

Universiteit Leiden

Master Computer Science

Understanding Role-Playing Mechanisms in Zero-Shot LLM Rankers

Name: Ruishu Li
Student ID: s4331184

Date: 26/06/2026

Specialisation: Data Science

1st supervisor: Suzan Verberne
2nd supervisor: Michiel van der Meer
Daily supervisor: Yumeng Wang

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
2333 CA Leiden
The Netherlands

Abstract

This thesis investigates why role-playing prompts can harm zero-shot LLM rankers and how these effects are represented inside the model. Following prior work, I construct Random20, a set of ten positively and ten negatively framed role prompts, and evaluate it using a pointwise setup in which the model predicts whether a passage is relevant to a query. I then conduct controlled behavioral experiments that separate adjective polarity, adverb polarity, role identity, and their interactions across six retrieval datasets.

The behavioral results show that the largest drops in ranking performance are driven primarily by adjective–adverb combinations rather than by isolated negative words. Role-only controls on TREC-COVID indicate that semantic role identity alone is insufficient to cause severe ranking failures. However, role identity can modulate harmful lexical framing when the two are combined, although this effect is less consistent across datasets than the lexical effect itself.

To examine the internal mechanism, I apply residual-stream and attention-head activation patching, followed by single-head and group zero ablation, to the Random20 prompts. The results show that differences between positive and negative role conditions are first recoverable around role-related prompt positions and later become concentrated at the final token used for the Yes/No relevance decision. A sparse set of attention heads causally affects this decision, but these heads are functionally mixed: some support the target direction, whereas others suppress or regulate it. The intervention results also reveal different sensitivities for relevance and irrelevance judgments, suggesting partially distinct internal computations.

Overall, harmful role-playing effects in zero-shot LLM ranking are driven by systematic interactions between prompt components and are routed through identifiable, but functionally heterogeneous, transformer components.

Contents

Chapter 1 Introduction **5**

Chapter 2 Background and Related Work **7**
2.1 Zero-Shot LLM Ranking Methods 7
2.2 Role-Playing and Prompt Strategies 7
2.3 Interpretability and Mechanistic Analysis of Ranking Models 8

Chapter 3 Data **8**
3.1 TREC-DL datasets 9
3.2 BEIR benchmarks 9
3.3 Dataset Usage Across Experimental Stages 10

Chapter 4 Methods **11**
4.1 Experimental Setup and Ranking Framework 11
4.2 Behavioral Role-Component Analysis 13
4.3 Mechanistic Interpretability Analysis 15

Chapter 5 Behavioral Analysis: What Makes Negative Role Prompts Harmful? **19**
5.1 Behavioral Experimental Setup 19
5.2 Reproduction of Random20 Ranking Results 20
5.3 Lexical Polarity and Cross-Domain Transfer 21
5.4 Role Identity and Role–Lexical Interaction 24

Chapter 6 Mechanistic Analysis: How Role-Prompt Information Is Routed **29**
6.1 Where is role information encoded and routed? 30
6.1.1 Attention Head Patching Across Datasets and Prompt Positions . 30
6.1.2 Residual Stream Component Analysis 33
6.2 Helpful Head Selection for Causal Intervention 35
6.2.1 Instruction-Level Helpful Heads 36
6.2.2 Last-Token Helpful Heads 36
6.2.3 Lexical Role-Phrase Helpful Heads 37
6.2.4 Summary of Recurring Helpful Heads 38
6.2.5 Final Top-10 Heads for Single-Head Zero Ablation 39
6.3 Do the selected heads causally affect the ranking decision? 40
6.3.1 Single-head zero ablation of the selected top-10 heads 41
6.3.2 Group ablation of top-20 candidate heads 42
6.3.3 Summary of causal intervention findings 45
6.4 Summary of Mechanistic Findings 45

Chapter 7 Discussion **47**
7.1 Prompt-level harmfulness 47
7.2 Internal routing of role-prompt information 47
7.3 Asymmetry between relevance and irrelevance 48
7.4 Implications and limitations 48

Chapter 8 Conclusion **49**

Appendix A Additional Behavioral Selection Tables **56**
A.1 Role-only control results on TREC-COVID 56
A.2 Selection of harmful adjective–adverb pairs 58
A.3 Selection of role–adjective–adverb combinations 58

Appendix B Extended Activation Patching Results **60**

Appendix C Component-level activation patching **61**
C.1 TREC-COVID Component-Level Results 61
C.2 SciFact Component-Level Results 61
C.3 FIQA Component-Level Results 61
C.4 Climate-Fever Component-Level Results 61

Appendix D Attention head-level activation patching **64**
D.1 Additional Role-Decomposition Head Patching Results 64
D.2 SciFact Head-Level Results 64
D.2.1 SciFact: Relevant 64
D.2.2 SciFact: Irrelevant 64
D.3 FIQA Head-Level Results 69
D.3.1 FIQA: Relevant 69
D.3.2 FIQA: Irrelevant 69
D.4 Climate-Fever Head-Level Results 72
D.4.1 Climate-Fever: Relevant 72
D.4.2 Climate-Fever: Irrelevant 72
D.5 TREC-DL 2019 Head-Level Results 75
D.5.1 DL19: Relevant 75
D.5.2 Trec DL 2019: Irrelevant 75
D.6 TREC-DL 2020 Head-Level Results 78
D.6.1 DL20: Relevant 78
D.6.2 Trec DL 2020: Irrelevant 78
D.7 TREC-COVID Head-Level Results 81
D.7.1 TREC-COVID: Relevant 81
D.7.2 TREC-COVID: Irrelevant 81

Appendix E Additional Head-Selection and Ablation Tables **84**
E.1 Additional Top-10 Helpful-Head Results 84
E.2 Mean-ablation results for top-20 candidate heads 87
E.3 Complete Top-20 Head Rankings from Activation Patching 88

Chapter 1 Introduction

Large Language Models (LLMs) are increasingly used as zero-shot rankers for information retrieval (IR). Earlier work showed that pretrained sequence-to-sequence models can be used for document ranking [12]. More recent work has extended this idea to instruction-following LLMs, where ranking can be formulated through natural-language prompts [17, 14]. In a pointwise ranking setup, an LLM receives a query and candidate passage and is asked to judge whether the passage is relevant. This pointwise setup is one member of a broader family of LLM-based reranking methods, including pairwise, setwise, and listwise formulations [17, 14]. Because the model is not fine-tuned for the target ranking task, the prompt serves as part of the task specification from which document scores are derived. Consequently, even small changes in prompt wording can alter the scores assigned to candidate documents and, in turn, their resulting rankings [17].

One common prompt strategy is *role-playing*, where the model is assigned a role or persona before making a relevance judgment. At first sight, such role descriptions may appear to be stylistic additions: they seem to change how the model presents itself, rather than the underlying relevance computation. However, earlier work suggests that this assumption is not always valid. Wang et al. (2025) showed that role-playing prompts can substantially affect zero-shot LLM rankers: some roles improve effectiveness, while others cause large performance drops [21]. This raises an important reliability problem. If a role description can change ranking quality, then role prompts should be treated as part of the ranking method itself rather than as harmless stylistic instructions. This thesis studies why and how role-playing prompts can produce harmful ranking behavior. The main hypothesis is that harmful role-playing effects do not arise from a role prompt as a single indivisible instruction. Instead, a role prompt combines multiple components, including lexical polarity and role identity. Lexical polarity appears in negative adjectives and adverbs such as *awful*, *incorrect*, or *falsely*. Role identity refers to the semantic persona assigned to the model, such as a service worker, an entertainment-related role, or another task-mismatched identity. I separate these components and study how they affect ranking behavior individually and jointly. Rather than assuming that role identity is harmful by itself, I treat it as a potential contextual factor whose effect may depend on the surrounding lexical framing. The thesis has two connected parts. The first part is behavioral. Following the role-prompt ranking setup of Wang et al. [21], I construct a set of 20 role prompts, called Random20, by randomly pairing adjective and adverb candidates without replacement within their respective polarity categories. I first evaluate this prompt set as a behavioral baseline and then construct controlled prompt variants to identify which components cause the strongest degradation. These variants include adjective-only prompts, adverb-only prompts, adjective–adverb paired prompts, fixed-lexical role-identity prompts, and combined role–lexical prompts. This part studies how isolated negative words, lexical combinations, role identity, and role–lexical interactions contribute to ranking degradation. The second part addresses the internal mechanism. Behavioral ranking metrics can show that a prompt is harmful, but they cannot explain how role-prompt information affects the model internally. I therefore use activation patching, attention-head patching, helpful-head selection, single-head zero ablation, and top-20 group ablation to trace where role information is encoded and how it is routed toward the final Yes/No relevance decision. The mechanistic experiments use the same Random20 prompt set, rather than the newly constructed worst adjective–adverb prompts. This keeps the mechanistic analysis

aligned with the initial Random20 evaluation while allowing the behavioral experiments to separately identify harmful prompt components. This thesis addresses four research questions:

RQ1: How do isolated negative adjectives, isolated negative adverbs, and adjective–adverb combinations differ in their effects on zero-shot ranking performance?

RQ2: To what extent do harmful adjective–adverb combinations transfer across retrieval datasets?

RQ3: What role does semantic role identity play by itself and when combined with negative lexical framing?

RQ4: How is role-prompt information represented and routed through internal transformer components toward the final Yes/No relevance decision?

To answer these questions, I evaluate role-prompt effects across six retrieval datasets: TREC-DL 2019, TREC-DL 2020, TREC-COVID, SciFact, FIQA, and Climate-FEVER. Ranking effectiveness is measured using NDCG@10. The mechanistic interpretability experiments use label-aware logit-difference recovery, single-head zero ablation, and top-20 group ablation with a role-conditioned correct score to identify where role-conditioned information is represented and which attention heads causally affect the final decision.

Contributions. This thesis makes four main contributions.
First, I replicate the role-prompt ranking setup of Wang et al. using the Random20 prompt set constructed in this thesis, and confirm that negatively framed role prompts can cause substantial degradation in zero-shot LLM ranking.
Second, I show that the strongest harmful effects are compositional: adjective–adverb interactions produce stronger degradation than isolated negative adjectives or adverbs.
Third, I show that role identity alone is not sufficient to explain harmful ranking behavior. On TREC-COVID, meaningful, unrelated, object-like, and meaningless role-only prompts remain relatively stable and do not cause severe degradation. However, when role identity is combined with harmful adjective–adverb framing, it can modulate or amplify the degradation. This suggests that role semantics is secondary to lexical framing rather than the main source of harmfulness.
Fourth, I provide a mechanistic explanation of role-playing effects in zero-shot LLM rankers. Activation patching shows that role information is recoverable near role-related prompt positions and later becomes concentrated at the final decision token. Attention-head patching and zero ablation further show that this process involves a sparse set of decision-relevant heads. However, these heads are not all helpful in the same way: some support the correct label-aware direction, while others suppress or regulate it. This reveals partially distinct internal pathways for relevance and irrelevance judgments. Group zero ablation with the role-conditioned correct score further reveals condition-specific position-level effects. Positive-role rejection of irrelevant documents depends strongly on the last-token group. For negative roles, ablating the `Role-All`-selected group weakens the harmful role-induced `Yes` tendency on irrelevant documents, while the broader effects of last-token ablation are consistent with final decision integration.

Chapter 2 Background and Related Work

In recent years, large-scale pre-trained language models (LLMs) have made significant progress in natural language processing and have increasingly been applied to information retrieval (IR). Unlike traditional learning-to-rank methods, LLMs can perform document relevance assessment in a zero-shot setting, meaning that they do not require task-specific supervised fine-tuning. This makes prompt design especially important.

2.1 Zero-Shot LLM Ranking Methods

Existing research has shown that competitive ranking results can be achieved by presenting queries and candidate documents to an LLM in natural language and asking it to judge relevance [17, 14]. Compared with traditional ranking models, these methods rely on the knowledge and reasoning capabilities learned during pre-training rather than on dataset-specific relevance supervision.

Zero-shot LLM ranking methods can be formulated in several different paradigms. In the pointwise paradigm, the model receives one query–document pair and predicts a relevance label or relevance score for that document [23]. In the pairwise paradigm, the model compares two candidate documents and decides which one is more relevant to the query [15]. In the listwise paradigm, the model receives a list of candidate documents and is asked to produce a ranked ordering [18]. In the setwise paradigm, the model compares a set of candidate documents at each step, improving the efficiency of zero-shot LLM ranking compared with exhaustive pairwise comparisons [24].

This thesis focuses on the pointwise ranking paradigm. This choice follows the experimental setting of Wang et al. (2025), where the LLM receives a query and a candidate passage and answers whether the passage is relevant [21]. The pointwise setup is also suitable for mechanistic analysis because each query–document pair produces a single Yes/No decision, making it possible to compare the internal activations leading to relevance and irrelevance judgments.

2.2 Role-Playing and Prompt Strategies

Prompt wording, instruction style, and input structure can substantially affect ranking performance. Sun et al. (2025) systematically studied prompt variation in zero-shot LLM ranking and showed that minor changes in prompt formulation can lead to substantial differences in effectiveness [17]. This finding highlights the central importance of prompt design in zero-shot ranking systems.

Among prompt design strategies, role-playing has received increasing attention. In this setting, the model is asked to adopt a particular role, identity, or behavioral style before producing a judgment. Previous work has shown that role-playing can influence not only output style, but also task performance and reasoning behavior [7].

Wang et al. (2025) analyzed the effects of role-playing prompts on zero-shot LLM ranking and showed that some role prompts improve ranking effectiveness, whereas others cause substantial degradation [21]. In particular, their study identified positive and negative adjective and adverb candidates associated with helpful and harmful role-prompt behavior, and several negatively framed prompts produced severe ranking failures across multiple datasets.

Existing work shows that role-playing prompts can strongly affect zero-shot LLM ranking, but it does not fully explain which parts of a role prompt are responsible for harmful behavior. In particular, it remains unclear whether degradation is driven mainly by negative lexical wording, by the semantic identity of the assigned role, or by interactions between these components. This thesis builds on this gap by separating lexical polarity from role identity and later analyzing how these prompt-level effects are represented inside the model.

2.3 Interpretability and Mechanistic Analysis of Ranking Models

To better understand ranking behavior in LLMs, recent interpretability work has moved from output-level comparisons to the analysis of internal model representations. Mechanistic interpretability studies aim to explain model behavior by analyzing internal representations, residual streams, attention heads, and causal circuits inside transformer models [5, 10]. Within information retrieval, causal interventions have also been used to reverse engineer relevance-related computations in neural retrieval models [1]. Recent work has also examined methodological best practices for activation patching in information retrieval, emphasizing careful intervention design and metric interpretation [13]. Liu et al. (2025) studied how LLMs compute relevance from a mechanistic interpretability perspective and showed that activation patching can identify how query, document, instruction, and final-token representations contribute to relevance judgments [8]. Their subsequent attention-head ablation experiments further demonstrated that components selected through patching can have causal effects on downstream relevance assessment.
Tools such as *TransformerLens* make it possible to inspect, cache, and intervene on internal activations in transformer language models [11]. TransformerLens has been widely used as an implementation framework for mechanistic interpretability studies, including work on activation patching, circuit discovery, and attention-head analysis. For example, prior work uses activation patching to localize behaviorally relevant internal representations and studies how methodological choices affect the reliability of patching results [22, 6]. In the ranking setting, Liu et al. (2025) use activation patching and attention-head ablation to analyze how query, document, instruction, and final-token representations contribute to relevance judgments [8]. These studies motivate the use of TransformerLens-style interventions in this thesis to trace how role-prompt information is represented and routed toward the final Yes/No relevance decision.
Building on these tools, I perform mechanistic analysis using the Random20 prompts. Importantly, the mechanistic experiments are not conducted on the newly constructed worst adjective–adverb combinations. Instead, the full Random20 prompt set is retained during patching in order to analyze the same prompt family used in the initial ranking evaluation. This separation is important because the lexical ablation experiments are designed for controlled behavioral analysis, whereas the patching experiments are intended to explain the original role-playing mechanism itself.

Chapter 3 Data

I evaluate prompt effects across six benchmark datasets covering scientific verification, biomedical retrieval, open-domain web search, financial question answering, and climate claim verification.

Table 1: Official relevance-label distribution for TREC-DL 2019.

Label	Meaning	Count
0	Not relevant	5158
1	Relevant	1601
2	Highly relevant	1804
3	Perfectly relevant	697

Table 2: Official relevance-label distribution for TREC-DL 2020.

Label	Meaning	Count
0	Not relevant	7780
1	Relevant	1940
2	Highly relevant	1020
3	Perfectly relevant	646

All datasets are used under their official evaluation protocols. Where applicable, I access the collections through the BEIR framework [19]. I report NDCG@10 using the official relevance judgments (qrels) provided for each dataset.
The datasets differ substantially in corpus size and relevance scale. TREC-DL evaluates ranking over 8.8 million passages, TREC-COVID contains 171 thousand scientific articles, and the remaining BEIR datasets contain at most tens of thousands of documents. These differences are important when interpreting cross-dataset comparisons.

3.1 TREC-DL datasets

TREC-DL 2019 TREC Deep Learning 2019 (DL19) [3] is based on the MS MARCO Passage Ranking collection. The evaluation set contains 43 official queries and a corpus of 8,841,823 passages.
Relevance judgments follow a four-level graded scale consisting of not relevant, relevant, highly relevant, and perfectly relevant labels.
Table 1 summarizes the official relevance-label distribution.
The average number of positive passages per query is approximately 95.4.

TREC-DL 2020 TREC Deep Learning 2020 (DL20) [2] extends DL19 with 54 evaluation queries while using the same MS MARCO corpus of 8,841,823 passages.
The relevance scale is identical to DL19. Table 2 summarizes the official label distribution.
The average number of positive passages per query is approximately 66.8.

3.2 BEIR benchmarks

I use the following datasets from the BEIR benchmark.

TREC-COVID TREC-COVID [16] is a biomedical retrieval benchmark built on the CORD-19 collection.

Table 3: Summary statistics of datasets used in this thesis.

Dataset	Queries	Documents	Relevance Scale	Domain
TREC-DL 2019	43	8.8M	0–3	Web Retrieval
TREC-DL 2020	54	8.8M	0–3	Web Retrieval
TREC-COVID	50	171K	0–2	Biomedical Retrieval
SciFact	300	5.1K	Binary	Scientific Verification
FIQA-2018	648	57K	0–2	Financial Retrieval
Climate-Fever	1535	5.4K	Binary	Climate Verification

The dataset contains 50 expert-formulated queries and a corpus of 171,332 scientific articles. Relevance judgments follow a three-level graded scale consisting of not relevant, partially relevant, and highly relevant labels.
TREC-COVID serves as the primary dataset for the lexical ablation experiments because it provides a relatively stable biomedical retrieval setting and manageable evaluation size.

SciFact SciFact [20] is a scientific claim verification dataset containing 300 claims and 5,183 scientific abstracts.
The original labels consist of SUPPORTS, REFUTES, and NOT ENOUGH INFO. For retrieval evaluation, SUPPORTS and REFUTES are treated as relevant, resulting in a binary relevance setting.

FiQA-2018 FiQA-2018 [9] is a financial question answering and retrieval dataset containing 648 queries and 57,638 documents.
Relevance judgments follow a three-level graded scale consisting of not relevant, relevant, and highly relevant labels.

Climate-Fever Climate-Fever [4] is a climate claim verification dataset consisting of 1,535 claims and 5,416 Wikipedia paragraphs.
The dataset uses a binary relevance setting consisting of relevant and not relevant labels.
The datasets intentionally cover multiple retrieval domains, including web search, biomedical retrieval, scientific verification, financial retrieval, and climate claim verification.
This diversity is important because prompt sensitivity may depend strongly on the retrieval domain and the structure of relevance judgments.
TREC-DL represents large-scale open-domain ranking, while SciFact and Climate-Fever emphasize factual consistency and evidence verification. FIQA focuses on financial question answering, whereas TREC-COVID evaluates biomedical retrieval under expert-defined information needs.
This broad evaluation setting allows us to study both the stability and transferability of harmful prompt effects across domains. Table 3 summarizes the dataset statistics.

3.3 Dataset Usage Across Experimental Stages

Although the thesis introduces all six datasets in this chapter, not every dataset is used in the same way in every experiment. Table 4 summarizes how each dataset is used across the behavioral and mechanistic stages.

Table 4: Dataset usage across experimental stages.

Experiment stage	DL19	DL20	COVID	SciFact	FIQA	Climate-Fever
Random20 replication	✓	✓	✓	–	–	–
Lexical ablation	–	–	✓	–	–	–
Harmful adjective–adverb pair transfer	✓	✓	✓	✓	✓	✓
Role-only control	–	–	✓	–	–	–
Fixed-lexical role-identity analysis	–	–	✓	✓	✓	✓
Role–lexical interaction analysis	–	–	✓	✓	✓	✓
Activation patching	✓	✓	✓	✓	✓	✓
Single-head zero ablation	✓	✓	✓	✓	✓	✓
Top-20 group ablation	✓	✓	✓	✓	✓	✓

This distinction is important for interpreting the experimental design. TREC-COVID is used as the main source dataset for controlled lexical ablation because it has a manageable query set and a stable biomedical retrieval setting. The remaining datasets are then used to test whether the harmful prompt effects and mechanistic patterns transfer across domains.

Chapter 4 Methods

This thesis uses a zero-shot pointwise ranking framework to analyze how harmful prompt effects arise from lexical polarity and role identity. The overall methodology consists of two connected stages: behavioral ranking experiments and mechanistic interpretability analysis. The behavioral and mechanistic stages use Random20 in different ways. The initial ranking evaluation includes all ten positive and ten negative Random20 prompts. The subsequent behavioral decomposition focuses on negative prompts and newly constructed harmful variants, because its purpose is to determine how severe the degradation can become and which lexical and role components are responsible. In contrast, the mechanistic analyses use the full Random20 set, including both positive and negative prompts, to examine how role-prompt information is represented, routed, and integrated into the final Yes/No decision.

4.1 Experimental Setup and Ranking Framework

Progressive experimental design. The experiments in this thesis are designed progressively. In the behavioral decomposition stage, rather than treating complete negative role prompts as indivisible units, I systematically decompose harmful prompt behavior into smaller components and analyze them in isolation before studying their interaction. The overall behavioral pipeline follows six stages: replication of the role-prompt ranking experiment using Random20, lexical polarity ablation, cross-domain transfer of harmful lexical pairs, role-only control, fixed-lexical role-identity analysis, and role–lexical interaction analysis.

The mechanistic pipeline then uses residual-stream and attention-head activation patching, head selection based on recovery scores, single-head zero ablation of the selected top-10 relevance and irrelevance head sets, and top-20 group ablation to test how the Random20 role-prompt effects are internally represented and causally routed. The group-ablation stage uses a role-conditioned correct score to distinguish ordinary relevance ac-

curacy from agreement with the answer direction induced by positive or negative roles. This progression is important because each stage motivates the next one. The Random20 evaluation first shows that positively and negatively framed role prompts can produce substantially different ranking outcomes, with several negative prompts causing severe degradation. The lexical ablation experiments then test whether these failures originate from specific adjective–adverb combinations. The role-only control tests whether role identity alone is sufficient to cause degradation. The fixed-lexical role-identity experiment then asks whether role semantics can modulate degradation when harmful lexical framing is held constant. Finally, activation patching and ablation experiments examine how role-prompt information across the full Random20 set is represented and routed inside the transformer.

Zero-shot pointwise ranking setup. I use a zero-shot pointwise ranking setup. For each query–document pair, the model independently judges whether the document is relevant to the query.
All experiments use `meta-llama/Llama-3.1-8B-Instruct` without task-specific fine-tuning. The scoring method follows the pointwise Yes/No relevance-judgment formulation used in Wang et al. [21]. In this formulation, the model receives one query–document pair at a time and is asked whether the document is relevant to the query. This is also consistent with common pointwise LLM ranking practice, where the model generates a binary relevance answer and the likelihood of `Yes` is used as the ranking score.
For each candidate document d and query q, I extract the first-token logits of the answer tokens `Yes` and `No`:

$$\ell_{\text{Yes}} = \text{logit}(\text{Yes}), \qquad \ell_{\text{No}} = \text{logit}(\text{No}).$$

The document score is computed as the two-way normalized probability assigned to `Yes`:

$$s(q, d) = \frac{\exp(\ell_{\text{Yes}})}{\exp(\ell_{\text{Yes}}) + \exp(\ell_{\text{No}})}.$$

Candidate documents are ranked in descending order of $s(q, d)$. The same candidate set is kept fixed across prompt variants within each dataset, so differences in NDCG@10 reflect changes caused by the prompt rather than by candidate retrieval.
Ranking effectiveness is evaluated using NDCG@10 with the official qrels for each dataset. Logit differences are used later for mechanistic analysis, but the standard ranking experiments use the Yes probability defined above.

