

How does Instruction Hierarchy Training mitigate prompt injections: Preliminary results from an attentional study

Aman Neelappa

Amey Muke

June 22, 2026

Abstract

Enforcing an instruction hierarchy is a promising approach to mitigating a large class of prompt injection attacks. Recent work has shown that instruction-hierarchy training (IHT) is not only effective in preventing unauthorized instruction execution but also generalizes well to out-of-distribution domains [8]. In this work we study the impact of such training on attention patterns in an effort to gain a mechanistic understanding of how it works. We fine-tune a Qwen3-8B model [6] with an IHT-analog GRPO objective and show that it places significantly less attentional weight on instructions embedded in tool/data responses in an indirect prompt injection dataset [10]. On a paired evaluation over 1,000 examples, IHT reduces the injected-channel probe-attention fraction from 0.229 to 0.220 ($p \approx 8 \times 10^{-15}$) and the injected-channel PVCC from 1.60 to 1.49 ($p \approx 9 \times 10^{-7}$), while slightly *increasing* attention in the legitimate instruction channel—a channel-specific shift. Behaviourally, injection ASR falls from 71.1% to 67.6% (−3.5pp, McNemar $p < 10^{-3}$) as utility rises from 82.2% to 85.2%. Crucially, the causal attention→injection pathway and the decodability of the injection signature are preserved: attention knockout still removes ~18pp of injection ASR in the IHT model and a leakage-free per-head probe stays highly separable (AUROC 0.82 → 0.87). IHT therefore shifts the operating point *along* the same attention mechanism rather than re-routing it.

1 Introduction and Main Contribution

Indirect prompt injection succeeds when a model treats text arriving in a *data* channel as if it were an *instruction* [3, 5]. A natural mechanistic signature is *where the model attends* while generating: if an injected instruction has hijacked control, the model should attend to it more than to the surrounding benign data. We operationalize this with the **Probe-vs-Clean-Context ratio (PVCC)** [1]: with the injected instruction as the *probe* and the legitimate content as *clean data*, over the first $m=32$ response tokens we compute the per-token-averaged attention share (PTAS) to each span and set $PVCC = PTAS_{\text{probe}}/PTAS_{\text{clean}}$. Token normalization prevents longer spans being mechanically favored, and

the early window targets the “commit” step at which the model selects which instruction to follow.

This report establishes the *baseline* attention behaviour of the open Qwen3-8B model [6], against which an instruction-hierarchy-trained (IHT) counterpart will later be compared. **Our contributions are:**

1. A recomputed, CI-bearing characterisation of the PVCC effect size across two dataset families (Section 4.1), showing a directionally robust but strongly dataset-dependent signal.
2. Per-layer and per-head attention probes (Section 4.2) that lift detection from AUROC 0.57 (PVCC alone) to 0.84, isolate a *leakage-free* probe-only signal (AUROC 0.83, clean-control at chance), localise it to the mid-stack (layer 18), and show it transfers across dataset families.
3. A **causal** attention-knockout/amplification experiment (Section 4.3) that upgrades PVCC from a correlate to a mechanism: editing response→probe attention moves injection success monotonically, with a null random-span control.

Together these turn an attention *correlate* of injection into a *causally validated* mechanism, providing the baseline for measuring whether instruction-hierarchy training acts by reshaping this attention.

2 IHT Training Details

The instruction-hierarchy model studied here is an open-model *analog* of IHT: a GRPO-trained LoRA adapter applied to the same Qwen/Qwen3-8B base used throughout this report. It is not an exact reproduction of any private recipe; it is a separation/instruction-priority replication on a clean, validated attack corpus. All facts below are read directly from the released adapter (`adapter.config.json`, model card) and the experiment protocol on disk.

The reward is reinforcement-learning-style: the policy is rolled out on the `openai/ih-challenge` task skeletons [4] and scored by a deterministic per-example Python grader that rewards following the higher-priority (system) instruction while *resisting* a lower-priority injected attack—i.e. the behavioural separation target reported in Section 5. Attacks are synthesised once by a frozen

Table 1: IHT (IH-analog) fine-tune configuration, as recorded on disk.

