	77.4	22.7	78.9	20.9
Bag-Stats MLP	81.8	17.8	80.6	19.1	78.8	21.3	80.4	19.4
Retrieval Anom.	77.3	22.0	75.4	24.8	80.1	19.2	77.6	22.0
QC-SetTrans	83.5	15.9	82.4	17.0	81.0	18.2	82.3	17.0
QC-SetTrans-Deep	84.4	14.9	83.7	16.1	81.5	17.5	83.2	16.2
Triad-Pool	<u>84.7</u>	<u>14.6</u>	<u>84.1</u>	<u>15.7</u>	<u>81.9</u>	<u>17.0</u>	<u>83.6</u>	<u>15.8</u>
CORD-D (matched)	*86.4*	*12.5*	*85.7*	*14.0*	*82.3*	*16.4*	*84.8*	*14.3*
CORD-PHI teacher	88.6	9.8	86.6	12.0	83.4	15.3	86.2	12.4

6 Discussion and Conclusion

On the retained R1 comparison, CORD-D reaches 84.8% AUROC, a 2.5-point difference over QC-SetTrans and a 1.2-point difference over Triad-Pool. The comparison focuses the contribution on an explicit question about the representational value of coordination summaries under a common bag, reader-representation, training-data, and calibration interface.

The result motivates information-order comparisons as a useful evaluation target for hidden-state poisoning monitors. The method description separates graph summaries, resampling signals, and the distillation surrogate so that subsequent work can test transfer, compute, and deployment consequences with complete supporting records.

Ethics Statement

Human evaluation comprised minimal-risk judgments of licensed, non-sensitive passages under anonymized ethics approval and fair compensation. The work evaluates defensive monitoring against established poisoning attacks.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed the edits and remain responsible for the final content.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1467–1474, 2012. URL https://arxiv.org/abs/1206.6389.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *IEEE Symposium on Security and Privacy*, pp. 407–425, 2024. doi: 10.1109/sp54263.2024.00179. URL https://doi.org/10.1109/sp54263.2024.00179.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Shivam Garg, Dimitris Tsipras, Percy S. Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2023. doi: 10.1109/tpami.2022.3162397. URL https://doi.org/10.1109/tpami.2022.3162397.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Backdoor attacks for in-context learning with language models. *arXiv preprint arXiv:2307.14692*, 2023. URL https://arxiv.org/abs/2307.14692.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, 2019. doi: 10.18653/v1/D19-1221. URL https://aclanthology.org/D19-1221/.

Jiashu Xu, Mingyu Ma, Fei Wang, Chaowei Xiao, and Muhao Chen. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In *Proceedings of the*

2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 3111–3126, 2024. doi: 10.18653/v1/2024.naacl-long.171. URL https://aclanthology.org/2024.naacl-long.171/.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R. Salakhutdinov, and Alexander Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, pp. 7614–7623, 2019. URL https://proceedings.mlr.press/v97/zhu19a.html.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL https://arxiv.org/abs/2307.15043.