Replication of the role-prompt ranking experiment using Random20. Before introducing the controlled behavioral experiments, I first replicate the role-prompt ranking setup of Wang et al. [21] using a set of 20 prompts constructed for this thesis, termed Random20.
Random20 is formed using adjective and adverb candidates selected from Wang et al.'s study. These candidates are randomly paired within their respective polarity categories to produce ten positive and ten negative role prompts. A separate no-role prompt is used as the baseline and is not counted as part of Random20.
This initial experiment replicates the role-prompt ranking results of Wang et al. [21] using the same adjective and adverb candidates, model, scoring method, and evaluation metric. Random20 is evaluated on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The

results confirm the general trend reported in prior work: positively framed role prompts tend to maintain or improve ranking effectiveness, whereas several negatively framed prompts cause substantial degradation.
This replication serves as the empirical starting point for the remainder of the thesis. After establishing that the harmful role-prompt effect can be reproduced qualitatively, I conduct controlled experiments to identify which lexical and role components are responsible for the degradation.

4.2 Behavioral Role-Component Analysis

Prompt families and role identities. The behavioral experiments use four related prompt families.
First, Random20 is used for the initial ranking evaluation and for all mechanistic interpretability experiments. The set contains ten positive and ten negative role prompts, while a separate no-role prompt is used as the baseline. The mechanistic analyses use all 20 Random20 prompts unless a specific subset is explicitly stated. It provides the common reference point between the reproduced ranking behavior and the internal activation-patching analysis.
Second, controlled lexical prompts are constructed to isolate the effects of adjectives and adverbs. These prompts are used in the behavioral decomposition experiments. Starting from the same pointwise ranking setup, I evaluate three lexical variants on TREC-COVID: adjective-only prompts, adverb-only prompts, and adjective–adverb paired prompts. This experiment tests whether the performance degradation is mainly caused by negative adjectives, negative adverbs, or by their interaction when both are present in the same prompt.
Third, role-only prompts are constructed to test whether role identity alone is sufficient to cause large ranking changes. These prompts contain a role phrase but remove the negative adjective and adverb components. They include domain-specific medical roles, non-domain semantic occupations, unrelated semantic roles, object-like unusual roles, and meaningless string roles.
Fourth, selected role–lexical prompts are constructed through a two-stage selection procedure. After the adjective-only, adverb-only, and adjective–adverb paired ranking experiment, I use the COVID results to identify the ten lowest-performing adjective–adverb combinations. These ten harmful lexical pairs are then crossed with selected role identities, producing a role-by-lexical grid. The lowest-performing role–adjective–adverb combinations from this grid are used in the cross-dataset behavioral experiments. This procedure ensures that the harmful role–lexical prompts are selected empirically rather than manually.
The selected role identities cover non-expert, service, retail, entertainment, assistant-like, and socially grounded occupations that are semantically mismatched with the ranking task. This allows the experiments to test whether the assigned role contributes to degradation beyond negative lexical wording alone.
The general role-prompt template can be written as:

```
You are a/an [ADJ] [ROLE] that [MODAL] rank passages [ADV],
based on their relevance to a query.
```

Here, [MODAL] denotes the modal verb used in the role prompt, such as *will*, *shall*, *can*, *could*, or *might*. In the adjective-only condition, [ADJ] is varied while the adverb component is removed. In the adverb-only condition, [ADV] is varied while the adjective component is removed. In the adjective–adverb paired condition, both [ADJ] and [ADV] are present simultaneously. In the fixed-lexical role-identity condition, the adjective and adverb are fixed, while only the [ROLE] identity is varied. In the implementation used for the main experiments, the role prompt is placed at the beginning of the input. The default adjusted pointwise prompt has the following structure:

[ROLE] + instruction + query + document + output format.

The output format asks the model to answer with Yes or No. This fixed prompt structure is kept constant across role conditions so that performance differences can be attributed to role-prompt variation rather than to changes in the base ranking instruction.

Lexical polarity ablation. To investigate which lexical components drive harmful prompt effects, I construct three controlled prompt settings: adjective-only prompts, adverb-only prompts, and adjective–adverb paired prompts:

- For adjective-only prompts, the role description contains a negative adjective while the adverb component is removed.
- For adverb-only prompts, the role description remains neutral while the ranking instruction contains a negative adverb.
- For the paired setting, both negative adjectives and negative adverbs are included simultaneously.

I select ten negative adjectives and ten negative adverbs. All pairwise combinations are generated, resulting in 100 distinct adjective–adverb prompts. These experiments are conducted primarily on TREC-COVID for the selection stage, following the source-dataset rationale described above.

Cross-domain transfer of harmful lexical pairs. After evaluating all 100 adjective–adverb combinations on TREC-COVID, I sort the combinations by NDCG@10 in ascending order and select the ten lowest-scoring combinations as harmful lexical pairs.
TREC-COVID is therefore the source dataset for selecting the harmful pairs. The selected pairs are then evaluated on five additional target datasets: TREC-DL 2019, TREC-DL 2020, SciFact, FIQA, and Climate-Fever. The original TREC-COVID results are retained as the source-dataset reference.
This stage evaluates whether harmful lexical interactions generalize across retrieval domains rather than remaining specific to the dataset on which they were selected.

Role-only control. Before evaluating role identity under harmful lexical framing, I also run a role-only control experiment on TREC-COVID. In this setting, the prompt contains a role phrase but no negative adjective or adverb. I compare domain-specific medical roles, non-domain semantic occupations, unusual object-like roles, and meaningless string roles. This control tests whether role identity alone is sufficient to cause large ranking changes, or whether harmful behavior mainly requires negative lexical framing.

Role identity and role–lexical interaction. After the role-only control, I isolate semantic role effects under harmful lexical framing by fixing the adjective *incorrect* and the adverb *falsely*. I then vary only the role identity itself. This means that every prompt shares the same negative lexical framing, and any remaining performance difference can be attributed to the semantic identity of the role.
Twenty alternative role identities are evaluated under this fixed lexical framing. These roles are selected to cover non-expert, service, retail, entertainment, assistant-like, and socially grounded occupations that are semantically mismatched with the ranking task. This design allows me to test whether role semantics independently contributes to harmful ranking behavior even when lexical negativity is held constant.
Finally, I combine selected role identities with the selected harmful adjective–adverb pairs. This interaction analysis tests whether role semantics amplifies or mitigates the harmful effects of lexical polarity. The resulting prompts are evaluated across multiple datasets to analyze cross-domain consistency.

Evaluation of harmfulness. For the transfer and interaction analyses, harmfulness is measured as the absolute performance drop relative to the baseline:

$$\text{Drop} = \text{NDCG@10}_{\text{baseline}} - \text{NDCG@10}_{\text{prompt}}.$$

Under this definition, larger positive values indicate stronger harmful effects, whereas negative values indicate improvements over the baseline.
For the role-identity ablation experiments, the reference baseline is slightly different. Instead of the no-role baseline, I use the fixed adjective–adverb prompt itself as the lexical baseline:

$$\text{Drop}(r) = \text{NDCG@10}_{\text{adj+adv baseline}} - \text{NDCG@10}_{r+\text{adj+adv}}.$$

This allows me to isolate the additional contribution introduced by role identity alone.

4.3 Mechanistic Interpretability Analysis

Mechanistic setup. To explain the behavioral effects identified in the preceding experiments, I use activation patching and attention-head interventions to examine how differences between positive and negative role prompts are represented and causally routed inside the model.
Activation patching is a causal intervention technique used to test whether a specific internal activation contributes to a model behavior [6, 22, 8]. The basic idea is to compare two forward passes: a *clean* run, which represents the behavior to be recovered, and a *corrupted* run, which represents the alternative or degraded prompt condition. After both prompts are passed through the transformer, an activation from a selected layer, attention head, or token position in the clean run is inserted into the corresponding location of the corrupted run. If this replacement shifts the corrupted output toward the clean output, the patched activation is interpreted as causally contributing to the behavioral difference between the two conditions.
As illustrated in Figure 1, each patching experiment compares a positive role prompt with its corresponding negative role prompt. The positive-role prompt is used for the clean run, which provides the reference activations and output behavior, whereas the matched negative-role prompt is used for the corrupted run, whose degraded behavior is the target of the intervention.

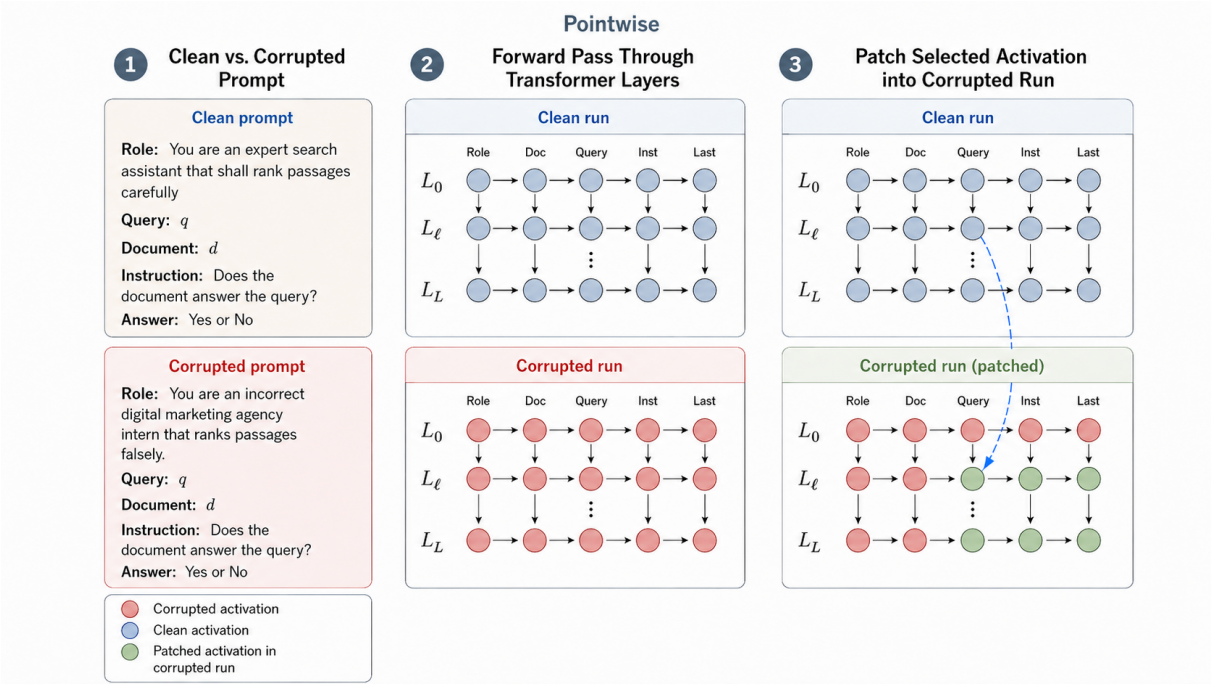


Figure 1: Overview of the activation-patching procedure used in the mechanistic analysis. A positive-role prompt is used for the clean run, and its corresponding negative-role prompt is used for the corrupted run. A selected activation from the clean run is then inserted into the corresponding location of the corrupted run. The resulting change in the label-aware `Yes`/`No` logit difference measures the extent to which the intervention recovers the clean positive-role behavior.

I patch either residual-stream activations or attention-head outputs at selected prompt positions, including the role phrase, instruction segment, query segment, document segment, and final token. For each intervention, an activation from the clean positive-role run is inserted into the corresponding location of the corrupted negative-role run. The effect is measured by the extent to which the patched run recovers the clean–corrupted difference in the label-aware logit difference. This makes it possible to identify where differences between positive and negative role conditions are represented and which internal components causally route these differences toward the final decision. Importantly, the mechanistic experiments use the Random20 prompts rather than the newly constructed worst adjective–adverb combinations. This design choice ensures that the interpretability analysis directly explains the same role-playing mechanism used in the reproduced Wang et al. experiments.

The analysis is performed using the same pointwise Yes/No ranking framework as the behavioral experiments. However, instead of ranking documents by Yes probability, the mechanistic experiments measure changes in label-aware logit difference.

For relevant examples, the target direction is:

$$\mathrm{LD}_{rel} = \mathrm{logit}(\mathrm{Yes}) - \mathrm{logit}(\mathrm{No}).$$

For irrelevant examples, the target direction is:

$$\mathrm{LD}_{irrel} = \mathrm{logit}(\mathrm{No}) - \mathrm{logit}(\mathrm{Yes}).$$

This label-aware formulation prevents opposite decision directions from being mixed together.
Following established activation-patching practice, I compute normalized logit-difference recovery from the clean, corrupted, and patched runs [6, 22]:

$$\text{LD Recovery} = \frac{\text{LD}_{\text{patched}} - \text{LD}_{\text{corrupted}}}{\text{LD}_{\text{clean}} - \text{LD}_{\text{corrupted}}}.$$

A higher positive recovery value indicates that the patched activation restores more of the clean behavior.

Residual stream patching. For residual-stream analysis, I patch `resid_pre` and `resid_post` activations. Here, `resid_pre` denotes the residual-stream representation entering a transformer block, whereas `resid_post` denotes the representation after the block's attention and MLP updates have been added. Comparing the two therefore helps distinguish where information is present before a layer's computation and how it changes after that layer has processed the input. This provides a component-level view of where role-conditioned information is represented across the transformer stack.
For each layer and target prompt position, I replace the activation in the corrupted run with the corresponding activation from the clean run. The target positions include `Role-All`, `Role-Adj`, `Role-Adv`, `Role-Adj-Adv`, `Instruction`, `Query`, `Document`, and the `Last` token.
The effect of each intervention is measured using the normalized label-aware logit-difference recovery defined above. This analysis identifies where positive–negative role differences are recoverable in the residual stream and whether these differences are later consolidated at the final prediction position.

Attention head patching and zero ablation. For head-level analysis, I patch the attention-head output $z^{(\ell,h)}$, where ℓ denotes the layer and h denotes the attention head. Compared with residual-stream patching, this allows me to identify which specific heads causally route role-prompt information.
The head-level patching experiments use the same clean-to-corrupted intervention as the residual-stream experiments. For each selected prompt position, the clean activation of an attention head is inserted into the corrupted run, and recovery is measured using the same normalized logit-difference recovery metric.
In addition to activation patching, I perform targeted zero-ablation analysis, following prior work that uses ablation to test whether components identified through patching are causally necessary for relevance-related behavior [8, 21]. For a selected attention head (ℓ, h), the head output at a target position is replaced with zero during inference. I then compare the ablated output with the clean output.
For relevance examples, the logit difference is:

$$\text{LD}_{rel} = \text{logit(Yes)} - \text{logit(No)}.$$

For irrelevance examples, the direction is reversed:

$$\text{LD}_{irrel} = \text{logit(No)} - \text{logit(Yes)}.$$

The ablation delta is reported as:

$$\Delta = \mathrm{LD}_{\mathrm{ablated}} - \mathrm{LD}_{\mathrm{clean}}.$$

Under this definition, a negative Δ means that ablating the head weakens the correct decision direction. Such a head can therefore be interpreted as supporting the target behavior. A positive Δ means that ablating the head increases the target logit direction, suggesting that the head may suppress the target behavior under normal inference.
For the causal intervention stage, I use two levels of ablation. First, I perform single-head zero ablation on the final top-10 relevance heads and top-10 irrelevance heads selected from the activation-patching rankings. These single-head results are averaged over all 20 Random20 role prompts, so that the results are not dependent on one specific positive–negative role pair.
Second, I perform group ablation on the broader top-20 candidate heads selected for each dataset, relevance direction, and target position. The resulting interventions are evaluated separately for positive and negative role groups.
The main-text analysis uses zero ablation because directly removing the selected head outputs produces the clearest causal contrast with the intact model. Mean ablation is retained as a supplementary robustness analysis and is reported in Appendix E.2. Both interventions use the same top-20 head groups selected from the activation-patching rankings.
Importantly, labels such as `Role-All`, `Role-AdjAdv`, `Inst`, and `Last` describe the target position used for head selection and intervention. They do not imply that every selected head has one fixed semantic function.
The same head indices can produce different activation patterns under positive and negative role prompts, and each top-20 group may contain both supporting and suppressive components. The observed group-ablation effect should therefore be interpreted as the net causal contribution of the selected group in a particular role and relevance context.
The purpose of this experiment is to test whether the selected heads support the answer direction associated with the current role condition. I therefore use a role-conditioned correct score for the group-ablation analysis. For positive roles, the role-conditioned target direction follows the ground-truth relevance label: relevant passages are associated with `Yes`, and irrelevant passages are associated with `No`. For negative roles, the analysis reverses this direction in order to measure agreement with the harmful role-induced behavior: relevant passages are associated with `No`, and irrelevant passages are associated with `Yes`.
Let $\mathrm{LD}_{\mathrm{std}}$ denote the standard label-aware logit difference. For relevant passages, this is logit(Yes) − logit(No), and for irrelevant passages, this is logit(No) − logit(Yes). I then define a role sign s, where $s = +1$ for positive roles and $s = -1$ for negative roles:

$$\mathrm{LD}_{\mathrm{role}} = s \cdot \mathrm{LD}_{\mathrm{std}}.$$

The role-conditioned correct score is computed as:

$$\mathrm{CorrectScore} = \sigma(\mathrm{LD}_{\mathrm{role}}),$$

where σ is the sigmoid function.
A higher score indicates stronger support for the role-conditioned target direction. For positive roles, this corresponds to stronger support for the ground-truth relevance judgment. For negative roles, it corresponds to stronger support for the reversed, role-induced

answer direction. Therefore, for negative roles, a high role-conditioned correct score should not be interpreted as high relevance accuracy; it instead indicates stronger agreement with the harmful behavior induced by the negative role prompt.
The group-ablation analysis compares the intact and ablated role-conditioned correct scores. Here, the intact score is obtained from the unablated forward pass under the same role prompt and query–document input. A decrease after ablation means that the selected head group contributes positively to the role-conditioned target direction in that condition. An increase means that removing the group strengthens that direction.
Because each intervention removes twenty functionally mixed heads simultaneously, the observed change should be interpreted as the net effect of the group. An increase should not automatically be interpreted as evidence that every selected head suppresses the role-conditioned target direction.

Chapter 5 Behavioral Analysis: What Makes Negative Role Prompts Harmful?

This chapter analyzes harmful role-playing behavior at the behavioral level. The goal is to determine which prompt components are responsible for ranking degradation with negative roles before studying the internal model mechanism. I therefore decompose negative role prompts into lexical polarity, role identity, and their interaction. This chapter answers the first three research questions. It first tests whether harmfulness is mainly caused by adjective–adverb combinations and whether these combinations transfer across datasets. It then examines RQ3 by asking whether role identity alone is sufficient to cause degradation, and whether it modulates harmful effects under negative lexical framing.

5.1 Behavioral Experimental Setup

I evaluate ranking effectiveness using NDCG@10. All behavioral experiments use the zero-shot pointwise Yes/No ranking setup described in Section 4.1.
The behavioral analysis is organized progressively. First, I replicate the role-prompt ranking setup of Wang et al. using Random20 on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. Second, I conduct controlled lexical ablation on TREC-COVID to compare adjective-only, adverb-only, and adjective–adverb paired prompts. Third, I transfer the harmful adjective–adverb pairs selected on TREC-COVID to other datasets. Fourth, I run a role-only control on TREC-COVID to test whether role identity alone causes large ranking changes. Fifth, I isolate role-identity effects under fixed negative lexical framing by varying only the role identity. Finally, I combine harmful lexical cues with selected role identities to test whether lexical polarity and role semantics reinforce each other.
TREC-COVID is used as the main source dataset for controlled lexical analysis because it provides a stable biomedical retrieval setting and a manageable query set. The remaining datasets are used to evaluate whether the observed harmful prompt effects transfer across retrieval domains.

Table 5: Results of the role-prompt ranking replication using Random20 on the three core datasets. Values are NDCG@10. Role0 is the no-role baseline. Roles 1–10 are positive role prompts, and roles 11–20 are negative role prompts.

Group	Role	Prompt	TREC-DL 2019	TREC-DL 2020	TREC-COVID
Baseline	role0	no role	0.6464	0.6025	0.7287
Positive	role1	expert/carefully	0.6641	0.6316	0.7854
	role2	reliable/accurately	0.6751	0.6432	0.7807
	role3	superb/swiftly	0.6652	0.6367	0.7870
	role4	logical/clearly	0.6734	0.6512	0.7716
	role5	capable/firmly	0.6622	0.6372	0.7882
	role6	brilliant/wisely	0.6678	0.6369	0.7865
	role7	focused/carefully	0.6657	0.6341	0.7838
	role8	able/nicely	0.6623	0.6309	0.7862
	role9	knowledgeable/perfectly	0.6566	0.6377	0.7920
	role10	strong/sharply	0.6729	0.6358	0.7878
Negative	role11	faulty/badly	0.5467	0.5656	0.7703
	role12	confused/wrongly	0.6377	0.6030	0.7878
	role13	clumsy/mistakenly	0.5484	0.5405	0.7748
	role14	messy/improperly	0.6574	0.6132	0.7927
	role15	incorrect/falsely	0.1880	0.2017	0.6005
	role16	awful/terribly	0.2530	0.2243	0.6071
	role17	unreliable/poorly	0.5830	0.5770	0.7651
	role18	problematic/incorrectly	0.3988	0.4293	0.7292
	role19	hopeless/sadly	0.6448	0.6241	0.7886
	role20	flawed/horribly	0.6228	0.6112	0.7951

5.2 Reproduction of Random20 Ranking Results

Before constructing new prompt ablations, I first replicate the role-prompt ranking experiment of Wang et al. [21] using Random20. The replication is conducted on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. These datasets are used to examine whether the main qualitative role-prompt pattern reported in prior work can also be observed with the Random20 prompts constructed for this thesis.

Table 5 shows that the Random20 results reproduce the broad qualitative pattern reported by Wang et al. in their Figure 2 [21]. In both studies, ranking effectiveness varies substantially across role-play wordings: positively framed roles generally maintain or improve performance relative to the no-role baseline, whereas some negatively framed roles cause large reductions in NDCG@10. Positive role prompts generally improve over the no-role baseline on all three datasets. Most positive roles obtain higher NDCG@10 than role0 on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. These results support the observation that positively framed adjective–adverb combinations can improve or preserve ranking effectiveness under the present pointwise ranking setup.

In contrast, several negative role prompts produce large performance drops. The strongest harmful case is role15, which combines the adjective–adverb pair *incorrect/falsely*. It causes severe degradation on all three datasets, especially on TREC-DL 2019 and TREC-DL 2020. Role16, with *awful/terribly*, also produces substantial degradation across the three datasets. Role18, with *problematic/incorrectly*, strongly degrades TREC-DL 2019

and TREC-DL 2020, although its effect on TREC-COVID is much weaker. These results suggest that harmful role-prompt effects are closely related to negative lexical framing.
However, the table also shows that not all negative prompts are equally harmful. Some negative roles, such as role14, role19, and role20, remain close to or even above the no-role baseline on one or more datasets. This means that degradation cannot be explained by negative lexical wording *in general*. The observed variation may arise because some lexical items provide clearer or stronger task-relevant signals to the model than others, or because the effect depends on interactions among multiple prompt components.
Overall, the reproduction stage serves two purposes. First, it confirms, using Random20, that role-playing prompts can substantially change zero-shot ranking effectiveness. Second, it identifies a clear empirical pattern: the most harmful prompts in the reproduced setting often contain strong adjective–adverb lexical framing, especially combinations such as *incorrect/falsely*, *awful/terribly*, and *problematic/incorrectly*. This observation motivates the controlled lexical decomposition experiments introduced in the following section, where adjectives, adverbs, adjective–adverb pairs, and role identities are analyzed separately.