Field	Value
Base model	Qwen/Qwen3-8B
Adapter	qwen3-8b-ih-grpo-seed42
Method	GRPO RL, PEFT LoRA
LoRA rank / α	32 / 32 (dropout 0)
LoRA targets	q,k,v,o,gate,up,down_proj, lm_head
Reward	deterministic IH grader (separation)
Group shape	8 groups $\times$ 8 rollouts
Learning rate	2×10^{-5}
KL coefficient	0.01
Seed	42
Steps / selected	500 / step-440 (best)
Train / test rows	1,024 / 254 (openai/ih-challenge)

attacker model and validated (no refusal-like or control-header-leaking attacks; no train/test overlap). Training uses GRPO with 8 groups $\times$ 8 rollouts, the Qwen3 non-thinking renderer, and seed 42; of the 500-step run we use the released best checkpoint, step 440 (held-out grader accuracy 82.68%, 210/254).

For the mechanistic comparison (Section 5) the architecture, tokenizer, prompt template (index 0), greedy decoding, example ordering, span resolver, and attention implementation are held fixed; the LoRA adapter is the only model-side difference. Both models are evaluated paired on the *same* first 1,000 SEP examples [10], seed 42. Hyperparameters not recorded on disk (e.g. optimizer schedule, batch construction beyond the group shape) are not stated here.

3 Dataset and Metrics

Datasets. Two families differ deliberately in how the injection is constructed and how often it succeeds. The **validation family** is built on SQuAD reading-comprehension passages [7]: a grounded QA task supplies the clean data and an explicit adversarial probe is inserted into it; because the probe competes against a concrete task, attacks succeed at a relatively low $\sim 36\text{--}40\%$ prompt-in-data ASR. The **SEP family** [10] (“Should it be Executed or Processed?”) pairs GPT-generated task/data with fixed probe instructions appended in controlled positions; its softer primary tasks yield a much higher $\sim 71\text{--}72\%$ ASR. SEP records carry covariates (probe insistence, append position, task type) used for stratification.

Metrics. Each example is run in two conditions: **data_with_probe** (probe in the data channel; following it is an *injection success*, scored by **prompt_in_data_asr**) and **probe_with_task** (probe legitimately in the instruction channel; a utility proxy, **probe_in_instruct_asr**). Success uses deterministic witness-substring matching,

Table 2: Attack success rate (ASR) by dataset family and split. In-data = **prompt_in_data_asr** (injection); in-instr = **probe_in_instruct_asr** (utility).

Dataset / split	n	In-data	In-instr
Validation (ASR eval)	100	40.0%	44.0%
Validation (attn subset)	1000	36.1%	–
SEP (ASR eval)	9160	70.6%	84.7%
SEP (attn subset)	1000	71.6%	–

Table 3: PVCC effect size (injection success vs. failure), recomputed from per-record values with analytic 95% CIs ($d > 0$: success attends more to the probe). The effect is large on grounded validation data, modest on SEP.

Run	n	ASR	Cohen’s d	95% CI	MWU p
val-100	100	40.0%	0.742	[0.33, 1.16]	1e-3
val-1000	1000	36.1%	0.748	[0.62, 0.88]	2e-28
SEP-1000	1000	71.6%	0.235	[0.10, 0.37]	3e-4

which is model-agnostic and cheap. The PVCC attention metric is defined in Section 1; injection success is the target label throughout. Probes are trained CPU-only with L_2 logistic regression (5-fold out-of-fold scoring, C by inner CV, permutation nulls).

The two $n=1000$ attention subsets used below track their parent ASR corpora closely (validation 36.1% vs. 40.0%; SEP 71.6% vs. 70.6%), so the attention measurements are representative of each family (Table 2).

4 Baseline Metrics

This section reports the baseline attention behaviour of the open Qwen3-8B model: the PVCC effect size, attention-probe detection, and a causal knockout test.

4.1 PVCC

We compute PVCC for every **data_with_probe** record and compare successful against failed injections; effect size is Cohen’s d (positive $\Rightarrow$ success has higher PVCC) with an analytic 95% CI and a two-sided Mann–Whitney U test, recomputed from per-record values (Table 3, Figure 1).

The *direction* is robust—successful injections attend relatively more to the probe in every run—but the *magnitude* is dataset-dependent: large on grounded validation data ($d=0.74\text{--}0.75$) and only modest on synthetic SEP ($d=0.24$). The two validation estimates agree almost exactly, so the gap reflects a genuine distributional difference, not sampling noise. Within SEP, the tempting “lower ASR $\Rightarrow$ larger d ” explanation fails across 11 strata (Pearson $r=+0.28$, $p=0.41$); probe insistence is the one consistent axis (non-insistent $d=0.43$ vs. insistent $d=0.24$). PVCC is therefore a real but weak scalar

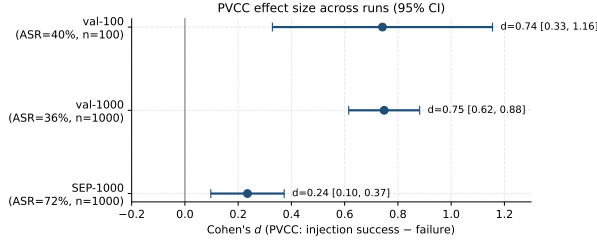


Figure 1: PVCC effect size (Cohen’s d , success – failure) with 95% CIs. All runs are strictly positive, but the SEP magnitude is far smaller than the validation magnitude.

Table 4: Feature-set ablation for SEP-1000 injection-success prediction (5-fold out-of-fold L_2 logistic regression). Per-layer/head structure (F2–F4) beats collapsed scalars (F0–F1). The probe-fraction-only variants (F2p/F3p) keep the skill while driving the matched clean-control AUROC to chance, isolating a leakage-free attention-to-probe signal.

Feature set	#feat	OOF AUROC	Clean-ctrl
F0: PVCC scalar	1	0.570	0.500
F1: 14 scalars	14	0.705	0.591
F2: per-layer	108	0.791	0.732
F3: per-head	2304	0.840	0.794
F4: per-head+std	4608	0.854	0.733
F2p : layer probe-only	36	0.750	0.500
F3p : head probe-only	1152	0.830	0.493

signal—motivating the structured probes next.

4.2 Attention Probe Training and Results

If the signal is more than a single ratio, structured attention features should predict injection success. We extract per-(layer,head) early-response attention to the probe and clean data ($L=36$, $H=32$) and train the same logistic pipeline on a ladder of feature sets: **F0** (PVCC scalar), **F1** (14 aggregated scalars), **F2** (per-layer), **F3** (per-head), **F4** (F3 + per-head std), plus probe-fraction-only variants **F2p/F3p** that zero every clean-data feature.

Out-of-fold AUROC climbs monotonically with structure (Table 4): F0 0.57 → F1 0.70 → F2 0.79 → F3 0.84 → F4 0.85, all far above their permutation nulls (≈ 0.50). PVCC alone is weak; the predictive content lives in *where* attention lands. Crucially, the richer sets also absorb structural leakage (clean-control AUROC rises to 0.73–0.79), so the decisive control is the probe-fraction-only ladder: **F3p reaches AUROC 0.83 with its matched clean-control at chance (0.49)**—comfortably above the leaky scalar 0.70 and its 0.62 structural floor. The signal localises to the mid-stack (per-layer peak at layer 18, Figure 2) and transfers across dataset families: per-head

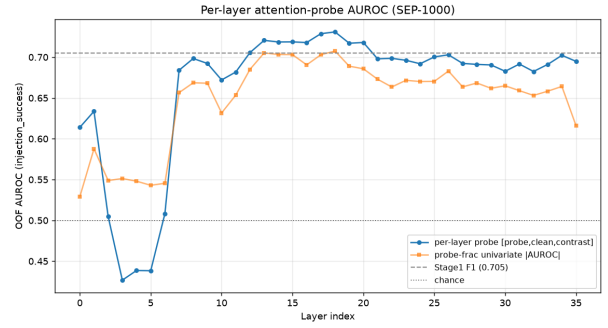


Figure 2: Per-layer out-of-fold AUROC. The injection signature is concentrated in the mid-stack, peaking at layer 18, with a broad plateau from layer ~ 12 onward and near-chance early layers.