5.3 Lexical Polarity and Cross-Domain Transfer

Lexical ablation on TREC-COVID. After reproducing the Random20 ranking behavior, I next investigate which lexical components are responsible for the strongest negative effects. Although all negative Random20 prompts contain both a negative adjective and a negative adverb, their effects vary substantially: some remain close to the no-role baseline, whereas others cause severe degradation. However, the Random20 prompts do not allow us to determine whether this variation is caused by adjectives alone, adverbs alone, or their combination. This motivates a controlled lexical ablation study. I therefore compare three prompt settings: adjective-only prompts, adverb-only prompts, and adjective–adverb paired prompts, following the templates defined in Section 4.2. In the adjective-only setting, the adverb slot is removed; in the adverb-only setting, the adjective slot is removed; and in the paired setting, both components are retained. All other parts of the prompt remain unchanged. The experiments are conducted on TREC-COVID. For the paired setting, I generate all 100 possible adjective–adverb combinations. Figure 2 shows the resulting NDCG@10 distributions.
The adjective-only and adverb-only prompts both remain relatively stable and perform above the no-role baseline when used in isolation. Neither isolated component produces the strongest degradation observed in the experiment. This indicates that individual negative lexical cues, by themselves, are not sufficient to explain the most harmful ranking failures.
In contrast, the adjective–adverb paired prompts show a much wider distribution. This setting contains both relatively stable combinations and severely harmful combinations. Most importantly, the lowest NDCG@10 scores appear only in the paired condition. Several adjective–adverb combinations collapse dramatically below the baseline, whereas the isolated adjective-only and adverb-only settings do not show the same level of degradation. This result demonstrates that harmful prompt behavior is not driven simply by the presence of a single negative word. Instead, the strongest degradation emerges when negative adjectives and negative adverbs are combined within the same prompt. The main empirical contrast can therefore be summarized as:

$$\text{Adj} + \text{Adv} \gg \{\text{Adj}, \text{Adv}\}.$$

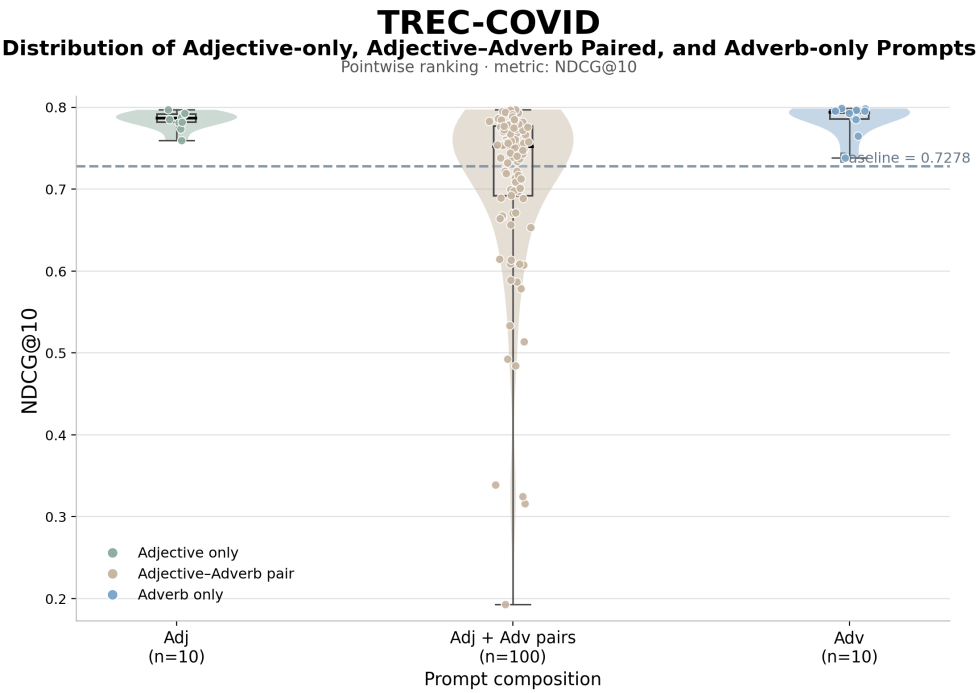


Figure 2: Distribution of ranking performance for adjective-only, adverb-only, and adjective–adverb paired prompts on TREC-COVID.

This means that the model does not respond to adjective and adverb polarity as two independent additive effects. Rather, certain adjective–adverb combinations create a much stronger harmful conditioning effect than either component produces alone.
Another important observation is the large variance within the paired setting itself. Not all negative adjective–adverb combinations are equally harmful. Some pairs remain close to the baseline, whereas others produce catastrophic degradation. This suggests that harmfulness depends on the specific interaction between adjective and adverb choices, rather than on lexical negativity in isolation.
This finding directly motivates the next step of the behavioral analysis. Because the paired condition contains the strongest harmful cases and also shows large internal variation, I select the lowest-performing adjective–adverb pairs from this distribution for cross-domain transfer experiments.
The lexical ablation results therefore provide the first major conclusion of this thesis: the strongest harmful prompt effects are compositionally structured. Negative ranking behavior emerges most strongly when multiple polarity-bearing lexical cues interact within the same prompt, which explains why the following analysis focuses on selecting and transferring the most harmful adjective–adverb pairs.

Selection of harmful adjective–adverb pairs. After evaluating all 100 adjective–adverb combinations on TREC-COVID, I select the worst-performing combinations for further analysis.
The selection is based on NDCG@10 ranking performance. I identify ten combinations from the low-performing tail of the distribution:

Pair ID	Adjective–Adverb Combination
1	awful falsely
2	hopeless wrongly
3	unreliable horribly
4	problematic poorly
5	incorrect improperly
6	messy incorrectly
7	clumsy terribly
8	confused mistakenly
9	faulty badly
10	flawed sadly

Note that these combinations are not selected manually but they emerge from the lexical ablation results as the most consistently harmful prompt constructions. Importantly, the strongest harmful combinations are not always those containing the most obviously negative individual words. Rather, the results suggest that interaction structure matters more than isolated lexical negativity. For example, strongly negative adjectives such as *hopeless* and *incorrect* remain relatively stable when paired with the weaker adverb *sadly*, achieving NDCG@10 scores of 0.7911 and 0.7831, respectively, on TREC-COVID. In contrast, the combination *awful falsely* produces catastrophic degradation, reducing NDCG@10 to 0.1930. The complete lexical-ablation results are reported in Appendix A.2, Table 12.
This further supports the hypothesis that harmfulness is fundamentally compositional rather than additive.

Cross-domain transfer of harmful lexical pairs. I next investigate whether the harmful lexical effects identified on TREC-COVID generalize to other retrieval domains. The ten selected adjective–adverb combinations are evaluated on TREC-DL 2019, TREC-DL 2020, SciFact, FIQA, and Climate-Fever, while the original TREC-COVID results are retained as the source-dataset reference. Table 6 reports the resulting NDCG@10 scores. Several patterns are immediately visible. First, all selected negative combinations perform substantially worse than the positive Random20 prompts across every dataset. Second, most combinations perform worse than the no-role baseline on several datasets, indicating that the harmful lexical effects identified on TREC-COVID often transfer beyond the source dataset, although not uniformly for every pair. Third, the degree of degradation varies substantially across datasets. The same prompt can be moderately harmful on one dataset and catastrophic on another.
The strongest and most consistent degradation is produced by:

awful falsely

This combination produces the lowest scores across nearly all datasets, reaching near-zero NDCG@10 on FIQA and Climate-FEVER and very low scores on DL19, DL20, and SciFact.
Figure 3 visualizes the corresponding cross-domain distributions. The results reveal several important trends. Open-domain web retrieval datasets such as DL19 and DL20 exhibit large absolute degradation, but factual verification datasets such as SciFact and Climate-Fever appear even more sensitive to harmful lexical framing. SciFact is particularly unstable: multiple combinations produce near-collapse behavior. By contrast, TREC-COVID

Table 6: Performance of the ten selected harmful adjective–adverb pairs across six datasets.

Group	Prompt / Pair	TREC-DL19	TREC-DL20	TREC-COVID	SciFact	FIQA	Climate-Fever
Baseline	role0	0.6464	0.6025	0.7287	0.5655	0.2583	0.0999
Positive	expert/carefully	0.6641	0.6316	0.7854	0.6494	0.2915	0.1310
Positive	reliable/accurately	0.6751	0.6432	0.7807	0.6549	0.2934	0.1227
Positive	superb/swiftly	0.6652	0.6367	0.7870	0.6479	0.2951	0.1241
Positive	logical/clearly	0.6734	0.6512	0.7716	0.6498	0.2890	0.1257
Positive	capable/firmly	0.6622	0.6372	0.7882	0.6571	0.2939	0.1298
Positive	brilliant/wisely	0.6678	0.6369	0.7865	0.6465	0.2929	0.1246
Positive	focused/carefully	0.6657	0.6341	0.7838	0.6545	0.2923	0.1256
Positive	able/nicely	0.6623	0.6309	0.7862	0.6480	0.2929	0.1278
Positive	knowledgeable/perfectly	0.6566	0.6377	0.7920	0.6508	0.2995	0.1301
Positive	strong/sharply	0.6729	0.6358	0.7878	0.6515	0.2914	0.1234
Negative	awful falsely	0.0767	0.0378	0.1930	0.0150	0.0049	0.0052
Negative	hopeless wrongly	0.1891	0.1771	0.5336	0.1017	0.0317	0.0139
Negative	unreliable horribly	0.3018	0.2795	0.6643	0.3498	0.0942	0.0415
Negative	problematic poorly	0.4010	0.4558	0.6927	0.5671	0.1742	0.0885
Negative	incorrect improperly	0.4107	0.4306	0.7219	0.3199	0.1230	0.0301
Negative	messy incorrectly	0.4958	0.5002	0.7541	0.4845	0.1818	0.0535
Negative	clumsy terribly	0.5565	0.5349	0.7577	0.5975	0.2168	0.0944
Negative	confused mistakenly	0.5147	0.5234	0.7692	0.5643	0.1923	0.0626
Negative	faulty badly	0.5400	0.5567	0.7713	0.6058	0.2287	0.1019
Negative	flawed sadly	0.6433	0.6235	0.7886	0.6413	0.2655	0.1400

remains relatively more stable overall, even though it served as the original selection dataset. This further shows that sensitivity to harmful lexical framing varies substantially across datasets.

Taken together, the cross-domain experiments support three conclusions. First, harmful adjective–adverb interactions generalize across domains rather than remaining dataset-specific. Second, the magnitude of degradation varies strongly by retrieval setting. Third, only a small subset of lexical combinations transfers as consistently catastrophic across all datasets.

These findings motivate the next stage of analysis: determining whether role identity semantics independently contributes to harmful ranking behavior beyond lexical negativity alone.

5.4 Role Identity and Role–Lexical Interaction

The previous experiments demonstrate that harmful lexical interactions can strongly degrade ranking performance. However, lexical polarity alone may not fully explain harmful prompt behavior. I therefore next investigate whether role semantics independently contributes to degradation when lexical framing is held fixed.

Role-only control on TREC-COVID. Before testing role identity under harmful lexical framing, I first evaluate whether role identity alone has a strong effect on ranking performance. This control experiment removes the negative adjective and adverb components and keeps only the inserted role phrase. The goal is to determine whether semantic role identity by itself is sufficient to produce large ranking changes.

Table 7 summarizes the TREC-COVID role-only results at the group level. The complete role-level scores are reported in Appendix A.1. The no-role baseline obtains an NDCG@10

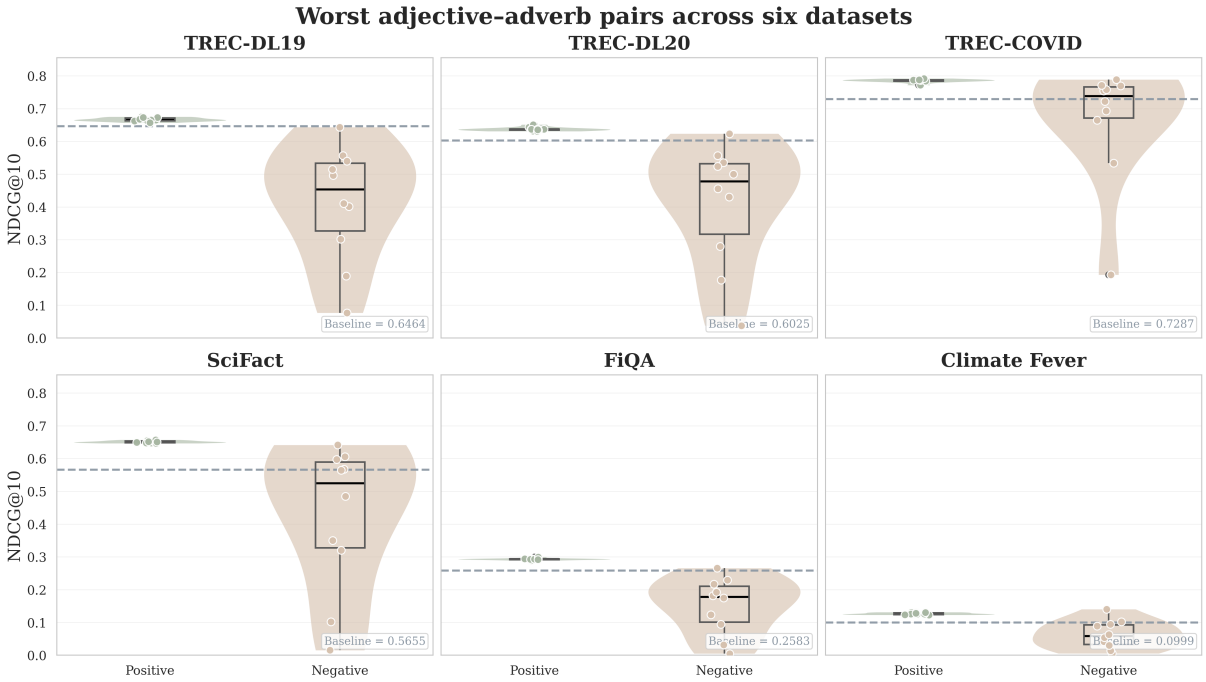


Figure 3: Distribution of the ten selected harmful adjective–adverb pairs across six datasets.

of 0.7278. In contrast, all role-only groups remain above the no-role baseline, and none of them shows the severe degradation observed for negative adjective–adverb prompts. Domain-specific medical roles obtain a mean NDCG@10 of 0.7905, while non-domain semantic occupations obtain a similar mean score of 0.7865. Unrelated semantic roles, object-like unusual roles, and meaningless string roles also remain in a similar performance range.

These results show that role identity alone does not produce the severe degradation observed in the negative role-prompt experiments. Even roles with no clear semantic relation to retrieval, and even meaningless role strings, remain close to the meaningful-role conditions. Thus, semantic role identity by itself cannot explain the harmful ranking behavior observed for negative role prompts.

However, this does not mean that role identity has no effect. The role-only control only shows that role identity is not strongly harmful in isolation.

Since role identity alone does not explain the large drops, the next question is whether it changes the strength of an already harmful lexical frame. To isolate this effect, I fix the adjective *incorrect* and the adverb *falsely*, and vary only the role description itself. Unlike the role-only control, this setting keeps the harmful lexical frame constant and tests whether different role identities amplify or mitigate its effect. Harmfulness is measured relative to a fixed lexical baseline consisting of the adjective–adverb prompt without any role identity:

$$\textit{incorrect} + \textit{falsely}.$$

For a role identity r in dataset d, the performance drop is defined as:

$$\text{Drop}_d(r) = \text{NDCG@10}_{d,\text{adj+adv baseline}} - \text{NDCG@10}_{d,r+\text{adj+adv}}.$$

Here, d denotes the dataset. Under this definition, positive values indicate that adding the

Table 7: Summary of TREC-COVID role-only control results. Values are NDCG@10. The table tests whether role identity alone, without negative adjective or adverb framing, is sufficient to cause large ranking degradation.

Role-only condition	Number of roles	Mean	Min	Max
No-role baseline	1	0.7278	–	–
Domain-specific medical roles	10	0.7905	0.7808	0.7984
Non-domain semantic occupations	10	0.7865	0.7697	0.7976
Unrelated semantic roles	10	0.7815	0.7757	0.7893
Object-like unusual roles	10	0.7806	0.7726	0.7921
Meaningless string roles I	10	0.7993	0.7874	0.8202
Meaningless string roles II	10	0.7895	0.7787	0.7980
All role-only prompts	60	0.7880	0.7697	0.8202

role identity makes performance worse than the fixed *incorrect–falsely* lexical baseline, whereas negative values indicate that the role identity partially mitigates the harmful lexical effect.

I first evaluate 20 candidate role identities on TREC-COVID. Figure 4a shows the resulting performance differences. The results reveal substantial variation even though the lexical framing remains identical across all prompts. Some roles produce additional degradation, while others partially stabilize or even improve the fixed lexical baseline.

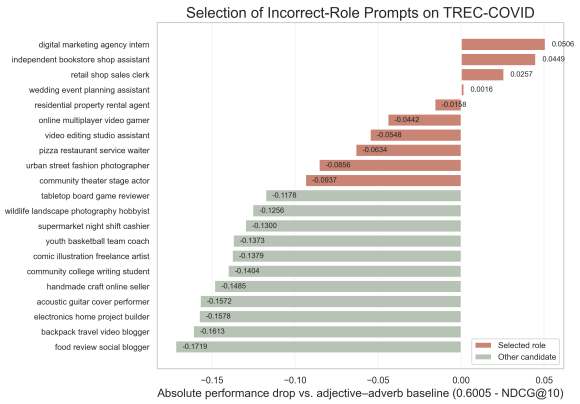
This finding suggests that role identity can affect ranking performance under harmful lexical framing, but its effect is weaker and less systematic than the effect of the harmful adjective–adverb frame itself. In other words, harmfulness cannot be explained purely by role identity. The model reacts strongly to the fixed negative lexical framing, and role identity mainly modulates this effect.

Several role categories appear among the stronger harmful cases on TREC-COVID. These include retail-related occupations, entertainment-related occupations, assistant-level service jobs, and socially informal identities. By contrast, some semantically neutral or less socially grounded roles produce much smaller effects despite sharing the exact same lexical framing.

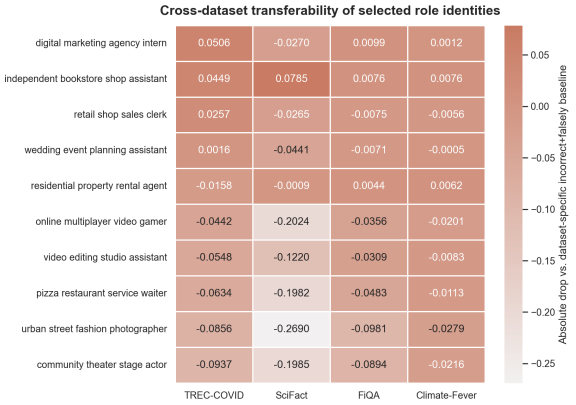
Based on these results, I select the ten role identities with the largest degradation values on TREC-COVID for further transfer analysis.

Cross-domain transfer of selected role identities I next test whether the role identities selected on TREC-COVID produce similar additional effects on other datasets. The selected roles are evaluated on SciFact, FIQA, and Climate-Fever, while the TREC-COVID results are retained as the source-dataset reference. Figure 4b presents the cross-dataset heatmap results. Each cell reports the NDCG@10 drop relative to the dataset-specific fixed *incorrect–falsely* lexical baseline.

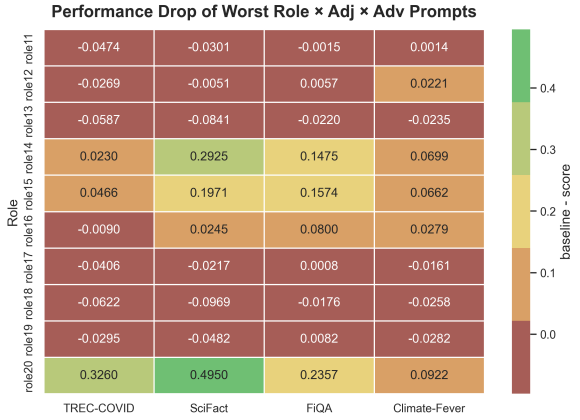
The results show that role identity effects do not transfer as uniformly harmful effects across datasets. On TREC-COVID, several selected roles produce substantial drops, which is expected because the roles were selected from this dataset. However, on SciFact, FIQA, and Climate-Fever, many drop values are close to zero or negative. As defined in Section 4.2, a negative drop indicates that the role-conditioned prompt achieves a higher NDCG@10 score than the dataset-specific *incorrect–falsely* lexical baseline. Thus, after



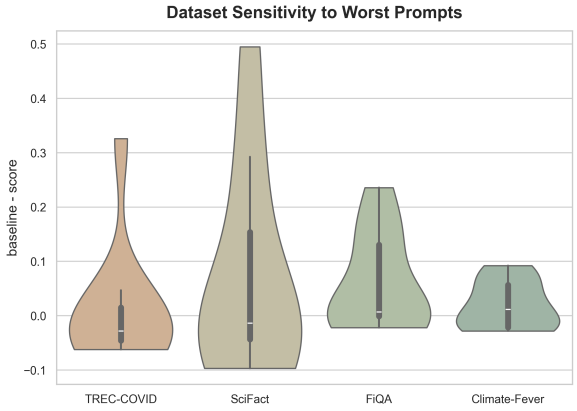
(a) TREC-COVID role-identity selection under fixed *incorrect–falsely* framing.



(b) Cross-dataset transfer of selected role identities.



(c) Performance drop for selected role–adjective–adverb prompts.



(d) Dataset sensitivity to selected harmful role–lexical prompts.

Figure 4: Summary of role-identity and role–lexical interaction analyses. The figure shows the TREC-COVID role-selection step, cross-dataset transfer under fixed lexical framing, performance drops for selected role–adjective–adverb prompts, and dataset-level sensitivity.

controlling for the fixed lexical framing, adding a role identity often does not further harm performance and can sometimes partially mitigate the degradation caused by the lexical prompt.

A small number of selected role identities show more consistent harmful transfer across datasets. The *independent bookstore shop assistant* role produces positive drops on all four datasets, while the *digital marketing agency intern* role produces positive drops on most datasets. Other roles show weaker or inconsistent transfer.

SciFact exhibits the largest variation in role effects. Several roles produce negative drops on SciFact, indicating that adding the role identity mitigates rather than amplifies the fixed harmful lexical framing. FIQA and Climate-Fever show smaller changes overall.

Overall, the results suggest that role identity can influence ranking behavior under fixed lexical framing, but its effect is not consistently harmful across domains. The negative adjective–adverb pair remains the main source of degradation, while role identity acts as a secondary modulating factor.

Table 8: Cross-dataset summary of selected role–adjective–adverb prompts. All values are NDCG@10.

Role ID	Prompt	TREC-COVID	SciFact	FIQA	Climate-Fever
role0	baseline	0.7278	0.5655	0.2583	0.0999
role11	confused retail shop sales clerk × mistakenly	0.7752	0.5992	0.2597	0.0993
role12	messy urban street fashion photographer × incorrectly	0.7547	0.5742	0.2525	0.0786
role13	clumsy online multiplayer video gamer × terribly	0.7865	0.6532	0.2802	0.1242
role14	incorrect independent bookstore shop assistant × improperly	0.7048	0.2766	0.1107	0.0308
role15	hopeless digital marketing agency intern × wrongly	0.6812	0.3720	0.1008	0.0345
role16	unreliable residential property rental agent × horribly	0.7368	0.5446	0.1782	0.0728
role17	flawed pizza restaurant service waiter × sadly	0.7684	0.5908	0.2574	0.1168
role18	faulty video editing studio assistant × badly	0.7900	0.6660	0.2758	0.1265
role19	problematic community theater stage actor × poorly	0.7573	0.6173	0.2500	0.1289
role20	awful wedding event planning assistant × falsely	0.4018	0.0741	0.0225	0.0085

These findings motivate the final behavioral experiment, which combines harmful lexical framing with selected role identities.

Interaction between lexical polarity and role identity The previous sections separately analyzed harmful lexical interactions and role-identity effects under fixed lexical framing. I now investigate how these two components interact when combined into a single prompt.

Specifically, I first take the ten lowest-performing adjective–adverb pairs selected from the TREC-COVID lexical ablation experiment. The complete adjective–adverb selection table is reported in Appendix A, Table 12. I then combine these pairs with ten selected role identities, producing a 10×10 role–lexical grid on TREC-COVID. The complete role–lexical interaction grid is reported in Appendix A, Table 13. From this grid, I select the strongest harmful role–adjective–adverb combinations and evaluate them across TREC-COVID, SciFact, FIQA, and Climate-Fever.

To maintain consistency with earlier sections, harmfulness is measured as:

$$\text{Drop} = \text{NDCG@10}_{\text{baseline}} - \text{NDCG@10}_{\text{prompt}}.$$

Table 8 reports the NDCG@10 scores of the selected role–adjective–adverb prompts, together with the no-role baseline for reference. The negative prompts exhibit substantial variation: some remain relatively close to the baseline, whereas others produce severe degradation across multiple datasets. This indicates that harmfulness cannot be explained by lexical negativity alone or role identity alone. Instead, the most severe failures stem from the interaction between harmful lexical frames and specific role identities. However, this effect is consistently less significant than that caused by adjective-adverb combinations, suggesting that role identity plays a supporting role.

Figure 4c visualizes the same results using performance drops relative to the baseline. The strongest harmful combination is clearly:

awful wedding event planning assistant × *falsely*

(role20). This prompt produces the largest degradation on every dataset simultaneously. The next strongest harmful prompts are:

incorrect independent bookstore shop assistant × *improperly*

(role14), and

hopeless digital marketing agency intern × *wrongly*

(role15). These combinations are especially destructive on SciFact and FIQA. Interestingly, not all role–lexical combinations remain harmful. Several prompts remain relatively close to baseline performance despite containing strongly negative wording. This again demonstrates that harmfulness is highly uneven and interaction-dependent.

Figure 4d summarizes dataset-level sensitivity distributions. SciFact exhibits the widest variance, the largest upper-range degradation, and the strongest instability overall. Among the evaluated datasets, SciFact shows the largest variation and some of the strongest degradation under the selected prompts. By contrast, TREC-COVID remains relatively more stable overall despite still exhibiting strong degradation for several prompt combinations.

Taken together, the interaction experiments support three major conclusions. First, harmful ranking behavior is fundamentally compositional. The degradation emerges when harmful lexical framing is combined with particular role identities.

Second, cross-domain transfer is selective rather than uniform. Only a small subset of role–lexical combinations remains consistently harmful across datasets, while most combinations are more dataset-dependent.

Third, dataset sensitivity varies substantially. SciFact shows particularly large degradation for several selected combinations, indicating that factual verification settings may be especially sensitive to harmful role–lexical prompts.

Overall, the behavioral experiments show that harmful role-playing behavior is driven primarily by negative adjective–adverb framing, while role identity changes the strength of this effect when it appears inside a harmful lexical frame.

However, behavioral ranking metrics do not show how role-conditioned information reaches the final Yes/No decision. The next chapter therefore turns to mechanistic analysis. Importantly, the mechanistic analysis uses the full Random20 prompt set, including both positive and negative prompts, rather than the newly constructed worst adjective–adverb combinations. This separates two goals: the behavioral experiments isolate which prompt components are harmful, while the mechanistic experiments analyze how the original role-playing mechanism is internally routed.

Chapter 6 Mechanistic Analysis: How Role-Prompt Information Is Routed

The behavioral experiments in Section 5 show that harmful role-playing effects are compositional. Lexical polarity is the primary driver of degradation, while role identity modulates the effect of harmful lexical framing. However, those experiments only measure the external ranking behavior. They do not explain where role-conditioned information is represented inside the model or how it reaches the final Yes/No decision.

This chapter studies that internal routing process using activation patching, attention-head patching, helpful-head selection, single-head zero ablation, and top-20 group ablation. The central question is how role-prompt information is encoded, propagated, and integrated into the final ranking decision.

The mechanistic experiments use the full Random20 prompt set, including both positive and negative prompts, rather than the newly constructed worst adjective–adverb combina-

tions from Section 5. This design choice is intentional. The behavioral experiments identify which controlled prompt components are harmful, whereas the mechanistic experiments analyze how the role-prompt effects observed in the initial Random20 evaluation are internally represented and routed. Together, the two parts connect prompt-level harmfulness with internal transformer components.

6.1 Where is role information encoded and routed?

To further investigate how role prompts influence the internal ranking behavior of the model, this section performs activation patching experiments on both attention heads and residual stream components across six datasets: TREC-DL 2019, TREC-DL 2020, TREC-COVID, SciFact, FIQA, and Climate-Fever.
For the mechanistic interpretability experiments, I construct controlled query–document pairs from the qrels. For each query, the document with the highest available relevance label is selected as the relevant document. An irrelevant document is selected from the label-0 documents using a fixed per-query random seed. This produces one relevant and one irrelevant document for each query whenever both are available.
This construction is used to separate relevance and irrelevance mechanisms. In the relevance condition, the selected relevant query–document pairs are used and the target direction is

$$\text{logit(Yes)} - \text{logit(No)}.$$

In the irrelevance condition, the selected label-0 or irrelevant query–document pairs are used and the target direction is

$$\text{logit(No)} - \text{logit(Yes)}.$$

This label-aware construction avoids mixing positive and negative evidence within the same mechanistic condition.
For activation patching, the sampling strategy differs across datasets because of computational cost. For TREC-DL 2019, TREC-DL 2020, and TREC-COVID, I use all available evaluation query–document pairs in the selected patching setting. For SciFact, FIQA, and Climate-Fever, I use a sampled subset of 50 query–document pairs per dataset due to computational constraints.
The activation patching analysis focuses on two levels of representation:

1. Attention head outputs, which identify localized information-processing circuits;
2. Residual stream components, which reveal how different prompt segments contribute across model layers.

The primary metric used in this section is `ld_recovery`, defined in Section 4.3. It measures how much a patched activation restores the label-aware logit difference between the clean and corrupted prompts. Higher positive values indicate stronger causal contribution of the patched component.

6.1.1 Attention Head Patching Across Datasets and Prompt Positions

To examine whether role-prompt effects rely on dataset-specific mechanisms or shared internal circuits, I patch attention-head outputs across the three full-query datasets: TREC-

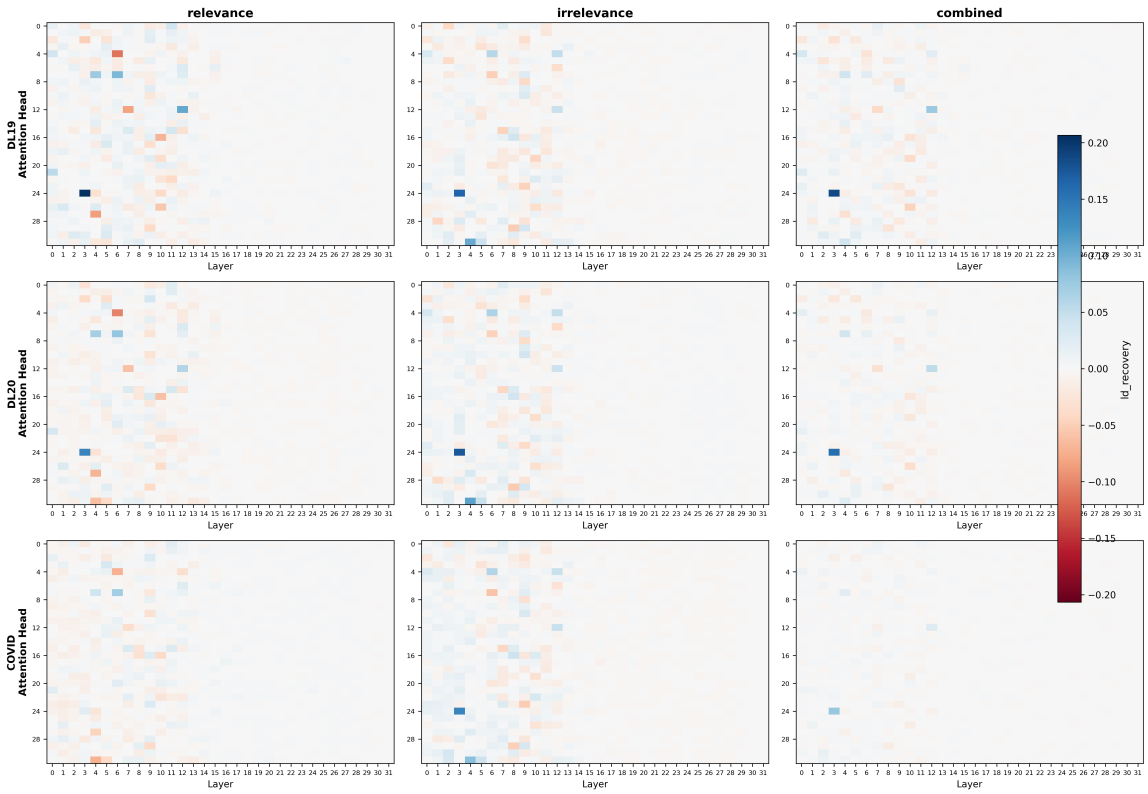


Figure 5: Cross-dataset attention head patching results for the `Role-Adj-Adv` prompt component using the `ld_recovery` metric. Columns correspond to relevance patching, irrelevance patching, and the averaged combined results. Rows correspond to TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

DL 2019, TREC-DL 2020, and TREC-COVID. The analysis reports two label conditions, relevance and irrelevance, together with their averaged combined result.

I focus on three prompt positions in the main text. First, `Role-Adj-Adv` patches the complete lexical role phrase, including both adjective and adverb information. Second, `Inst` patches the instruction segment. Third, `Last` patches the final token representation before the model produces the Yes/No answer. Together, these positions allow us to trace a possible routing path from role-phrase encoding to instruction-level steering and finally to the decision position.

Because each heatmap contains all layers and attention heads across three datasets and three target directions, reducing the figures further would make individual head patterns difficult to inspect. Therefore, the main text reports the three most informative positions, `Role-Adj-Adv`, `Inst`, and `Last`, at full width. The corresponding results for SciFact, FIQA, and Climate-Fever are reported in Appendix D, and the separate `Role-Adj` and `Role-Adv` results are reported in Appendix D.1.

Figures 5, 6, and 7 show a consistent routing pattern across TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The three datasets exhibit very similar layer-head structures, suggesting that these datasets share a broadly similar layer–head pattern.

The `Role-Adj-Adv` position highlights an early encoding stage. The most visible recurring head is `L3H24`, especially in the irrelevance and combined conditions. This suggests that

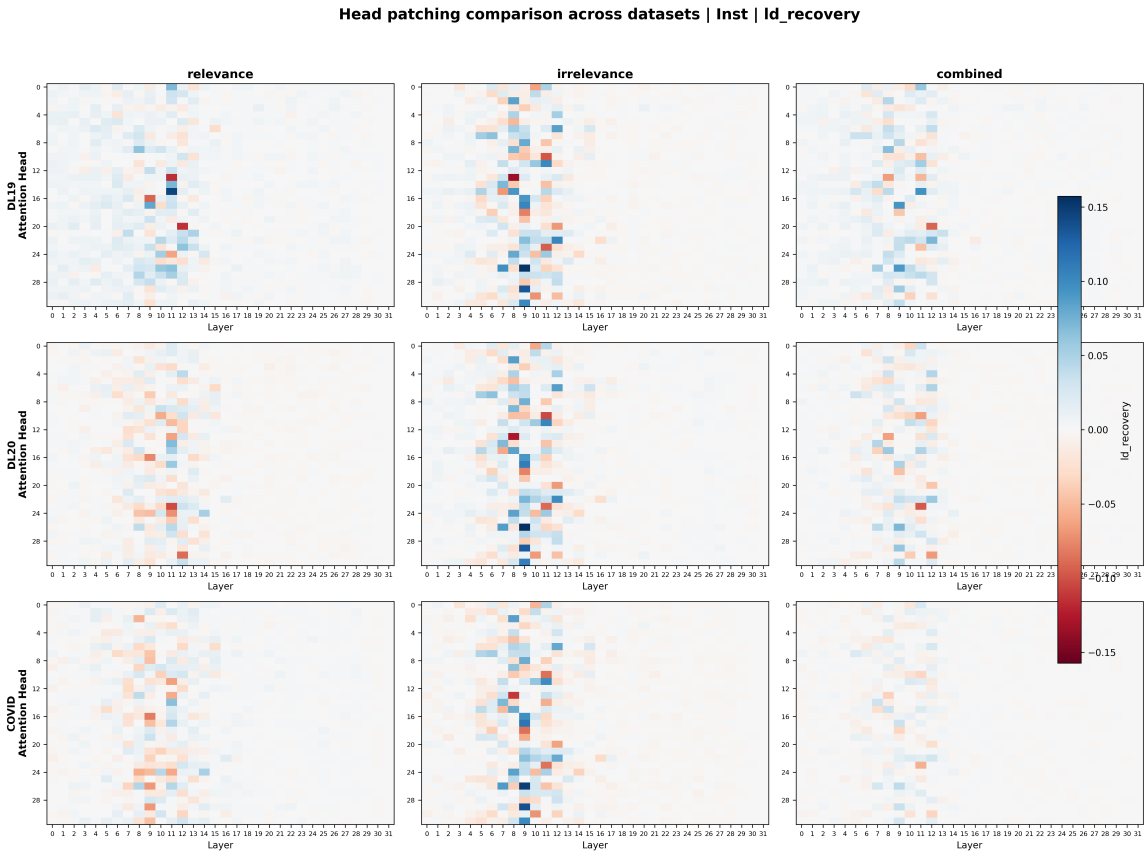


Figure 6: Cross-dataset attention head patching results for the `Inst` token position using the `ld_recovery` metric. Columns correspond to relevance patching, irrelevance patching, and the averaged combined results. Rows correspond to TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

the complete lexical role phrase is encoded in shallow layers before it affects later decision-related components.

The `Inst` position shows a different pattern. The strongest instruction-level effects are concentrated mainly around layers 9 and 11. In particular, heads such as `L9H26`, `L9H29`, `L9H17`, and `L11H11` are more visible in the irrelevance condition than in the relevance condition. This indicates that instruction-level role information is present in middle layers, but it is not the primary site where the final ranking decision is formed.

The clearest and strongest recovery appears at the `Last` token. Here, the irrelevance condition produces a much stronger and more concentrated signal than the relevance condition. The most visible heads are located around layers 14–16, including `L14H24`, `L15H4`, `L13H23`, `L15H8`, and `L16H29`. This suggests that role-conditioned ranking information is consolidated near the final prediction position before the model outputs the Yes/No decision.

The same qualitative relevance–irrelevance asymmetry also appears in the additional datasets reported in Appendix D. Although SciFact, FIQA, and Climate-Fever use sampled query–document pairs and differ in task domain, their appendix heatmaps show the same broad pattern: irrelevance patching produces stronger and more structured recovery than relevance patching, especially at the `Last` token. This indicates that the asymmetry between accepting relevant documents and rejecting irrelevant documents is not limited to the three full-query datasets.

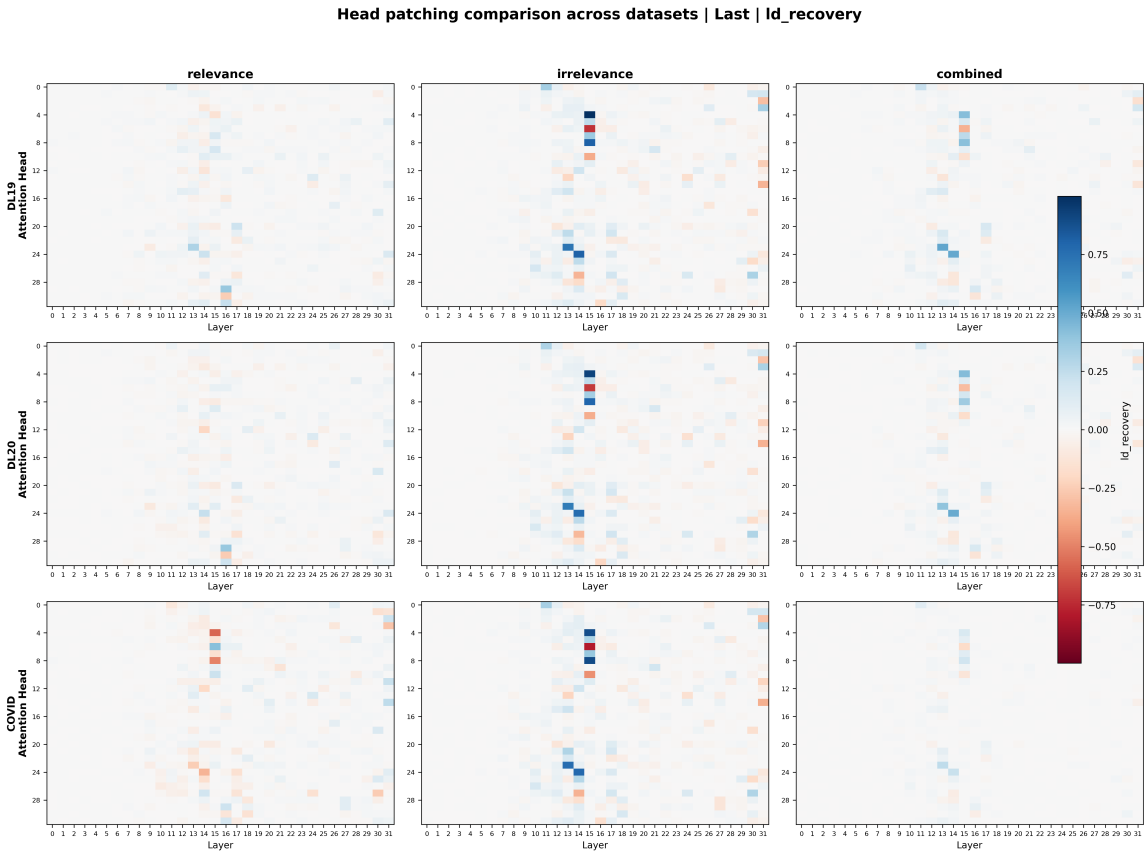


Figure 7: Cross-dataset attention head patching results for the `Last` token position using the `ld_recovery` metric. Columns correspond to relevance patching, irrelevance patching, and the averaged combined results. Rows correspond to TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

Taken together, the head-patching results suggest a staged routing pattern under the front-position role-prompt format used in this thesis. Because the role prompt appears before the instruction, query, and document, the early recovery around the role phrase may partly reflect this input ordering. Within this setting, the complete role phrase is recoverable in early heads such as `L3H24`, while instruction-level role information appears in middle-layer heads around layers 9–11. The strongest causal recovery is later concentrated at the `Last` token, where the model prepares the final Yes/No relevance decision. Thus, in the prompt ordering studied here, role-prompt information appears locally around the role phrase and later becomes available at decision-level attention heads.

6.1.2 Residual Stream Component Analysis

Beyond attention heads, residual stream patching provides a higher-level view of how different prompt segments contribute throughout the transformer stack.
Figures 8, 9, 10, and 11 show the residual stream patching results for TREC-DL 2019 and TREC-DL 2020 under both relevance and irrelevance settings. Due to space limitations, only DL19 and DL20 are shown in the main text, while the remaining datasets are included in Appendix C.
The residual stream results reveal several notable patterns.

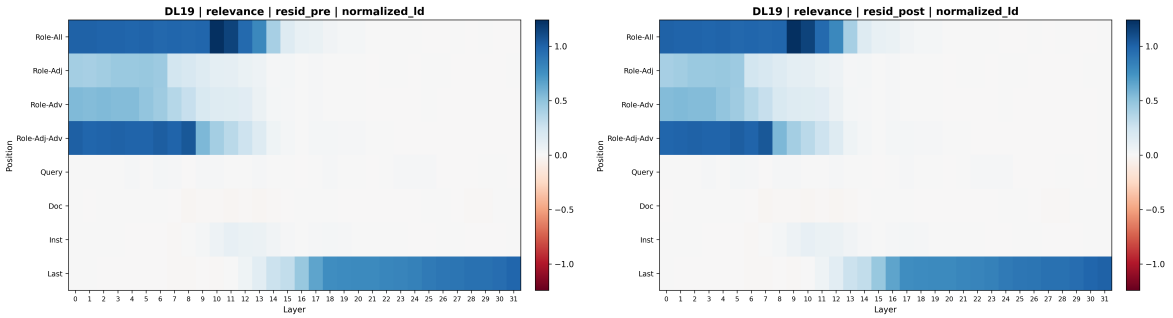


Figure 8: Residual stream patching results for TREC-DL 2019 under the relevance setting. Left: `resid_pre`. Right: `resid_post`.

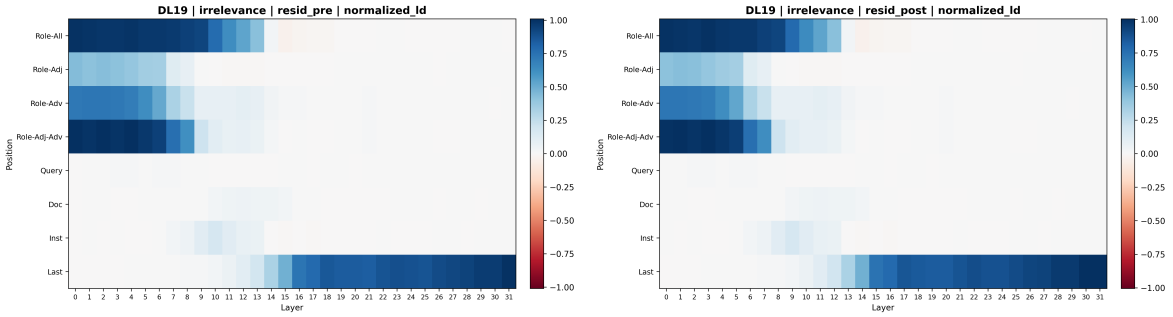


Figure 9: Residual stream patching results for TREC-DL 2019 under the irrelevance setting.

First, the strongest recovery consistently appears in the `Role-All` and `Role-Adj-Adv` conditions during early transformer layers. This indicates that the model encodes role-conditioning information very early in the forward pass.
Second, the `Last` token position shows extremely strong recovery in late layers across both relevance and irrelevance settings. This suggests that the final token acts as an aggregation point where ranking-relevant information is consolidated before output generation.
Third, the `Query` and `Doc` components exhibit comparatively weak recovery throughout most layers. This suggests that, under this patching metric, the largest recoverable differences are associated with prompt-conditioning signals rather than with the query or document positions themselves.
Finally, the irrelevance setting generally produces stronger and more persistent recovery than the relevance setting. In DL19 and DL20, the label-aware recovery for irrelevant documents extends across a broader range of layers, especially from approximately layers 14 to 31. This suggests that the model's rejection of irrelevant documents is more strongly affected by the patched positive–negative role differences than its acceptance of relevant documents.
Overall, the residual-stream and attention-head patching results reveal a broadly consistent routing pattern. Differences between positive and negative role conditions are recoverable around role-related positions in earlier layers and become strongly concentrated at the final decision token in later layers. The stronger recovery observed for irrelevant documents further suggests that relevance acceptance and irrelevance rejection rely on partially asymmetric internal computations.

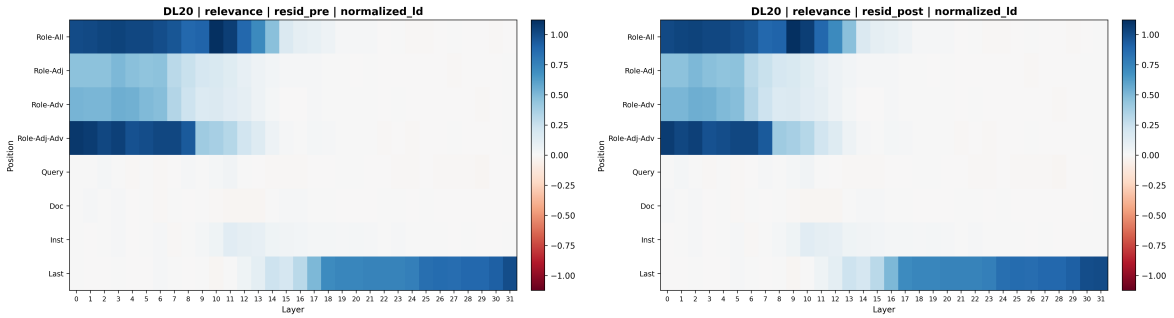


Figure 10: Residual stream patching results for TREC-DL 2020 under the relevance setting.

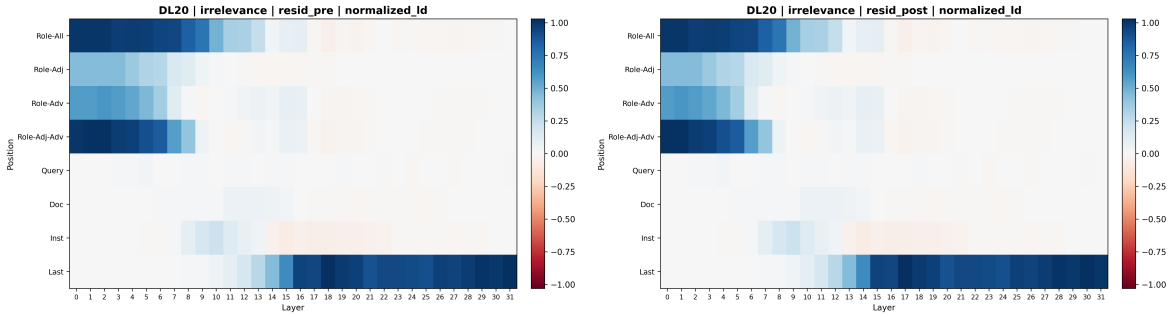


Figure 11: Residual stream patching results for TREC-DL 2020 under the irrelevance setting.

6.2 Helpful Head Selection for Causal Intervention

The activation patching results in the previous section show that role-prompt effects are not uniformly distributed across the model. Instead, the strongest recovery is concentrated in specific token positions and attention heads. To connect the patching analysis with causal intervention, I next identify a smaller set of consistently helpful heads and use them as candidates for the zero-ablation and group-ablation experiments in Section 6.3.