Table 5: Causal attention intervention on Qwen3-8B (base), SEP `data_with_probe`, $n=1000$, all heads, first 32 response steps. Editing post-softmax response → probe attention shifts injection ASR monotonically; the length-matched random-span knockout is the negative control. Δ is treatment – *none* in percentage points with paired bootstrap 95% CI; p is the exact paired (McNemar) test.

Intervention	ASR	Δ pp [CI]	p
None (baseline)	0.453	—	—
Random-span knockout	0.436	−1.7 [−3, 0]	0.068
Probe knockout	0.258	−19.5 [−22, −17]	2e-41
Probe knockout (uniform)	0.249	−20.4 [−23, −18]	4e-41
Probe amplify $\times 1.5$	0.484	+3.1 [1, 5]	5e-4
Probe amplify $\times 2$	0.506	+5.3 [3, 7]	8e-7
Probe amplify $\times 4$	0.567	+11.4 [9, 14]	3e-21

probes reach SEP → val AUROC 0.78–0.80 (F3p 0.80) where the scalar aggregate collapses to 0.64. Attention-to-probe is thus a real, localizable, and transferable predictor of injection success—not a structural artifact.

4.3 Causal Interventions

Probes are still *correlational*. To test causation we intervene on attention directly. On the base Qwen3-8B `data_with_probe` condition ($n=1000$), during generation with eager attention we edit the post-softmax attention weights from each of the first 32 response positions to the probe key positions (all heads), following the attention-knockout method of Geva et al. [2] with best-practice paired controls [9]: *knockout* sets those weights to 0 and renormalises the row; *amplification* multiplies them by $\gamma \in \{1.5, 2, 4\}$. The negative control knocks out a length-matched *random* benign span instead of the probe.

The result is decisive (Table 5, Figure 3). Probe knockout cuts injection ASR from 0.453 to 0.258 (−19.5pp, paired bootstrap CI [−22.3, −16.7], McNemar $p \approx 2 \times 10^{-41}$); 217 previously successful injections flip to failure versus only 22 the other way. The length-matched

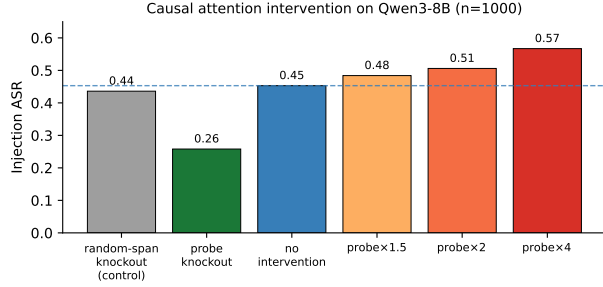


Figure 3: Injection ASR under each attention edit (base Qwen3-8B, $n=1000$, all heads, first 32 response steps). Knocking out probe attention roughly halves ASR; the length-matched random-span control is null; amplifying probe attention raises ASR monotonically. Dashed line: no-intervention baseline.

random-span knockout is essentially null (0.436, -1.7pp , $p=0.068$), so the effect is specific to probe attention and not to perturbing attention in general—the primary pre-registered contrast (probe vs. random knockout) is -17.8pp ($p \approx 5 \times 10^{-32}$). A uniform-renormalisation variant reproduces the knockout (0.249), ruling out a redistribution artifact. In the opposite direction, amplifying probe attention raises ASR monotonically in γ ($0.484 \rightarrow 0.506 \rightarrow 0.567$ at $\times 1.5/\times 2/\times 4$, all $p < 10^{-3}$). This is exactly the bidirectional, dose-dependent signature predicted if response $\rightarrow$ probe attention *causally mediates* injection success, establishing the attention-to-probe channel as a genuine mechanism rather than an epiphenomenon of PVCC.

5 IHT Metrics

We re-run the behavioural, attention, probe, and causal measurements of Section 4 on the IHT adapter and compare it to the base model on the *identical* paired SEP set ($n=1000$, template 0, seed 42). Table 6 collects the headline contrasts.