For each dataset and each target position, I rank attention heads by their label-aware logit-difference recovery score. The ranking is computed separately for relevance and irrelevance examples. To obtain the final top-10 head sets for single-head zero ablation, I aggregate the patching results across datasets and target positions. Specifically, for each head, I compute the mean `ld_recovery` score over all considered datasets and target positions, separately for relevance and irrelevance examples. The ten heads with the highest mean recovery in the relevance setting form the relevance top-10 set, and the ten heads with the highest mean recovery in the irrelevance setting form the irrelevance top-10 set. This selection criterion is therefore based on average recovery strength rather than on a single dataset or a single target position. For relevant documents, the logit-difference direction is defined as Yes–No, while for irrelevant documents it is defined as No–Yes. This distinction is important because the relevance and irrelevance settings correspond to different decision directions. Therefore, the head selection is also performed separately for relevance and irrelevance.

I use the patching rankings in two stages. First, I record the top-20 heads for each dataset, label condition, and target position. These top-20 lists provide the candidate pool for later

group ablation. Second, I select a smaller top-10 relevance-head set and top-10 irrelevance-head set for single-head zero ablation. This two-step design separates broad candidate discovery from focused causal testing.

This section focuses on four target positions used for head selection: `Role-All`, `Role-Adj-Adv`, `Inst`, and `Last`. `Role-All` measures the full role span. `Role-Adj-Adv` measures the complete lexical role phrase, including both adjective and adverb information. `Inst` measures whether role information is recoverable from the instruction region. `Last` measures whether role information directly affects the final decision position. The remaining target-position results, including `Role-Adj` and `Role-Adv`, are reported in Appendix D. The complete top-20 candidate-head rankings for all six datasets are reported in Appendix E.3.

Rather than listing the full candidate-head tables in the main text, I summarize the main patterns that motivate the subsequent causal interventions. Across datasets, role-related positions tend to emphasize earlier heads, such as `L3H24`, whereas the `Last` position more often selects middle-to-late heads associated with the final Yes/No decision, such as `L15H4`, `L15H8`, `L14H24`, and `L16H29`. The relevance and irrelevance rankings are also not identical. This supports the decision to select and ablate relevance-side and irrelevance-side head groups separately.

6.2.1 Instruction-Level Helpful Heads

Figure 12 shows the top-10 helpful heads for instruction-level patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The corresponding Climate-FEVER, FIQA, and SciFact results are reported in Appendix E.1. Here, "helpful" refers only to positive activation-patching recovery and does not imply that zero ablation of the head must reduce the ground-truth-aligned decision direction.

The instruction-level results show moderate recovery values compared with the later decision position. This suggests that the instruction region contains role-conditioned information, but it is not the primary location where the final Yes/No ranking decision is formed. Across datasets, irrelevance examples tend to produce stronger recovery than relevance examples.

Several heads recur across datasets in the instruction setting, including `L9H17`, `L11H15`, `L9H21`, `L11H0`, `L10H22`, and `L8H9`. Most of these heads are located in shallow or middle layers. This indicates that instruction-level role information is likely encoded early and then passed forward to later decision-related components.

6.2.2 Last-Token Helpful Heads

Figure 13 shows the top-10 helpful heads for last-token patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The corresponding Climate-FEVER, FIQA, and SciFact results are reported in Appendix E.1. This position is especially important because the next-token Yes/No decision is computed from the representation at the final prompt position [8, 21].

The last-token results produce the strongest recovery scores among the main target positions considered here. This concentration is consistent with earlier mechanistic analyses of LLM relevance judgment, which also identify late-layer attention heads at the final token as important for controlling the generated relevance label [8]. This indicates that role-prompt information is most causally concentrated at the final decision position. The

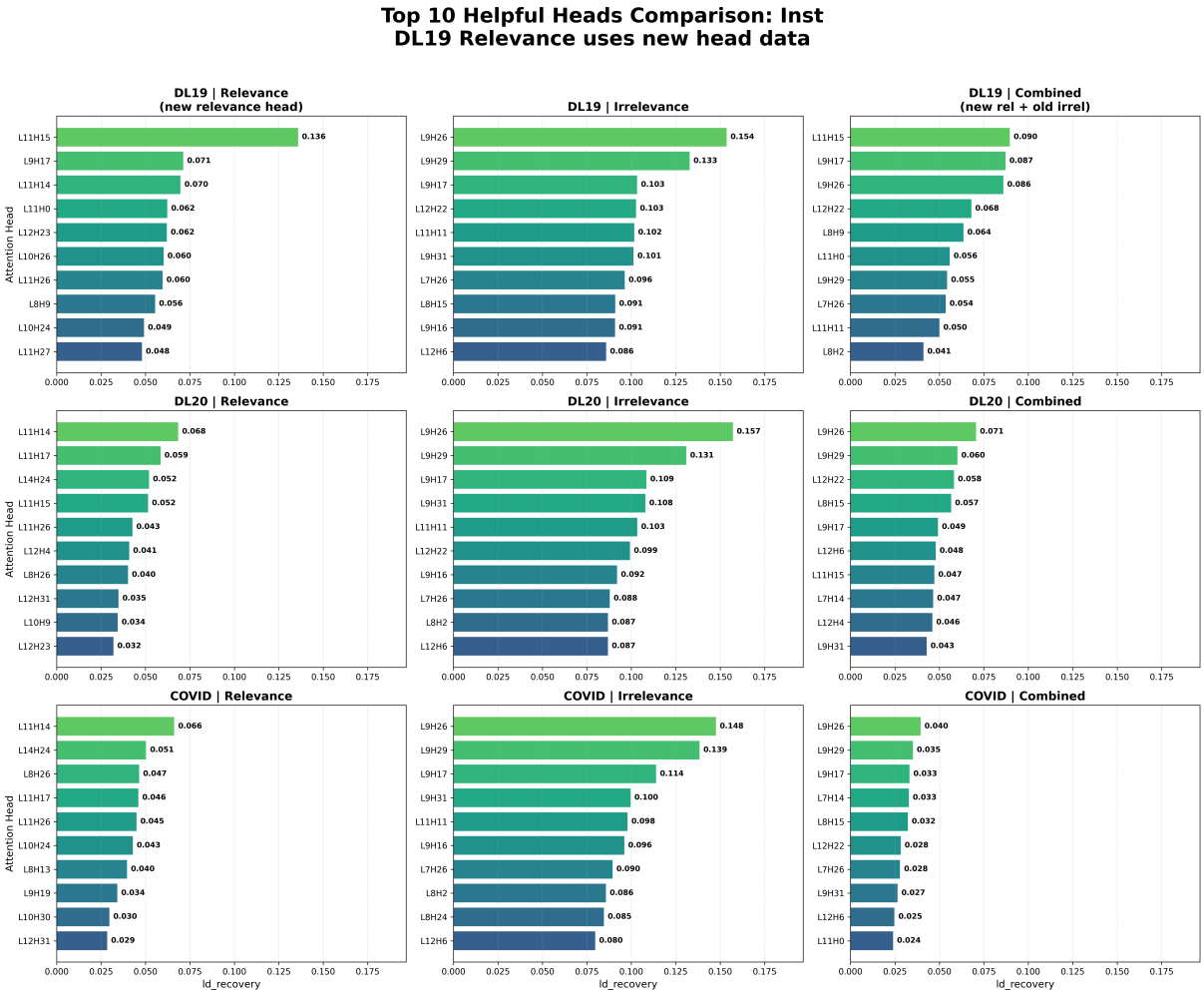


Figure 12: Top-10 helpful heads for instruction-level patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

effect is especially strong in the irrelevance setting, where several heads achieve much larger recovery scores than in the relevance setting.

The strongest relevance heads include `L16H29`, `L15H6`, `L15H10`, `L16H31`, and `L14H27`. These heads are mainly located in middle-to-late layers. In contrast, the strongest irrelevance heads include `L15H4`, `L14H24`, `L15H8`, `L13H23`, and `L15H7`. The difference between the relevance and irrelevance head groups suggests that accepting relevant documents and rejecting irrelevant documents rely on partially different internal circuits.

6.2.3 Lexical Role-Phrase Helpful Heads

Figure 14 shows the top-10 helpful heads for `Role-Adj-Adv` patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The corresponding Climate-FEVER, FIQA, and SciFact results are reported in Appendix E.1. This setting patches the full lexical role phrase, including adjective and adverb. It is therefore the most directly connected to the lexical role-prompt manipulations discussed earlier in the thesis.

The `Role-Adj-Adv` results differ from the last-token results. Instead of being dominated by late decision heads, this setting highlights shallow and middle-layer heads. The most

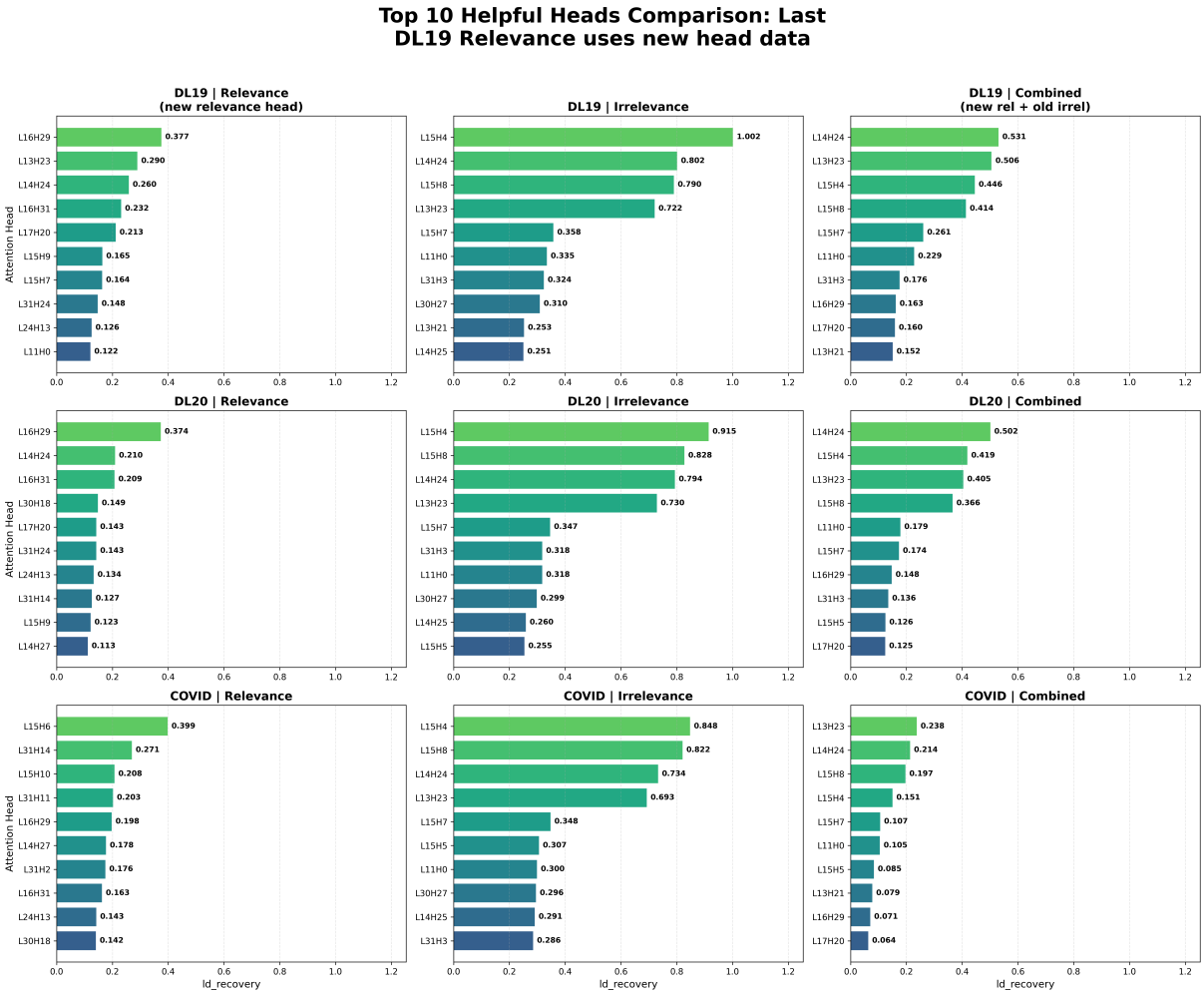


Figure 13: Top-10 helpful heads for last-token patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

stable head is `L3H24`, which appears as the strongest or one of the strongest heads across multiple datasets. Other recurring heads include `L12H12`, `L4H31`, `L4H7`, `L0H21`, and `L6H7`. This pattern suggests a two-stage mechanism. Lexical role information is first encoded in shallow or middle-layer heads, especially around the role phrase itself. It is then routed forward and becomes integrated into later decision-level heads at the final token. Thus, the role-prompt effect is not localized to a single token or layer. Instead, it appears to involve early lexical encoding followed by late decision aggregation.

6.2.4 Summary of Recurring Helpful Heads

Table 9 provides a compact overview of heads that appear prominently across the three main patching settings. Rather than serving as an additional selection criterion, the table summarizes the head-level patterns observed in the preceding patching results.

The table shows that the three patching settings are related but not identical. Some heads appear across multiple regions, suggesting partial overlap between role-related, instruction-level, and final-token routing. At the same time, the settings emphasize different parts of the model. `Role-Adj-Adv` patching highlights early heads associated with the

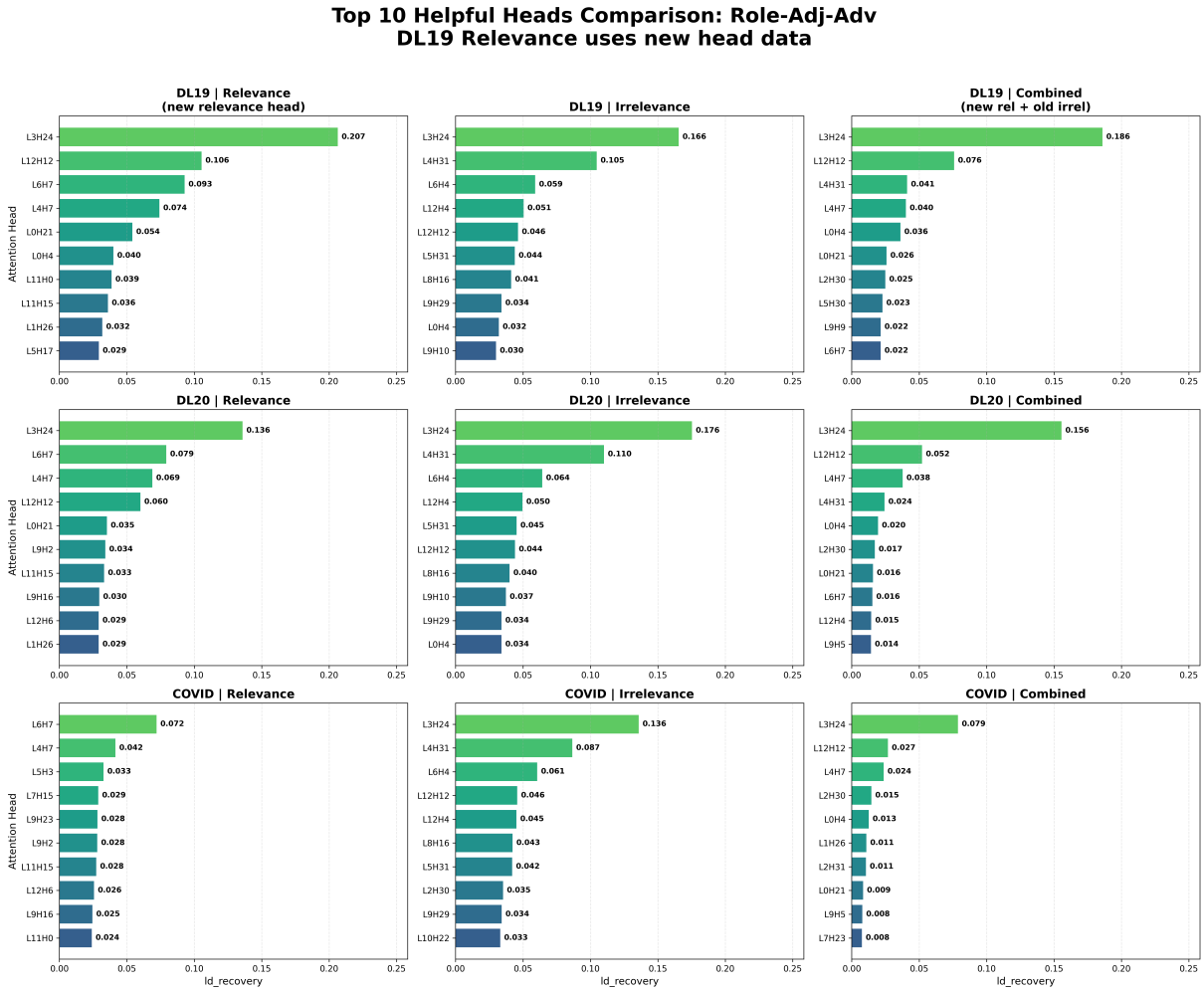


Figure 14: Top-10 helpful heads for joint adjective-adverb role-phrase patching on TREC-DL 2019, TREC-DL 2020, and TREC-COVID.

lexical role phrase, such as `L3H24`, whereas last-token patching highlights more middle-to-late heads that appear in the final-token recovery rankings. Instruction patching falls between these two patterns and mainly highlights heads associated with instruction-level role information. This overview motivates the following ablation experiments, where the selected heads are tested causally rather than only identified through patching recovery.

6.2.5 Final Top-10 Heads for Single-Head Zero Ablation

Using the mean `ld_recovery` aggregation described above, I select two separate groups of heads for the single-head zero-ablation experiments. The first group contains the top-10 heads for relevance-side effects. The second group contains the top-10 heads for irrelevance-side effects. These two groups are selected separately because relevance and irrelevance use different logit-difference directions and show different dominant heads. Table 10 lists the final 20 heads used in the following single-head ablation experiments. These two sets are used in Section 6.3. The purpose of the next section is to test whether the heads identified by activation patching are only recoverable indicators of role-prompt information, or whether they causally affect the final ranking decision when ablated.

Table 9: Overview of prominent heads across the three main patching settings.

Instruction patching	**Last-token patching**	**Role-Adj-Adv patching**
L9H17	L15H4	L3H24
L11H15	L14H24	L12H12
L9H21	L15H8	L4H31
L11H0	L13H23	L4H7
L10H22	L15H7	L0H21
L8H9	L11H0	L6H7
L11H11	L13H21	L4H4
L10H26	L14H25	L5H31
L11H24	L16H29	L2H30
L5H7	L13H13	L8H16

Table 10: Final selected heads for single-head zero-ablation experiments.

Top-10 relevance heads	**Top-10 irrelevance heads**
L16H29	L3H24
L15H6	L14H24
L15H10	L15H4
L16H31	L15H8
L14H27	L13H23
L31H14	L15H7
L31H11	L11H0
L24H13	L13H21
L30H18	L14H25
L13H23	L30H27

6.3 Do the selected heads causally affect the ranking decision?

After selecting the top-10 relevance and irrelevance heads from the helpful-head analysis, I next test whether these heads are causally important for the model's ranking decision. For this experiment, I perform zero ablation on each selected head. Specifically, the output of one attention head is replaced with zero during inference, and I measure how this changes the label-aware logit difference.

Unlike the preliminary intervention based on a single positive–negative role pair, the results in this section are computed over all 20 Random20 role prompts. This makes the intervention results less dependent on one specific role choice and provides a more robust test of whether the selected heads have stable causal effects.

For relevance examples, the target direction is:

$$\text{logit(Yes)} - \text{logit(No)}.$$

For irrelevance examples, the target direction is:

$$\text{logit(No)} - \text{logit(Yes)}.$$

The ablation score is defined as:

$$\Delta = \text{LD}_{\text{ablated}} - \text{LD}_{\text{clean}}.$$

Under this definition, a negative value means that ablating the head decreases the correct label-aware logit direction. In other words, the model becomes worse after the head is removed, so the head can be interpreted as supporting the target decision. A positive value means that ablating the head increases the correct direction, suggesting that the head may suppress or regulate the target decision under normal inference.

6.3.1 Single-head zero ablation of the selected top-10 heads

Figure 15 shows the zero-ablation results for the selected relevance heads on relevance examples, averaged over all 20 role prompts.

The relevance-side A figure shows that the selected heads are not functionally uniform. The most direct evidence for relevance-supporting heads comes from negative deltas, because a negative value means that ablating the head weakens the correct Yes-over-No direction.

Across the six datasets, `L16H31` and `L15H10` are the most stable relevance-supporting heads. Both heads produce negative deltas on all or nearly all datasets. This means that removing these heads consistently weakens the model's ability to support the correct relevance decision. Several other heads, such as `L24H13` and `L13H23`, also often show negative deltas, but their effects are smaller.

At the same time, some relevance-selected heads show positive deltas. For example, `L16H29` and `L15H6` are positive on most datasets. This suggests that these heads are not directly supporting the Yes-over-No direction under zero ablation. Instead, they may play a suppressive or regulatory role in the original computation.

The most dataset-dependent head is `L31H14`. On DL19 and DL20, ablating `L31H14` produces a negative delta, so it supports the relevance direction in the two MS MARCO-based datasets. However, on TREC-COVID, SciFact, FIQA, and Climate-FEVER, the same head produces a positive delta. This shows that the causal role of a selected head can depend strongly on the dataset and retrieval domain.

Figure 16 further decomposes the same relevance-head ablation effects by target position. The relevance-side B figure shows that the largest causal effects are concentrated at the *Last* token. Most other positions, including `Instruction`, `Query`, `Role-Adj`, `Role-Adv`, and `Role-Adj-Adv`, remain close to zero for most heads. This confirms the earlier activation-patching result that relevance-related information becomes most causally important at the final decision position.

The position-wise results also show that DL19 and DL20 have very similar trends. This is expected because both datasets are based on the same MS MARCO passage corpus. In both datasets, `L31H14`, `L16H31`, and `L15H10` show clear negative effects at the *Last* token, while `L16H29` and `L15H6` show positive or weak effects.

For TREC-COVID, SciFact, FIQA, and Climate-FEVER, the same last-token concentration remains visible, but the sign of some heads changes. In particular, `L31H14` becomes strongly positive at the *Last* token in these datasets. This means that removing `L31H14` strengthens the relevance direction outside DL19 and DL20, suggesting a dataset-dependent suppressive effect.

Figure 17 shows the zero-ablation results for the selected irrelevance heads on irrelevance examples, averaged over all 20 role prompts.

The irrelevance-side A figure shows a clearer and more stable split than the relevance-side figure. Again, negative deltas identify heads that support the correct No-over-Yes

direction, because removing these heads weakens the model's ability to reject irrelevant documents.
The strongest irrelevance-supporting heads are L15H4 and L15H8. Both heads produce negative deltas across all six datasets. The magnitude is especially large on SciFact, FIQA, and Climate-FEVER, but the same negative direction also appears on DL19, DL20, and TREC-COVID. L14H24 also produces negative deltas on all datasets, although its magnitude is smaller than that of L15H4 and L15H8. Together, these heads form the clearest group of No-over-Yes supporting heads.
In contrast, several irrelevance-selected heads produce positive deltas across all or most datasets. The most stable positive head is L11H0, which has the largest positive delta in every dataset. Other positive heads include L13H23, L15H7, L3H24, L14H25, and L13H21. Because a positive delta means that removing the head strengthens the No-over-Yes direction, these heads should not be interpreted as directly supporting irrelevance. Instead, they appear to suppress or regulate the irrelevance decision under normal inference.
Figure 18 shows the position-wise decomposition for the selected irrelevance heads.
The irrelevance-side B figure again shows that the strongest effects are concentrated at the *Last* token. This confirms that rejecting irrelevant documents is also finalized mainly at the final decision position.
The strongest last-token effect comes from L15H4. Ablating this head produces a large negative delta at the *Last* token across all datasets, especially on SciFact, FIQA, and Climate-FEVER. This means that L15H4 is one of the most important heads for supporting the No-over-Yes direction. L15H8 and L14H24 show a similar negative pattern, although their magnitudes are generally smaller.
The irrelevance-side position results also show clear dataset structure. DL19 and DL20 are highly similar, which is consistent with their shared MS MARCO corpus. TREC-COVID follows the same general direction but with intermediate magnitude. SciFact, FIQA, and Climate-FEVER show much larger negative last-token effects, especially for L15H4. This suggests that irrelevance judgments in the BEIR-style factual or domain-specific datasets are more sensitive to the selected irrelevance heads.
In addition to the last-token effect, L3H24 shows a visible positive response at the Role-All position, especially in SciFact, FIQA, and Climate-FEVER. This is consistent with the earlier activation-patching result that L3H24 is associated with role-phrase information. However, in the zero-ablation setting, this effect appears more clearly at the full role span than at the isolated adjective or adverb positions. This suggests that L3H24 may encode broader role-span information rather than functioning as a purely adjective-only or adverb-only component.
Overall, the top-10 single-head ablation results support the usefulness of the patching-based head-selection procedure. The heads selected from activation patching are not only correlated with recovery; they also causally affect the final ranking decision. At the same time, zero ablation shows that selected heads are not all helpful in the same way. Some heads directly support the correct decision direction, while others suppress or regulate it depending on the dataset and relevance condition.

6.3.2 Group ablation of top-20 candidate heads

The single-head zero-ablation experiments test the causal influence of individual selected heads. I next ablate the broader top-20 candidate groups selected from the activation-

patching rankings. For each dataset, relevance direction, and target position, the corresponding twenty head outputs are set to zero simultaneously. The results are then evaluated separately for positive and negative role prompts.
This analysis uses the role-conditioned correct score defined in Section 4.3. For positive roles, a high score indicates support for the standard relevance judgment: `Yes` for relevant documents and `No` for irrelevant documents. For negative roles, the expected direction is reversed: `No` for relevant documents and `Yes` for irrelevant documents. The score therefore measures agreement with the answer direction induced by the current role condition rather than ordinary classification accuracy.
Figure 19 reports the zero-ablation results for the four target positions: `Role-All`, `Role-AdjAdv`, `Inst`, and `Last`. The dashed lines show the corresponding clean-model scores. Mean-ablation results are reported separately in Appendix E.2.
The results should not be interpreted as assigning one fixed semantic function to each target-specific head group. The relevance-selected heads used for positive and negative relevant examples have the same indices, as do the corresponding irrelevance-selected heads. However, their activations depend on the prompt context. In addition, each top-20 set may contain both supporting and suppressive heads. The bars therefore represent the net effect of removing a functionally mixed group under a particular combination of role polarity and document relevance.

Positive-role results. For positive roles, the clearest and most consistent effect occurs at the `Last` token. Zeroing the last-token group substantially lowers the role-conditioned correct score on irrelevant documents across all six datasets. In this condition, a high score means support for the expected `No` answer. The large decrease therefore shows that the last-token group is causally important for rejecting irrelevant documents under positive role prompts.
The same last-token intervention also affects relevant documents, although the magnitude is generally less uniform than on irrelevant documents. Taken together, these results indicate that positive-role behavior is mainly finalized at the output decision position, where the model combines the preceding prompt and query–document information before producing `Yes` or `No`.
The effects of `Role-All` and `Role-AdjAdv` are much smaller for positive roles. On relevant documents, `Role-All` ablation often leaves the score close to the clean baseline and sometimes produces a small increase. Because the expected answer in this condition is `Yes`, the increase means that the ablated model becomes slightly more confident in the expected direction. However, this should not be interpreted as evidence that all heads selected from the `Role-All` ranking actively suppress `Yes`. The intervention removes twenty functionally mixed heads at once, and most of the observed changes are small. The more defensible conclusion is that this group is not necessary for the normal relevant-to-`Yes` judgment under positive roles.

Negative-role results. Negative roles show a different clean-score pattern. For relevant documents, the role-conditioned correct score is generally low, meaning that the model does not strongly follow the role-induced `No` direction. For irrelevant documents, the clean score is much higher, meaning that negative role prompts more strongly induce the model to support `Yes` for passages that should normally be rejected. In ranking terms, this is

a harmful false-positive tendency because irrelevant documents can receive elevated Yes probability and move upward in the ranked list.
The clearest role-span effect appears for negative roles on irrelevant documents. Zeroing the heads selected from the `Role-All` ranking produces a substantial decrease in the role-conditioned correct score across datasets. Because the negative-role expected answer for an irrelevant document is `Yes`, this decrease shows that the intact `Role-All`-selected group collectively contributes to the harmful irrelevant-to-`Yes` tendency.
This result does not imply that the `Role-All` group has a universal Yes-promoting function. On negative-role relevant examples, the clean score is already low, and the changes after ablation are smaller and less directly interpretable. Similarly, under positive roles the same selected head indices do not show the same strong pattern. The reliable conclusion is therefore condition-specific: the full-role-span group makes a clear causal contribution to the negative-role false-positive behavior on irrelevant documents.
The `Role-AdjAdv` intervention is weaker and less consistent than `Role-All`. This does not contradict the behavioral finding that adjective–adverb combinations are the primary trigger of ranking degradation. The behavioral experiment identifies which lexical construction initiates harmful behavior, whereas group ablation tests where the resulting role-conditioned signal is causally represented. The present results suggest that, once induced, the negative-role signal is represented across the broader role phrase rather than being confined to the adjective and adverb tokens alone.
The `Inst` group produces intermediate effects. It changes the role-conditioned score in several conditions, but its pattern is less consistent than the full-role-span effect for negative-role irrelevant documents and the last-token effect for positive-role irrelevant documents. This is consistent with the instruction region participating in the routing of role-conditioned information without serving as its primary encoding or final decision site.

The role of the final token. The `Last` intervention has broader effects across positive and negative roles and across relevant and irrelevant documents. Its effect is not confined to one role polarity or one document label. This wider sensitivity is consistent with the final token acting as a decision-integration site rather than as a fixed role-specific or label-specific circuit.
At the final token, the model has access to the accumulated role prompt, instruction, query, and document representations. The selected last-token heads can therefore combine role-conditioned information with query–document evidence before the model produces the final `Yes`/`No` answer. The direction of the ablation effect can vary across conditions because the answer expected under positive and negative roles is different and because the selected groups contain functionally mixed components.
Overall, the zero-ablation results suggest a distinction between role-conditioned signal representation and final decision integration. The clearest role-span contribution appears in the harmful negative-role response to irrelevant documents: removing the `Role-All`-selected group weakens the role-induced `Yes` tendency. By contrast, the `Last` group affects a broader range of role–label conditions and is especially important for the normal rejection of irrelevant documents under positive roles. Thus, `Role-All` identifies a condition-specific source of harmful role-conditioned behavior, whereas `Last` acts as the more general site at which role and relevance information are integrated into the final decision.

6.3.3 Summary of causal intervention findings

The causal intervention experiments support five conclusions.
First, the selected top-10 heads causally affect the label-aware logit difference across all six datasets and across the full Random20 prompt set. This confirms that heads selected by activation patching are not only correlated with recovery, but also causally influence the model's Yes/No decision direction.
Second, the single-head zero-ablation results show that selected heads are functionally mixed. Some heads directly support the target label-aware direction, while others suppress or regulate it. For relevance judgments, `L16H31` and `L15H10` are the most stable relevance-supporting heads. For irrelevance judgments, `L15H4`, `L15H8`, and `L14H24` are the clearest heads supporting the No-over-Yes direction. Other selected heads often produce positive ablation deltas, indicating suppressive or regulatory roles.
Third, the top-20 zero-ablation results show that the strongest position-level effects depend on both role polarity and document relevance. For positive roles, the most consistent decrease occurs when the `Last` group is ablated on irrelevant documents. Because the expected answer is `No`, this shows that last-token heads are causally important for the normal rejection of irrelevant passages.
Fourth, the clearest full-role-span effect appears for negative roles on irrelevant documents. Ablating the heads selected from the `Role-All` ranking substantially reduces support for the role-induced `Yes` answer. This demonstrates that the full-role-span group contributes to the harmful false-positive tendency produced by negative role prompts. However, the same group does not show one fixed answer-direction effect across all role–label conditions.
Fifth, the `Last` group has a broader effect than the role-related groups. Its intervention affects positive and negative roles as well as relevant and irrelevant documents, although the direction and magnitude differ across conditions. This is consistent with the final token serving as a decision-integration site where role-conditioned information is combined with query–document evidence.
Overall, the causal intervention results refine the head-selection analysis. Activation patching identifies heads and positions that recover role-conditioned information. Single-head zero ablation reveals whether individual heads support or suppress the label-aware decision direction. Top-20 group ablation with the role-conditioned correct score shows how broader target-position head groups support positive and negative role behavior differently.

6.4 Summary of Mechanistic Findings

Across datasets and patching settings, the mechanistic experiments reveal six main patterns.
First, role-prompt information is recoverable around role-related prompt positions. In both residual-stream and attention-head patching, the `Role-All` and `Role-Adj-Adv` components show clear recovery in early and middle layers. At the head level, the `Role-Adj-Adv` condition repeatedly highlights early heads such as `L3H24`, especially in the irrelevance setting. This suggests that the model first encodes lexical role information locally around the role phrase before this information affects later decision-related components.
Second, the final Yes/No decision position plays a central role. Both residual-stream patching and attention-head patching show that the strongest recovery is concentrated at the `Last` token. The last-token effects are especially clear in middle-to-late layers, including

heads such as L14H24, L15H4, L15H7, L15H8, and L16H29. This indicates that role-prompt information is not only represented locally in the prompt span, but is routed toward the final position where the model prepares its Yes/No judgment.

Third, the relevant and irrelevant settings show asymmetric internal behavior. In the activation-patching results, irrelevance patching generally produces stronger and more structured recovery than relevance patching, especially at the Last token.

However, the zero-ablation results show that this asymmetry is more nuanced than a simple "irrelevance is always stronger" pattern. For relevance examples, the largest position-wise ablation effects are observed mainly in DL19, DL20, and TREC-COVID, whereas stronger irrelevance-side effects appear in SciFact, FIQA, and Climate-Fever. This contrast shows that the relative sensitivity of relevance and irrelevance judgments varies across datasets.

Fourth, the influential heads are sparse but functionally mixed. The Top-10 helpful-head analysis identifies a small set of heads that repeatedly achieve high recovery scores. Zero ablation confirms that these heads causally affect the final label-aware logit difference. However, the ablation direction matters. Negative ablation deltas indicate heads that directly support the correct decision direction, because removing them weakens the target logit difference. For relevance judgments, heads such as L15H10 and L16H31 often behave as relevance-supporting heads. For irrelevance judgments, heads such as L15H4, L15H8, and L14H24 often support the No-over-Yes direction. By contrast, some selected heads produce positive ablation deltas, meaning that removing them strengthens the target direction. These heads are therefore better interpreted as suppressive or regulatory components rather than uniformly helpful heads.

Fifth, the top-20 zero-ablation results reveal condition-specific position-level effects. For positive roles, the strongest and most consistent effect appears at the Last token on irrelevant documents, where ablation weakens support for the expected No answer. For negative roles, the clearest full-role-span effect appears on irrelevant documents: ablating the heads selected from the Role-All ranking weakens the role-induced Yes tendency. This identifies a causal contribution to the harmful false-positive behavior produced by negative prompts. At the same time, the Role-All group does not exhibit one fixed answer-direction function across all conditions. The Last group affects a broader set of role–label conditions, consistent with its role as a final decision-integration site.

Sixth, the routing pattern generalizes across datasets but still reflects dataset structure. DL19 and DL20 show highly similar head-patching and zero-ablation patterns, which is expected because both are based on the MS MARCO passage corpus. TREC-COVID follows the same broad routing structure and also shows stronger sensitivity in the irrelevance-side zero-ablation setting than in the relevance-side setting. The additional BEIR-style datasets, especially SciFact, FIQA, and Climate-Fever, preserve the same relevance–irrelevance asymmetry, although the magnitude of the effects varies by dataset. Thus, the role-prompt mechanism is partly transferable across domains, while its strength depends on dataset and task characteristics.

Taken together, these findings show that role-playing effects in zero-shot LLM ranking are not only behavioral artifacts. They correspond to identifiable internal pathways: role information is recoverable around role-related prompt positions, routed through a sparse set of attention heads, and later concentrated at the last token where the Yes/No decision is formed. At the same time, the zero-ablation results show that the selected heads do not all support the target decision direction. Some heads support the label-aware di-

rection, because ablating them weakens the target logit difference. Other heads have a suppressive or regulatory effect, because ablating them strengthens the target direction. These supporting and suppressive effects also differ between relevance and irrelevance judgments.

Chapter 7 Discussion

The results of this thesis suggest that harmful role-playing in zero-shot LLM ranking should be understood at two connected levels. At the prompt level, the strongest degradation comes from negative adjective–adverb framing rather than from role identity alone. At the model level, role-conditioned information is recoverable around role-related prompt positions and later influences sparse decision-relevant attention heads near the final Yes/No prediction.

7.1 Prompt-level harmfulness

The behavioral experiments show that harmful role prompts are not explained by isolated negative words. Adjective-only and adverb-only prompts are less damaging than combined adjective–adverb prompts, and the strongest failures appear only when multiple negative lexical cues occur together. This indicates that the model is sensitive to prompt composition: harmfulness depends not only on whether a word is negative, but also on how negative lexical cues are combined inside the role prompt.

The role-only control further shows that role identity alone is not sufficient to explain the large ranking failures. Meaningful roles, unrelated roles, object-like roles, and meaningless role strings remain relatively stable on TREC-COVID when negative lexical framing is removed. However, when a fixed harmful lexical frame is present, changing the role identity can still change the degree of degradation. This suggests that role identity is best interpreted as a contextual modulator of harmful lexical framing, rather than as the primary source of harmfulness.

This distinction matters for prompt robustness. A role prompt may appear harmless when considered as a role name alone, but become harmful when combined with strongly negative adjective–adverb wording. Therefore, role-prompt robustness should be evaluated compositionally rather than only by testing individual role identities.

7.2 Internal routing of role-prompt information

The mechanistic results provide an internal explanation for these behavioral effects. Activation patching shows that role-related information is recoverable around the role span, especially at the full role phrase and the combined adjective–adverb component. The strongest recovery then appears at the final token, where the model prepares the Yes/No answer. This suggests that role information is first represented around the prompt region where it is introduced and later contributes to decision-level representations.

However, the ablation results show that these components are not uniformly helpful. Activation patching identifies heads whose activations are useful for recovering role-conditioned behavior, but zero ablation is needed to determine how those heads function. Some heads support the target logit direction, because ablating them weakens the label-aware decision. Other heads have the opposite effect, suggesting suppressive or regulatory

behavior. Thus, the selected heads should be interpreted as decision-relevant components rather than as a uniformly helpful circuit.
The group-ablation results refine this picture. For positive roles, the most important intervention effects appear at the `Last` position, especially for irrelevant documents. For negative roles, the full `Role-All` span is more important for the role-conditioned tendency to answer `Yes` on irrelevant documents. This is directly relevant for ranking, because increasing support for `Yes` on irrelevant passages can push non-relevant documents upward in the ranked list.

7.3 Asymmetry between relevance and irrelevance

Another important finding is that relevance and irrelevance are not controlled by exactly the same internal behavior. The selected relevance heads and irrelevance heads differ, and their zero-ablation effects are not symmetric.
For relevance examples, several heads mainly affect the Yes-over-No direction for relevant documents. Heads such as `L15H10` and `L16H31` often produce negative deltas when ablated, indicating that they support the correct relevance decision. However, other relevance-selected heads produce positive deltas, meaning that they may suppress or regulate the relevance direction.
For irrelevance examples, the selected heads show a clearer functional split. Heads such as `L15H4`, `L15H8`, and `L14H24` support the No-over-Yes direction, because ablating them weakens the rejection of irrelevant documents. Other heads, such as `L11H0`, `L13H23`, `L15H7`, and `L3H24`, often produce positive deltas, suggesting suppressive or regulatory roles.
This asymmetry is meaningful for pointwise ranking. A Yes/No ranker must both identify documents that provide relevant evidence and reject documents that do not provide sufficient evidence. The mechanistic results suggest that these two operations rely on partially distinct internal pathways rather than a single circuit used in opposite directions.

7.4 Implications and limitations

These findings suggest that role prompts should be treated as part of the ranking method itself, not as harmless stylistic additions. Changing the assigned role can change the document scores used for ranking, especially when the role prompt contains harmful lexical framing. For this reason, future evaluations of zero-shot LLM rankers should report the exact prompt wording, role description, lexical framing, and output format.
There are also several limitations. First, all experiments use a single instruction-tuned model, `Llama-3.1-8B-Instruct`, so the identified head indices are model-specific. Second, the ranking setup is pointwise and based on Yes/No relevance judgments; pairwise, listwise, or setwise ranking may show different prompt sensitivities. Third, the behavioral evaluation relies mainly on NDCG@10. This is appropriate for ranking because it focuses on top-ranked documents, but it does not directly measure calibration, binary classification accuracy, recall, or score stability across the full ranked list. Fourth, the mechanistic experiments focus on the Random20 prompt set rather than the newly constructed worst adjective–adverb prompts. This keeps the mechanistic analysis aligned with the initial role-prompt evaluation, but it also means that the strongest controlled lexical combinations are not fully traced mechanistically. Finally, the ablation experiments identify causal effects of selected heads, but they do not fully reveal the communication path

between early role-span representations and later decision heads. Future work could use path patching or circuit-level interventions to trace this pathway more directly.

Chapter 8 Conclusion

In this thesis, I investigated why and how role-playing prompts can harm zero-shot LLM rankers and how these effects are represented inside the model. The study combined behavioral ranking experiments with mechanistic interpretability analysis. The behavioral experiments decomposed harmful prompts into lexical polarity, role identity, and their interaction, while the mechanistic experiments traced how role-prompt information is encoded and routed through the transformer.

The results answer the four research questions as follows.

First, harmful negative prompt effects are driven more strongly by adjective–adverb combinations than by isolated lexical cues. Adjective-only prompts produce limited degradation, adverb-only prompts can be more harmful, and adjective–adverb paired prompts produce the strongest failures. This shows that harmful role-playing is compositionally structured rather than caused by a single negative word.

Second, harmful adjective–adverb combinations transfer across datasets. Pairs selected on TREC-COVID remain harmful on TREC-DL 2019, TREC-DL 2020, SciFact, FIQA, and Climate-Fever, although the magnitude of degradation differs by domain. Thus, harmful lexical effects are not purely dataset-specific, but their severity depends on the retrieval task.

Third, role identity is not the primary source of harmfulness by itself. The role-only control on TREC-COVID shows that meaningful, unrelated, object-like, and meaningless role prompts remain relatively stable and do not produce severe degradation. However, under fixed harmful lexical framing, changing the role identity can still change ranking performance. Thus, role identity mainly acts as a modulating factor, while adjective–adverb lexical framing remains the strongest driver of degradation.

Fourth, role-prompt information is routed through identifiable internal components. Activation patching shows that differences between positive and negative role conditions are first recoverable around role-related prompt positions and later become concentrated at the final decision token. Single-head zero ablation confirms that the selected attention heads causally affect the label-aware `Yes`/`No` decision, while top-20 group ablation reveals condition-specific effects at the `Role-All` and `Last` positions. In particular, the full-role-span group contributes to the harmful negative-role-induced `Yes` tendency on irrelevant documents, whereas the last-token group plays a broader role in integrating role and query–document information before the final decision.

The selected heads are sparse but functionally mixed: some support the ground-truth-aligned decision direction, whereas others suppress or regulate it. Relevance and irrelevance judgments also emphasize different heads and intervention effects, suggesting that accepting relevant documents and rejecting irrelevant documents rely on partially distinct internal computations.

Overall, this thesis shows that harmful role-playing in zero-shot LLM rankers is both prompt-structured and mechanistically localized. Role prompts do not merely change output style; they can change the internal computation underlying relevance judgment. The strongest harmful effects arise from compositional prompt features, and their influence

is routed through sparse but functionally mixed attention-head pathways.
Future work should trace the exact communication pathways between early role-token encoding and late decision heads using path patching. It should also test whether the strongest controlled adjective–adverb prompts use the same circuits as the Random20 prompts. Finally, future intervention work should distinguish between heads that support correct ranking behavior and heads that suppress or regulate it, so that harmful prompt sensitivity can be reduced without damaging normal relevance judgments.

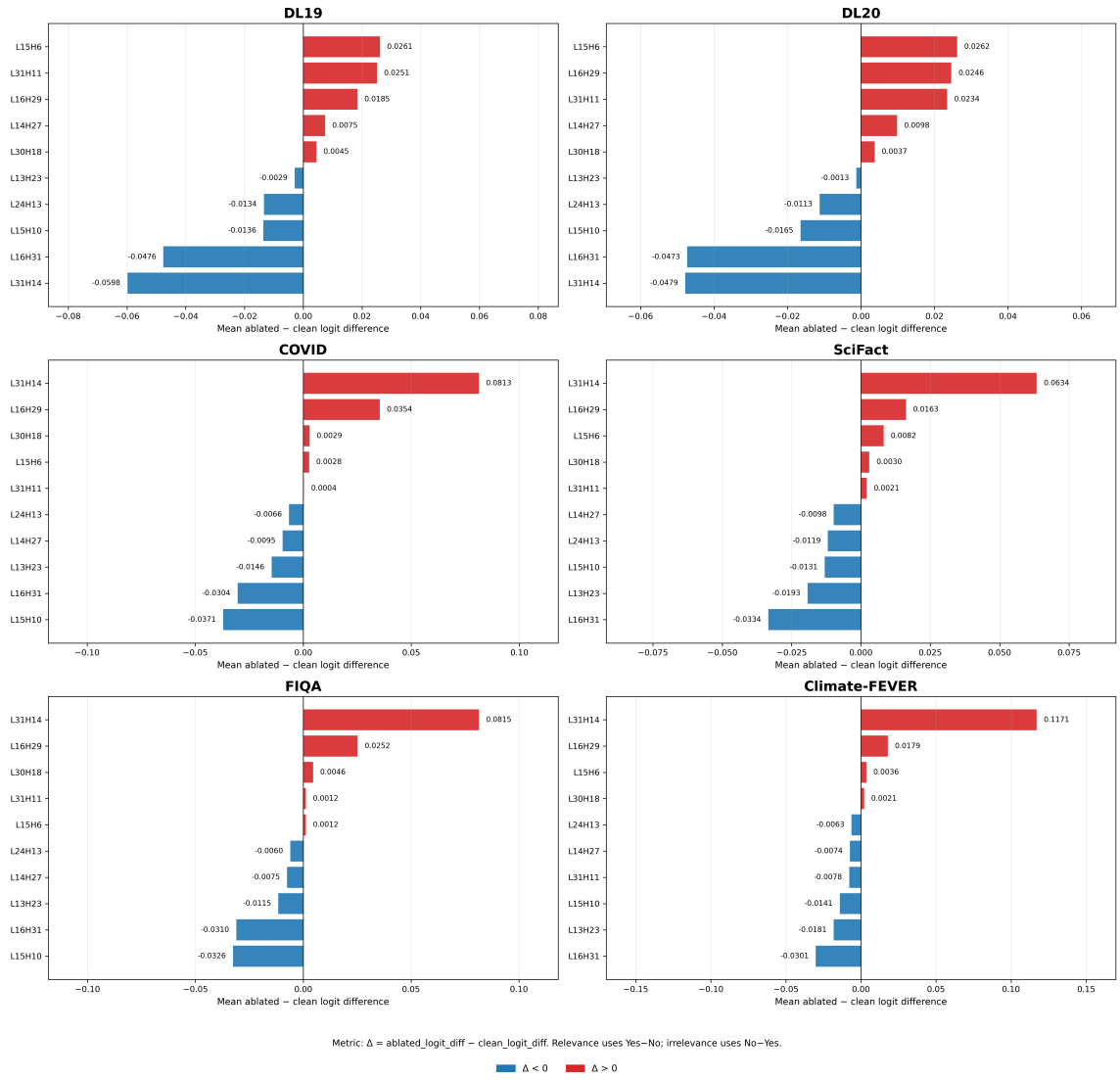


Figure 15: Zero-ablation delta for the selected relevance heads on relevance examples across six datasets, averaged over all 20 Random20 role prompts.

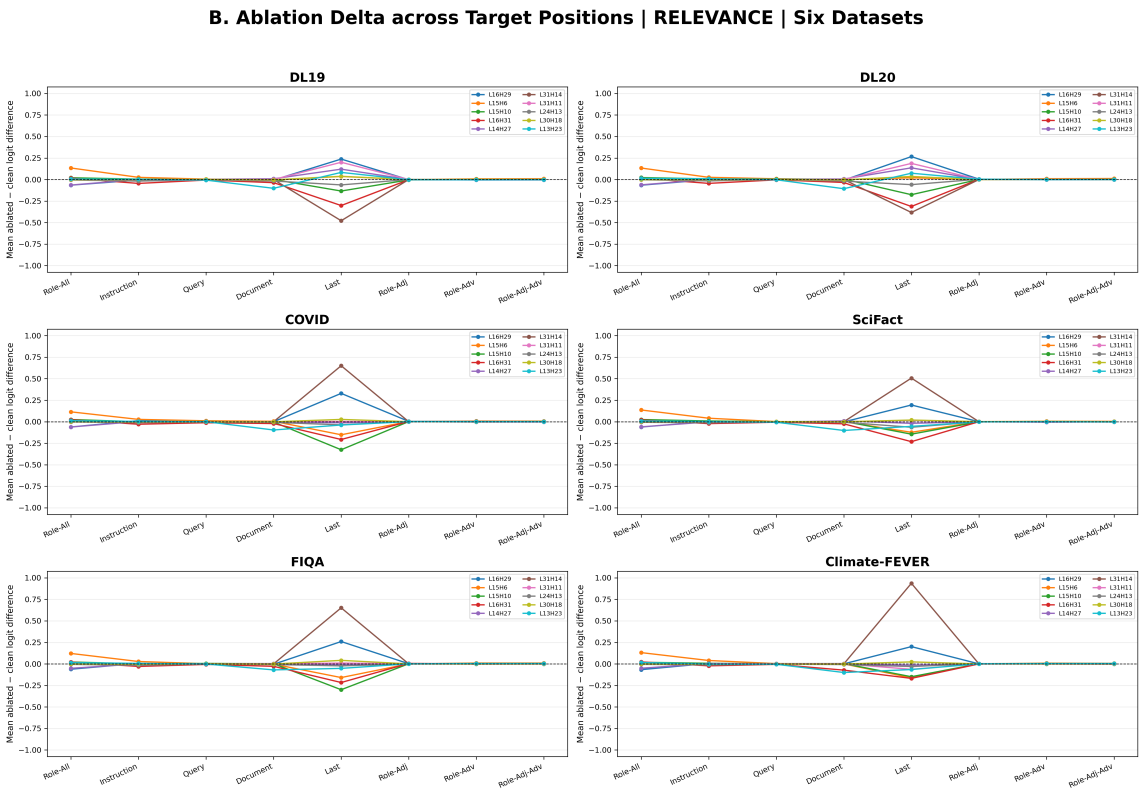


Figure 16: Zero-ablation delta of the selected relevance heads across target positions, averaged over all 20 Random20 role prompts.