Behaviour. IHT moves all three behavioural metrics in the intended direction: injection ASR drops 71.1% $\rightarrow$ 67.6% (-3.5pp , paired bootstrap CI $[-5.5, -1.5]$, exact McNemar $p=9.2 \times 10^{-4}$), utility rises 82.2% $\rightarrow$ 85.2% ($+3.0\text{pp}$, $p=2.9 \times 10^{-4}$), and the instruction/data separation metric rises $0.203 \rightarrow 0.263$ ($+5.97\text{pp}$). The effects are statistically clear but *small* in magnitude.

Attention. The mechanistic endpoint matches the prediction. In the injected data channel, IHT lowers the probe-attention fraction $0.229 \rightarrow 0.220$ (paired $d_z=-0.25$, $p=7.8 \times 10^{-15}$) and PVCC $1.598 \rightarrow 1.492$ ($d_z=-0.16$, $p=8.8 \times 10^{-7}$). The effect is *channel-specific*: in the legitimate instruction channel PVCC instead rises slightly

Table 6: Base vs. IHT (IH-analog GRPO LoRA) on the same paired SEP set ($n=1000$, template 0, seed 42). Behaviour and injected-channel attention move in the intended direction; the legitimate channel does not (specificity); probe separability and the causal knockout effect are preserved. p is the paired test (McNemar for rates/knockout, paired t for attention means).

Metric	Base	IHT	Δ (p)
<i>Behaviour</i>			
Injection ASR	0.711	0.676	-3.5pp ($9\text{e-}4$)
Utility	0.822	0.852	$+3.0\text{pp}$ ($3\text{e-}4$)
Separation	0.203	0.263	$+6.0\text{pp}$
<i>Attention, injected channel</i>			
PVCC	1.598	1.492	-0.106 ($9\text{e-}7$)
Probe-attn. frac.	0.229	0.220	-0.009 ($8\text{e-}15$)
<i>Attention, legitimate channel (control)</i>			
PVCC	1.798	1.835	$+0.037$ (0.03)
<i>Probe & causal (signature preserved)</i>			
Probe AUROC (F3p)	0.819	0.868	$+0.049$
Knockout ASR Δ	-19.5pp	-18.1pp	both $p < 1\text{e-}35$

($+0.037$, $p=0.034$), so IHT down-weights data-channel instructions without suppressing attention to genuine instructions. The reduction is broadly distributed—949 of 1152 head-level contrasts are significant after Benjamini–Hochberg correction ($q < 0.05$)—and concentrates in the mid-stack (largest per-head reductions in layers ~ 17 – 26), overlapping the layers that carry the baseline injection signature.

Signature and causal pathway are preserved.

Critically, IHT does not erase the mechanism. A leakage-free per-head attention probe is *at least as* separable on the IHT model (OOF AUROC $0.819 \rightarrow 0.868$), and probes transfer across the two models almost perfectly (F3p base $\rightarrow$ IHT AUROC 0.906 , IHT $\rightarrow$ base 0.928), so the injection signature lives in the same feature space. The causal attention-knockout result also survives: zeroing response $\rightarrow$ probe attention removes -19.5pp of injection ASR on base ($p \approx 2 \times 10^{-41}$) and -18.1pp on IHT ($p \approx 3 \times 10^{-36}$), with the length-matched random-span control null in both ($-1.7\text{pp}/-0.7\text{pp}$) and amplification monotone in both. Taken together—a channel-specific attention reduction, an unchanged causal attention $\rightarrow$ injection pathway, and preserved (slightly higher) separability—IHT shifts the *operating point* along the existing attention mechanism rather than re-routing injection through a different circuit.

6 Limitations

Single model and IH analog. All results are for Qwen3-8B; cross-model and cross-scale generalization are untested here. The IHT model (Sections 2–5) is