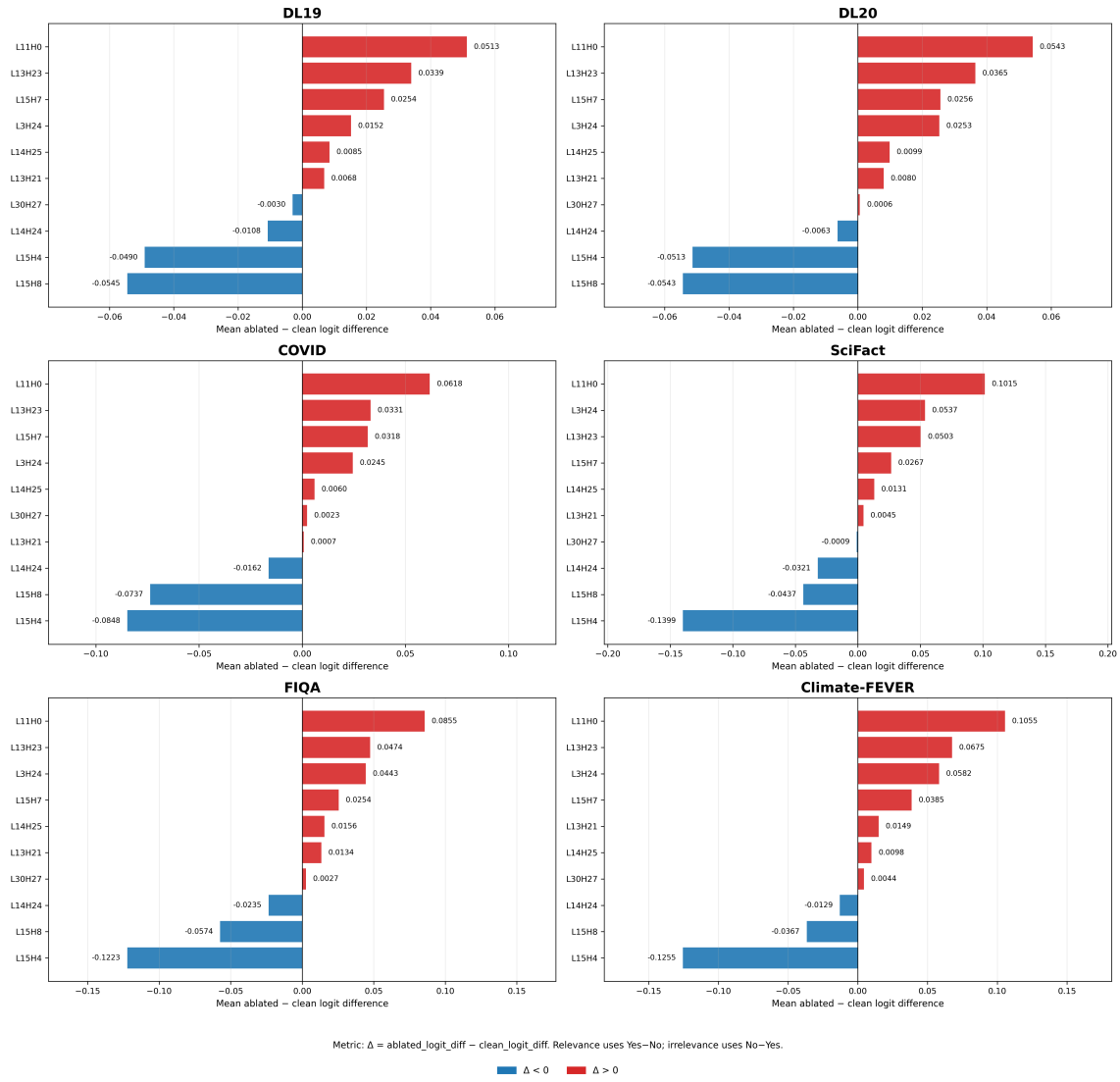


Figure 17: Zero-ablation delta for the selected irrelevance heads on irrelevance examples across six datasets, averaged over all 20 Random20 role prompts.

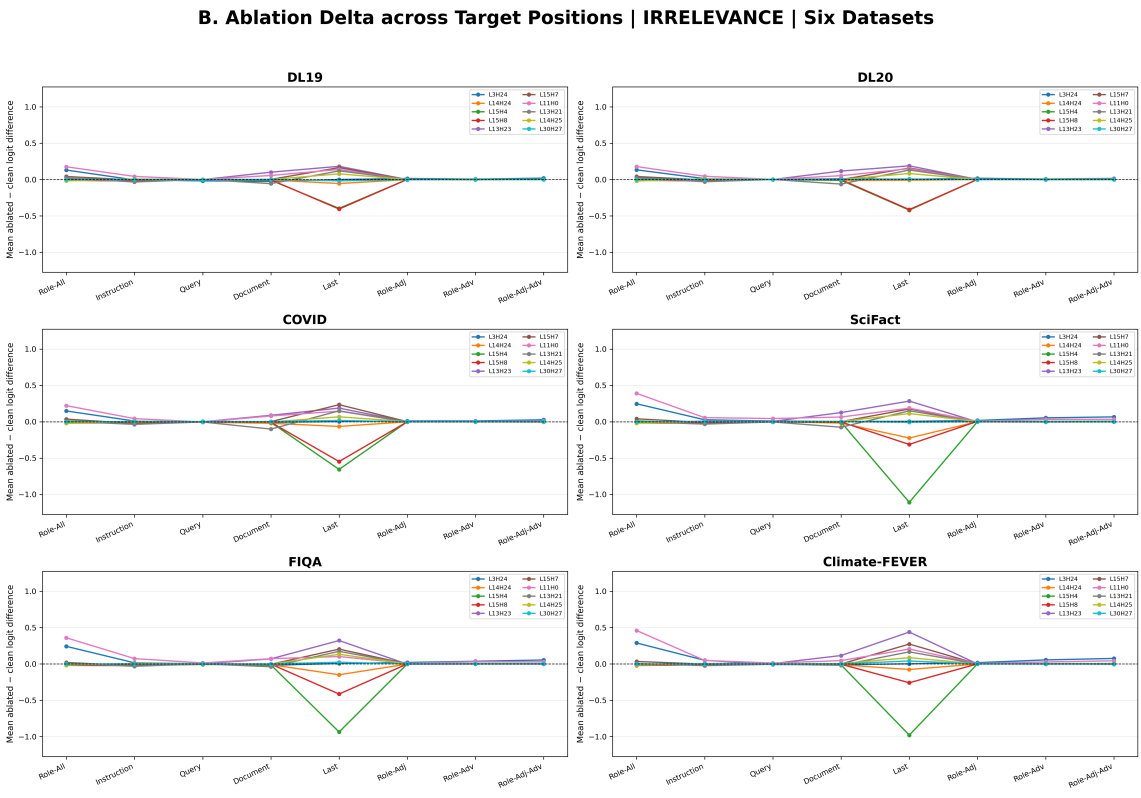


Figure 18: Zero-ablation delta of the selected irrelevance heads across target positions, averaged over all 20 Random20 role prompts.

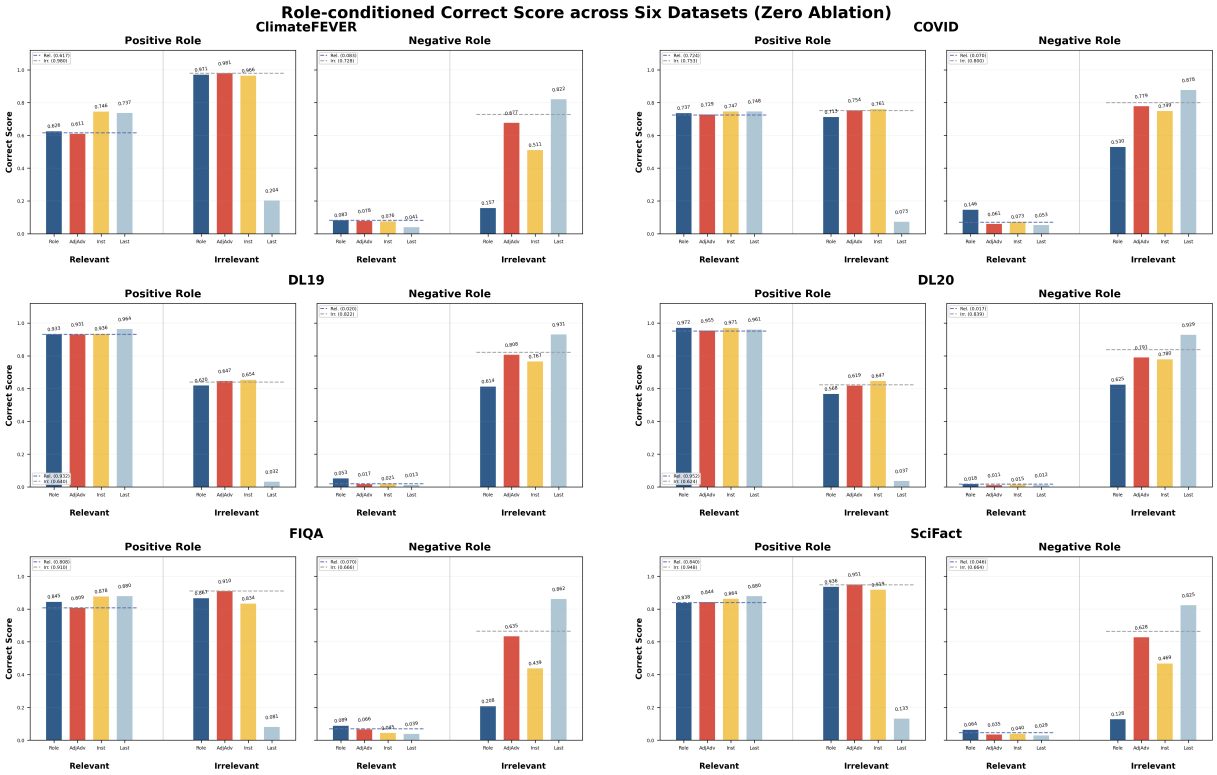


Figure 19: Role-conditioned correct score after zero ablation of the top-20 candidate heads across six datasets. Results are separated by positive and negative role prompts and by relevant and irrelevant documents. Dashed lines indicate the corresponding clean-model scores. A lower ablated score indicates that the selected head group supports the role-conditioned expected answer in that condition.

Appendix A Additional Behavioral Selection Tables

This appendix reports the intermediate selection tables used in the behavioral role–lexical interaction experiments. These tables complement the main-text analysis in Section 5. They show how the harmful adjective–adverb pairs and role–lexical combinations were selected before cross-dataset evaluation.

A.1 Role-only control results on TREC-COVID

This appendix reports the complete role-level results for the role-only control experiment on TREC-COVID. The purpose of this control is to test whether role identity alone is sufficient to cause large ranking degradation when no negative adjective or adverb is used. The experiment compares several role-only groups: domain-specific medical roles, non-domain semantic occupations, unrelated semantic roles, object-like unusual roles, and meaningless string roles. All prompts use the same pointwise ranking instruction and differ only in the inserted role phrase. The group-level summary and main interpretation are reported in Table 7 in the main text. Table 11 provides the complete role-level NDCG@10 scores used to compute that summary.

Overall, the full role-level results support the main-text conclusion: role identity alone does not cause the severe degradation observed for negative adjective–adverb prompts. Instead, most role-only prompts remain close to each other and above the no-role baseline.

Table 11: Complete TREC-COVID role-only control results. Values are NDCG@10.

Group	Role ID	NDCG@10	Role phrase
Baseline	role0	0.7278	no role
Domain-specific medical roles	role1	0.7972	infectious disease medical specialist
	role2	0.7984	respiratory medicine medical specialist
	role3	0.7927	epidemiology research medical specialist
	role4	0.7808	virology laboratory medical specialist
	role5	0.7850	immunology clinical medical specialist
	role6	0.7954	public health medical specialist
	role7	0.7960	global health medical specialist
	role8	0.7934	emergency clinical medical specialist
	role9	0.7837	critical care medical specialist
	role10	0.7824	vaccine research medical specialist
Non-domain semantic occupations	role11	0.7950	retail shop sales clerk
	role12	0.7697	urban street fashion photographer
	role13	0.7868	online multiplayer video gamer
	role14	0.7918	independent bookstore shop assistant
	role15	0.7862	digital marketing agency intern
	role16	0.7769	residential property rental agent
	role17	0.7904	pizza restaurant service waiter

Continued on next page

Table 11: Complete TREC-COVID role-only control results. Continued.

Group	Role ID	NDCG@10	Role phrase
	role18	0.7976	video editing studio assistant
	role19	0.7902	community theater stage actor
	role20	0.7807	wedding event planning assistant
Unrelated semantic roles	role1	0.7893	professional museum curator
	role2	0.7768	university history lecturer
	role3	0.7848	wildlife documentary filmmaker
	role4	0.7763	professional orchestra conductor
	role5	0.7826	marine biology research assistant
	role6	0.7837	space mission control operator
	role7	0.7862	railway systems engineer
	role8	0.7757	classical literature librarian
	role9	0.7801	professional chess tournament referee
	role10	0.7791	archaeological field researcher
Object-like unusual roles	role11	0.7815	volcanic mountain rock
	role12	0.7885	wandering desert cactus
	role13	0.7826	sleeping farm sheep
	role14	0.7727	broken kitchen microwave
	role15	0.7752	floating soap bubble
	role16	0.7726	melting ice cube
	role17	0.7921	wooden park bench
	role18	0.7778	tired household goldfish
	role19	0.7730	rusty bicycle chain
	role20	0.7897	drifting cloud in the sky
Meaningless string roles I	role1	0.7940	`xQ7m !@# vrK92`
	role2	0.7994	`zlp_88 %%& Travax`
	role3	0.7933	`Qm!ta 42xyz ^^^^`
	role4	0.8004	`blorf7 <>? kL_22`
	role5	0.8202	`R8R8R ~~~ pexton`
	role6	0.8037	`@@mix yora13 kkk!`
	role7	0.8048	`Nivrox ;;; y7_y7`
	role8	0.8001	`7delta qx!qx %room`
	role9	0.7874	`*flux9 Brrr_4 zeta?`
	role10	0.7896	`klm88 $$nova UuUu`
Meaningless string roles II	role11	0.7943	`brakton velmir qostel`
	role12	0.7946	`trunyik pelzor naxlin`
	role13	0.7839	`morvex quilta zarnok`
	role14	0.7854	`drim42 K?K?K? @void`
	role15	0.7939	`00zero qqq!! Brasta`

Continued on next page

Table 11: Complete TREC-COVID role-only control results. Continued.

Group	Role ID	NDCG@10	Role phrase
	role16	0.7787	`Vortex? 9nine9 ===`
	role17	0.7980	`pLuto_7 x!x!x! Morphi`
	role18	0.7926	`%%moon G7G7 Vraxen`
	role19	0.7927	`Tiki_88 ?!?! drongo`
	role20	0.7809	`haxxor7 zZz99 ##orbit`

A.2 Selection of harmful adjective–adverb pairs

Table 12 reports the TREC-COVID NDCG@10 scores for the adjective-only, adverb-only, and adjective–adverb paired prompts. The highlighted cells indicate the ten lowest-performing adjective–adverb pairs selected for the subsequent role–lexical interaction experiment.

A.3 Selection of role–adjective–adverb combinations

Table 13 reports the TREC-COVID NDCG@10 scores for the 10×10 role–lexical grid. Each column corresponds to one selected harmful adjective–adverb pair from Table 12. Each row corresponds to one selected role identity. The highlighted cells indicate the final role–adjective–adverb combinations selected for cross-dataset evaluation.

Table 12: TREC-COVID NDCG@10 scores for adjective–adverb lexical ablation. Boxed cells indicate the ten selected harmful adjective–adverb pairs.

Adv./Adj.	faulty	clumsy	incorrect	unreliab.	hopeless	confused	messy	awful	problem.	flawed	adv only
badly	0.7713	0.7862	0.7338	0.6977	0.7602	0.7905	0.7668	0.5789	0.7135	0.7939	0.7870
wrongly	0.7622	0.7530	0.6894	0.6565	0.5336	0.7381	0.7392	0.3163	0.6999	0.7919	0.7982
mistakenly	0.7770	0.7786	0.7507	0.6705	0.7536	0.7692	0.7782	0.3247	0.7485	0.7932	0.7968
improperly	0.7845	0.7952	0.7219	0.7137	0.7324	0.7819	0.7798	0.4846	0.7423	0.7973	0.7986
falsely	0.6657	0.6532	0.6076	0.5139	0.4926	0.6145	0.6677	0.1930	0.6086	0.7372	0.7385
terribly	0.7562	0.7577	0.7221	0.6713	0.7444	0.7780	0.7698	0.5871	0.7089	0.7965	0.7853
poorly	0.7576	0.7686	0.7434	0.6982	0.7606	0.7794	0.7592	0.5892	0.6927	0.7786	0.7927
incorrectly	0.7731	0.7722	0.7193	0.6890	0.6094	0.7679	0.7541	0.3390	0.7169	0.7937	0.7648
sadly	0.7800	0.7864	0.7831	0.7478	0.7911	0.7831	0.7759	0.7014	0.7761	0.7886	0.7955
horribly	0.7561	0.7707	0.7405	0.6643	0.7725	0.7853	0.7749	0.6133	0.7129	0.7770	0.7957
adj only	0.7969	0.7816	0.7910	0.7738	0.7854	0.7919	0.7877	0.7594	0.7819	0.7926	–

Table 13: TREC-COVID NDCG@10 scores for the role–adjective–adverb interaction grid. Boxed cells indicate the final selected combinations used for cross-dataset evaluation.

Role	awful falsely	hopeless wrongly	unreliable horribly	problem poorly	incorrect improperly	messy incorrectly	clumsy terribly	confused mistakenly	faulty badly	flawed sadly
R1	0.5657	0.7235	0.7761	0.7683	0.7342	0.7714	0.7950	0.7752	0.7830	0.7882
R2	0.6878	0.7349	0.7522	0.7550	0.7494	0.7547	0.7755	0.7769	0.7768	0.7889
R3	0.5215	0.7335	0.7558	0.7743	0.7619	0.7631	0.7865	0.8034	0.7901	0.7990
R4	0.4423	0.7102	0.7572	0.7701	0.7048	0.7744	0.7856	0.7641	0.7802	0.7806
R5	0.4734	0.6812	0.7130	0.7462	0.6998	0.7621	0.7820	0.7696	0.7729	0.7833
R6	0.5979	0.7300	0.7368	0.7708	0.7340	0.7807	0.7663	0.7819	0.7812	0.7827
R7	0.6273	0.7208	0.7571	0.7615	0.7499	0.7597	0.7894	0.7821	0.7821	0.7684
R8	0.4191	0.6857	0.7675	0.7793	0.7373	0.7794	0.7931	0.7953	0.7900	0.7993
R9	0.6597	0.7207	0.7458	0.7573	0.7448	0.7623	0.7739	0.7838	0.7806	0.7749
R10	0.4018	0.6981	0.7536	0.7729	0.7154	0.7778	0.7746	0.7869	0.7775	0.7885

Role identities: R1 = retail shop sales clerk; R2 = urban street fashion photographer; R3 = online multiplayer video gamer; R4 = independent bookstore shop assistant; R5 = digital marketing agency intern; R6 = residential property rental agent; R7 = pizza restaurant service waiter; R8 = video editing studio assistant; R9 = community theater stage actor; R10 = wedding event planning assistant.

Appendix B Extended Activation Patching Results

This appendix presents the complete activation patching results that complement the analyses discussed in Section 6. The goal of these experiments is to identify where and how role-playing information influences the ranking behaviour of the LLM reranker.
I report results for both relevant and irrelevant query–document pairs across multiple datasets, including SciFact, FIQA, Climate-Fever, TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The appendix is organized into component-level analyses and attention head-level analyses.
Throughout the appendix:

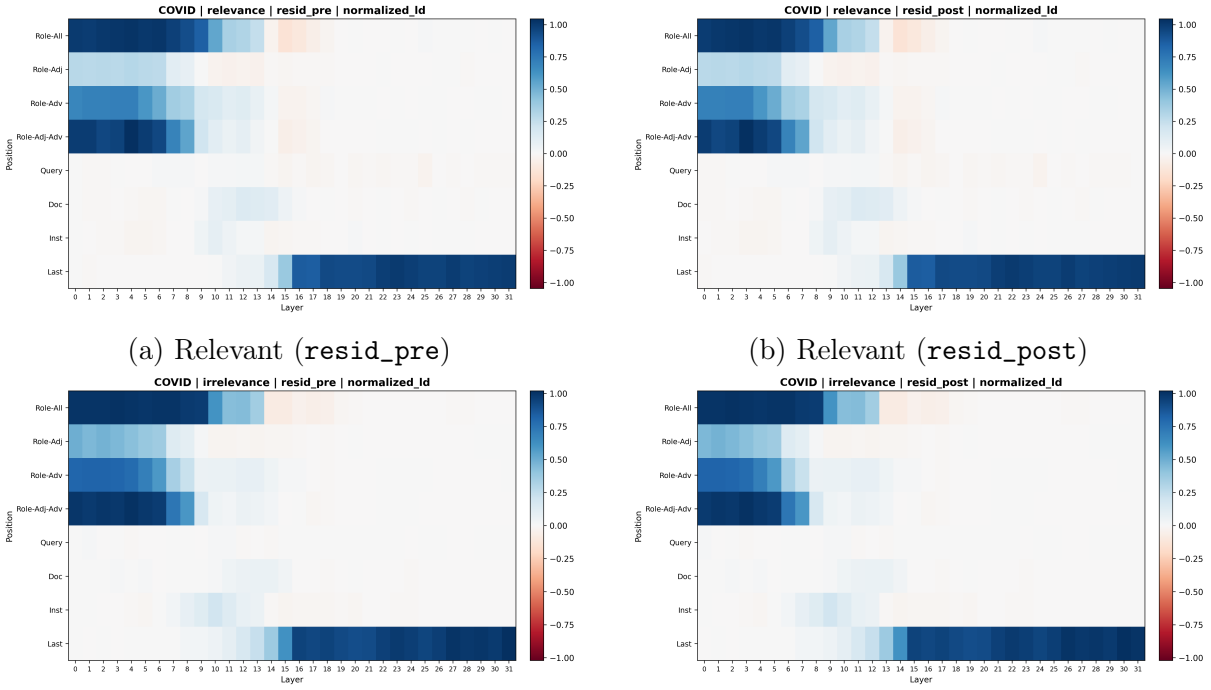
- **Relevance** refers to patching on relevant query–document pairs, where the target direction is logit(Yes) − logit(No).
- **Irrelevance** refers to patching on irrelevant query–document pairs, where the target direction is logit(No) − logit(Yes).
- **LD recovery** measures how much patching restores the label-aware logit difference between the clean and corrupted runs.
- **Role-All** denotes patching the entire role prompt span.
- **Role-Adj** denotes patching the adjective-related role span.
- **Role-Adv** denotes patching the adverb-related role span.
- **Role-Adj-Adv** denotes jointly patching the adjective and adverb role-prompt spans.
- **Inst** denotes the instruction segment.
- **Last** denotes the final token representation.

The results show that differences between positive and negative role conditions are recoverable around role-related positions in earlier layers and become concentrated at the final token in later layers.

Appendix C Component-level activation patching

This section reports component-level activation patching results using normalized logit difference recovery. Across datasets, the results demonstrate a highly consistent pattern: role-related information is concentrated in early layers while ranking-related information emerges in later layers.