an open-model *analog*—a GRPO separation LoRA on a clean attack corpus—not the original private IHT recipe, and the IHT effect sizes are *small* (-3.5pp ASR, $\Delta\text{PVCC}=-0.11$). **Scope of the IHT comparison.** For the IHT model we report the matched behavioural, attention, probe, and all-head causal contrasts; we do *not* have on disk a GRPO clean-control (e.g. a reward-shuffled or non-IH LoRA), the full F0–F4 attention-probe ladder for the IHT model (only F0/F2p/F3p), or Experiment-C activation/attention patching between base and IHT—so the claim is an operating-point shift along a shared mechanism, not a circuit-level identification. **Dataset-family dependence and base rates.** The PVCC magnitude differs sharply between families ($d \approx 0.75$ vs. ≈ 0.24); SEP’s high ASR (~ 0.72) compresses the failure group and shrinks d irrespective of true signal, so cross-dataset d comparisons partly reflect class balance. **Probe capacity.** High-capacity sets (F3/F4) memorize the training split (train–OOF gap ≈ 0.15) and rely on strong L_2 ; the full sets still absorb structural leakage (clean-control up to 0.79), which is why the leakage-free F2p/F3p variants carry the headline claim. **Causal scope.** The knockout experiment edits all heads over the first 32 response steps on the `data_with_probe` condition; per-head localisation of the causal effect (aiming the edit at the layer-18 mid-stack heads) and activation/attention patching are not yet run. The intervention’s own no-intervention baseline ASR (0.453) reflects its generation configuration and differs from the SEP attention-subset ASR (0.716); paired tests are computed within the intervention run, so this does not affect the causal conclusion. **Labeling.** Witness-substring matching is deterministic and model-agnostic but can miss paraphrased compliance or count incidental witness mentions. **Decoding/template scope.** Temperature 0, prompt template index 0 (no defensive instruction), single sample per condition; templates and temperature were not swept.

7 Conclusion and Future Work

On the open Qwen3-8B baseline, attention-to-probe is a directionally robust but dataset-dependent PVCC signal (Cohen’s $d \approx 0.75$ validation, ≈ 0.24 SEP), a real and *leakage-free* predictor of injection success once per-head structure is used (probe-only AUROC 0.83, clean-control at chance), localized to the mid-stack (layer 18) and transferable across dataset families (SEP→val 0.80). Most importantly, an attention-knockout intervention shows the signal is *causal*: removing response→probe attention roughly halves injection ASR ($0.453 \rightarrow 0.258$, $p \approx 2 \times 10^{-41}$) against a null random-span control, and amplifying it raises ASR monotonically. PVCC is therefore not merely a detector but a handle on the mechanism of injection.

We then run the matched instruction-hierarchy ablation (Sections 2, 5): an IH-analog GRPO LoRA on the

same base, evaluated paired on the same examples. IHT down-weights data-channel attention in a channel-specific way (injected probe-attention and PVCC drop; the legitimate channel does not) and lowers injection ASR, yet the causal attention→injection pathway and the decodable injection signature are preserved. IHT thus shifts the operating point along this mechanism rather than re-routing it.

Future work. The decisive next step is circuit-level localisation: activation/attention patching between base and IHT on the mid-stack heads identified here, to test whether the behavioural gain is *mediated* by the same causal heads. Further directions: a GRPO clean-control and the full F0–F4 ladder for the IHT model, per-head localisation of the causal knockout effect, cross-model/cross-scale replication, and validation on realistic indirect-injection settings [3] beyond short witness-verifiable SEP probes.

References

- [1] Anonymous. Rennervate: Attention-based detection of prompt injection in large language models. *arXiv preprint arXiv:2512.08417*, 2025.
- [2] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*, 2023. arXiv:2304.14767; introduces attention knockout.
- [3] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. arXiv:2302.12173.
- [4] Chuan Guo, Juan Felipe Ceron Uribe, Sicheng Zhu, Christopher A. Choquette-Choo, Steph Lin, Nikhil Kandpal, Milad Nasr, Rai, Sam Toyer, Miles Wang, Yaodong Yu, Alex Beutel, and Kai Xiao. IH-Challenge: A training dataset to improve instruction hierarchy on frontier LLMs. *arXiv preprint arXiv:2603.10521*, 2026.
- [5] Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium*, 2024. arXiv:2310.12815.
- [6] Qwen Team. Qwen3 technical report. <https://github.com/QwenLM/Qwen3>, 2025. Qwen3-8B instruction-tuned model.

- [7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of EMNLP*, 2016.
- [8] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.
- [9] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.16042.
- [10] Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H. Lampert. Can LLMs separate instructions from data? and what do we even mean by that? In *International Conference on Learning Representations (ICLR)*, 2025. SEP (Should it be Executed or Processed?) benchmark.