C.1 TREC-COVID Component-Level Results



(a) Relevant (`resid_pre`)

(b) Relevant (`resid_post`)

(c) Irrelevant (`resid_pre`)

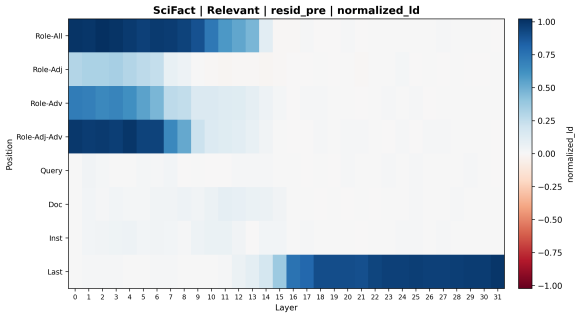
(d) Irrelevant (`resid_post`)

Figure 20: TREC-COVID component-level LD recovery showing the shift from role positions to the decision token.

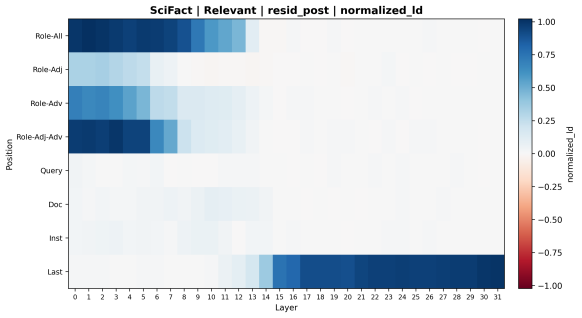
C.2 SciFact Component-Level Results

C.3 FIQA Component-Level Results

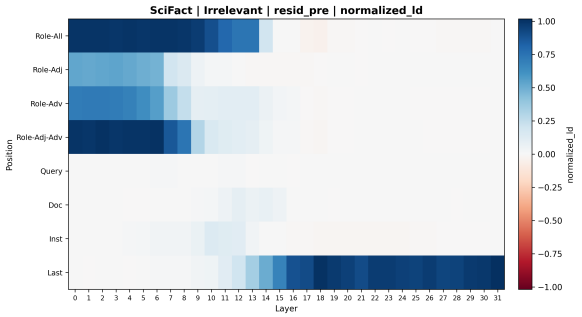
C.4 Climate-Fever Component-Level Results



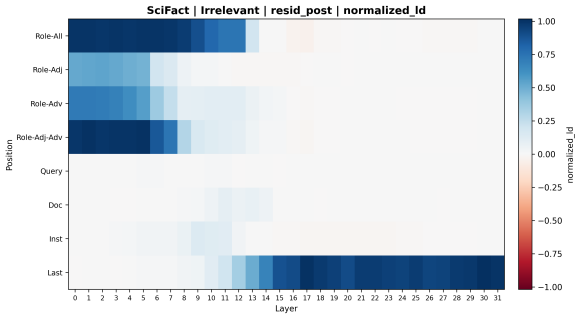
(a) Relevant (`resid_pre`)



(b) Relevant (`resid_post`)

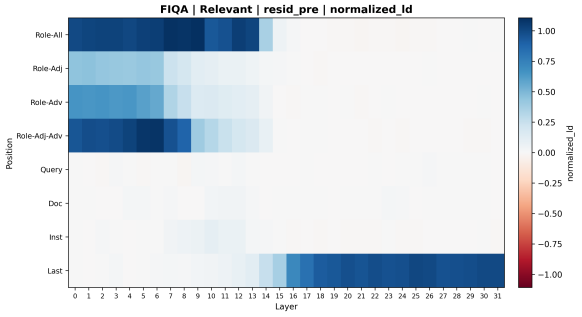


(c) Irrelevant (`resid_pre`)

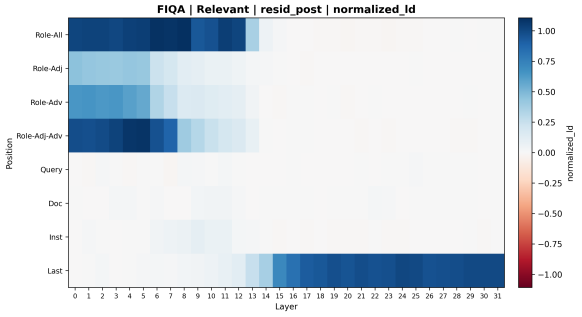


(d) Irrelevant (`resid_post`)

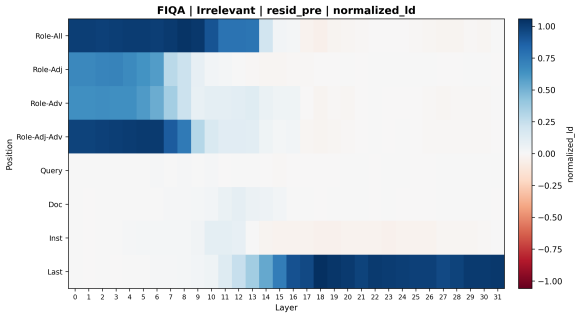
Figure 21: SciFact component-level LD recovery.



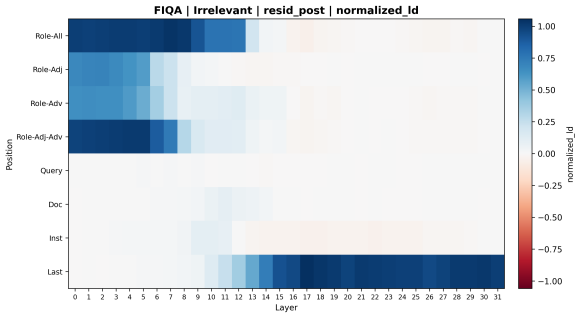
(a) Relevant (`resid_pre`)



(b) Relevant (`resid_post`)

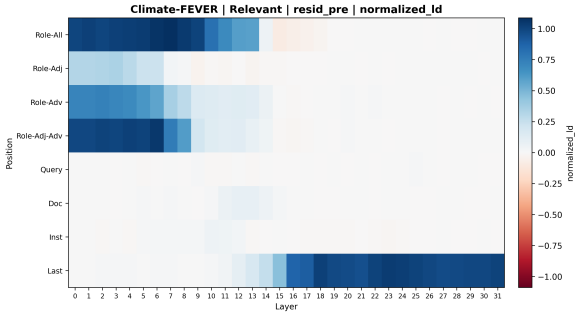


(c) Irrelevant (`resid_pre`)

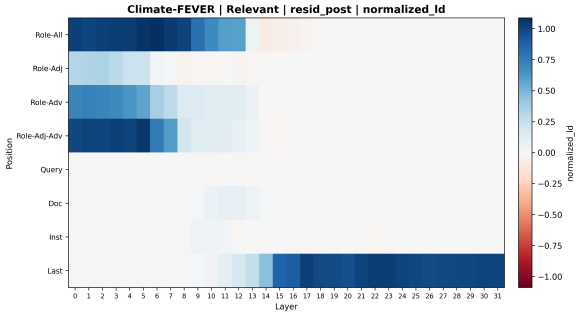


(d) Irrelevant (`resid_post`)

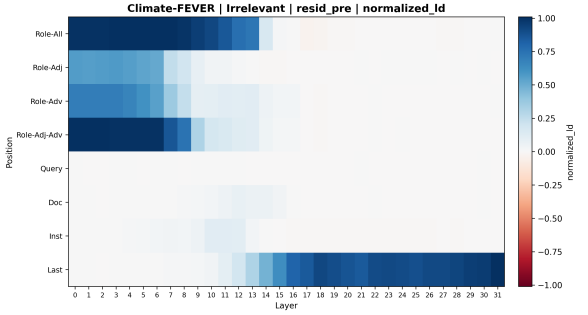
Figure 22: FIQA component-level LD recovery.



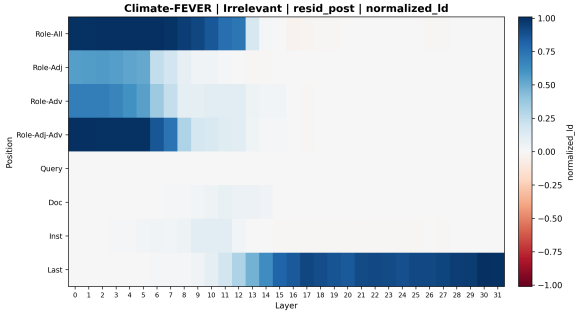
(a) Relevant (`resid_pre`)



(b) Relevant (`resid_post`)



(c) Irrelevant (`resid_pre`)



(d) Irrelevant (`resid_post`)

Figure 23: Climate-Fever component-level LD recovery.

Appendix D Attention head-level activation patching

This section presents the complete set of attention head-level activation patching results. For each dataset, I analyze both relevance and irrelevance query–document conditions, examining recovery across multiple prompt components: `Role-All`, `Role-Adj`, `Role-Adv`, `Role-Adj-Adv`, `Inst`, and `Last`.

The main text discusses the cross-dataset summary figures for TREC-DL 2019, TREC-DL 2020, and TREC-COVID. The figures below provide the corresponding detailed head-level results for the remaining datasets and prompt components. They are used to verify whether the relevance–irrelevance asymmetry and the routing pattern observed in the main text also appear in SciFact, FIQA, and Climate-Fever.

Overall, the appendix figures show the same qualitative pattern as the main text. The `Role-Adj-Adv` and role-related positions mainly highlight early or middle-layer heads, while the `Last` position produces the strongest and most structured recovery. The irrelevance condition is generally stronger than the relevance condition, indicating that rejecting irrelevant documents relies on a partially different and more pronounced set of heads.

D.1 Additional Role-Decomposition Head Patching Results

The main text reports the `Role-Adj-Adv` condition because it corresponds to the complete lexical role phrase and most directly matches the behavioral adjective–adverb analysis. For completeness, this appendix reports the separate `Role-Adj` and `Role-Adv` patching results.

These additional figures show that the isolated adjective and adverb spans produce weaker and less coherent recovery patterns than the joint `Role-Adj-Adv` condition. This supports the interpretation in the main text that the model does not treat the harmful prompt as two fully independent lexical cues. Instead, the combined role phrase is the more informative unit for recovering role-conditioned behavior.

D.2 SciFact Head-Level Results

D.2.1 SciFact: Relevant

D.2.2 SciFact: Irrelevant

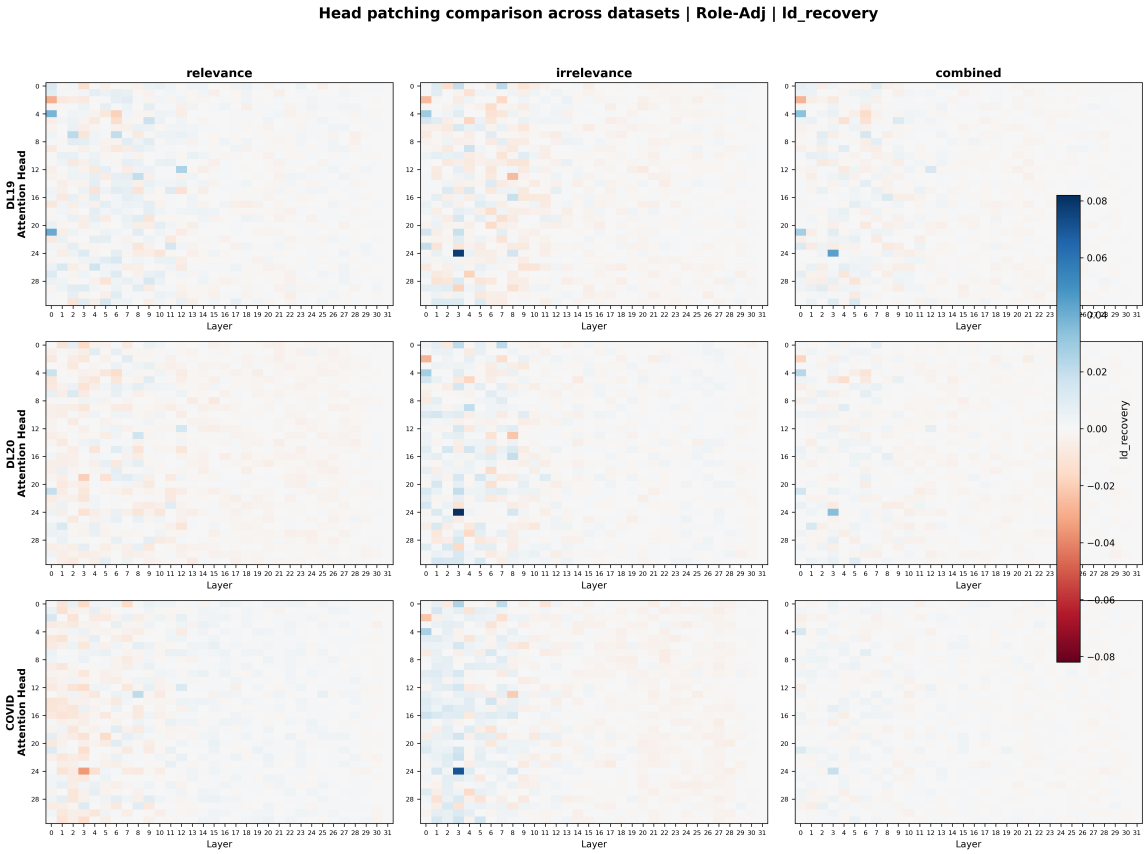


Figure 24: Cross-dataset attention head patching results for the `Role-Adj` prompt component.

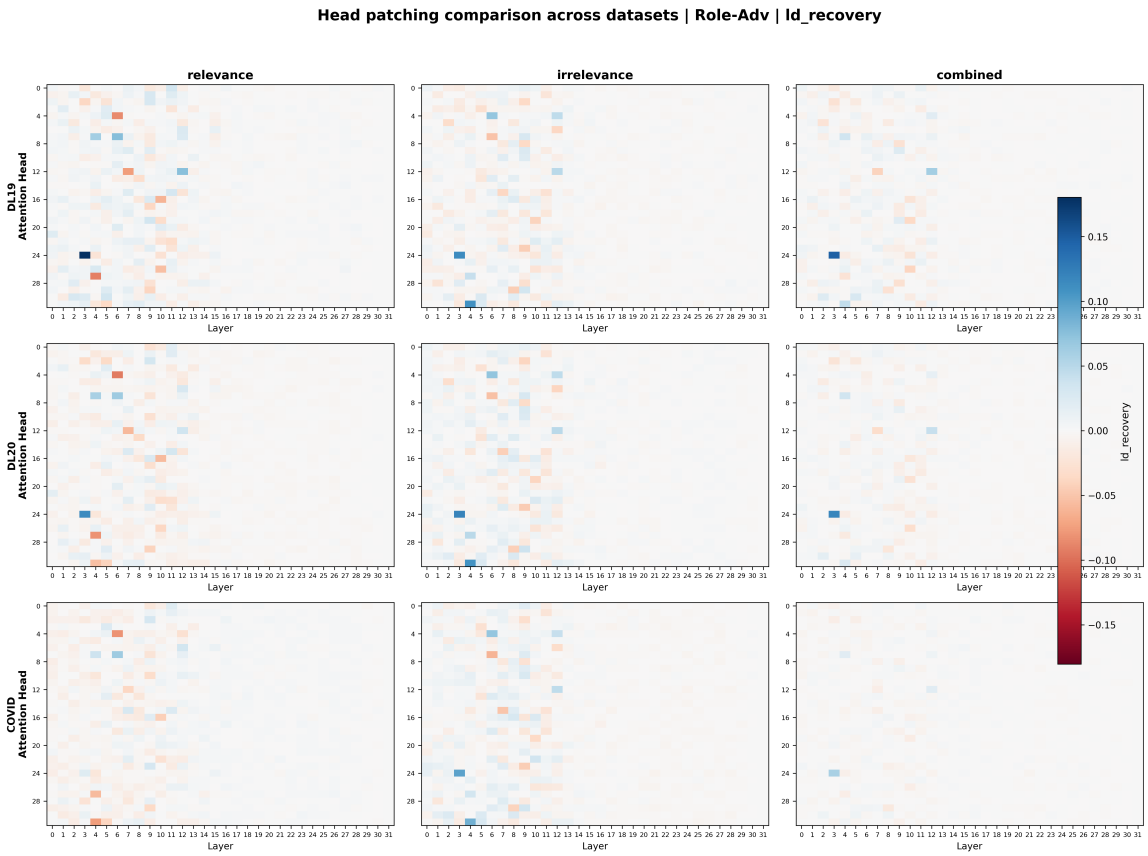
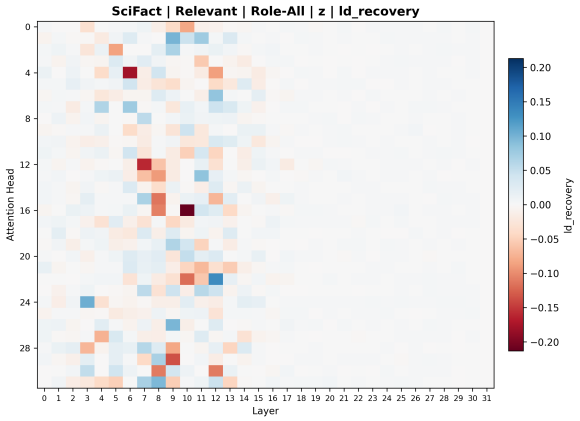
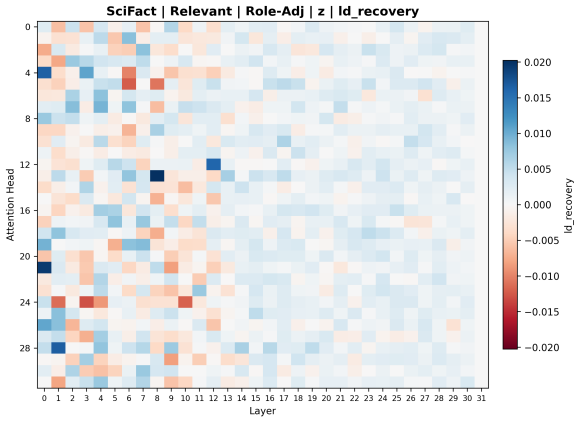


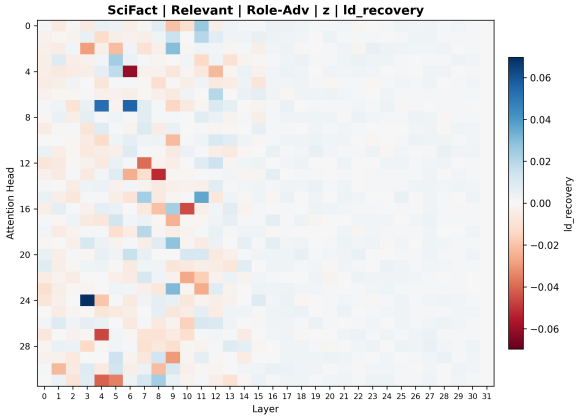
Figure 25: Cross-dataset attention head patching results for the `Role-Adv` prompt component.



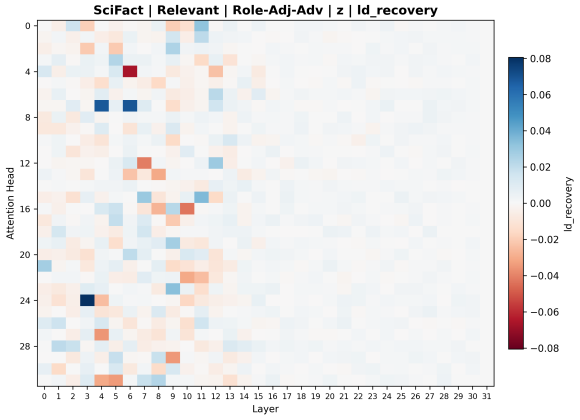
(a) Role-All



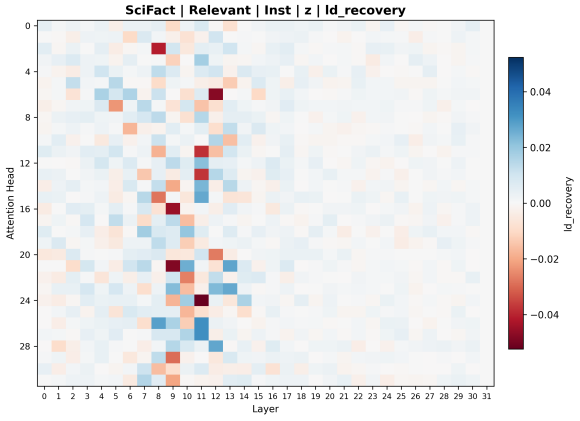
(b) Role-Adj



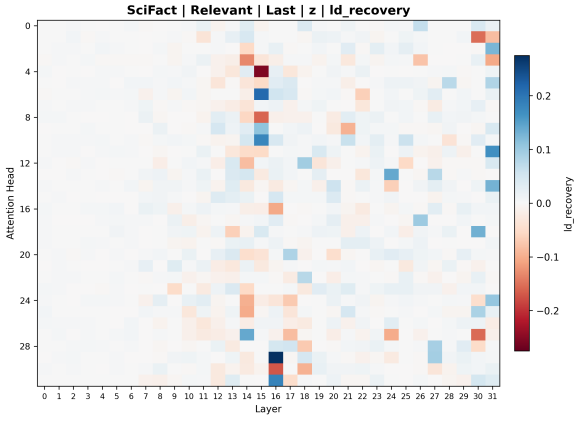
(c) Role-Adv



(d) Role-Adj-Adv

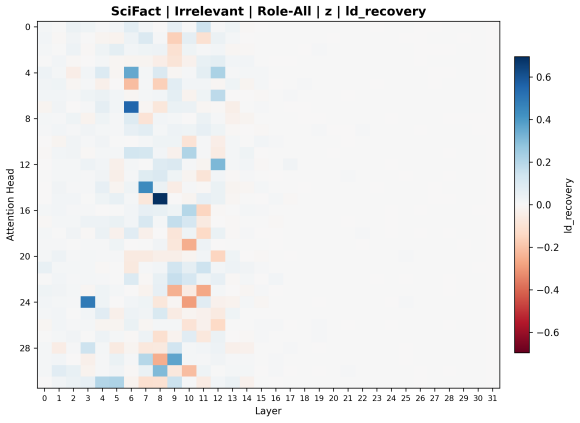


(e) Inst

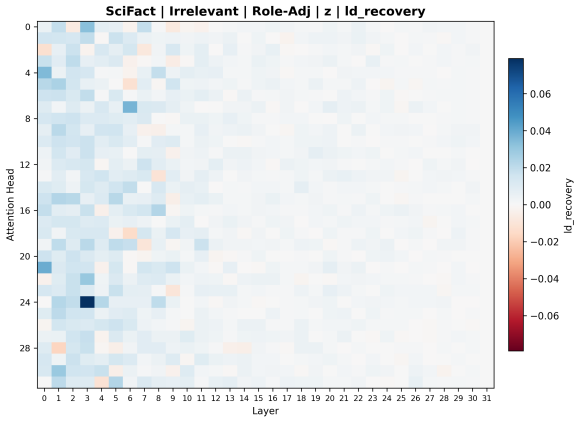


(f) Last

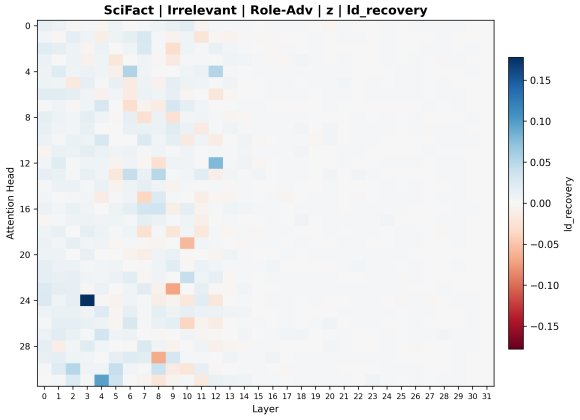
Figure 26: SciFact head-level LD recovery for **Relevant** role prompts.



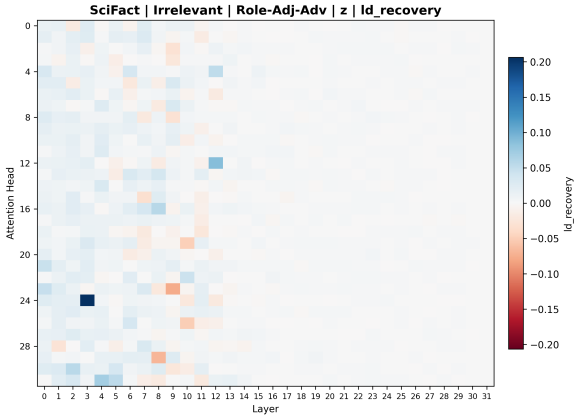
(a) Role-All



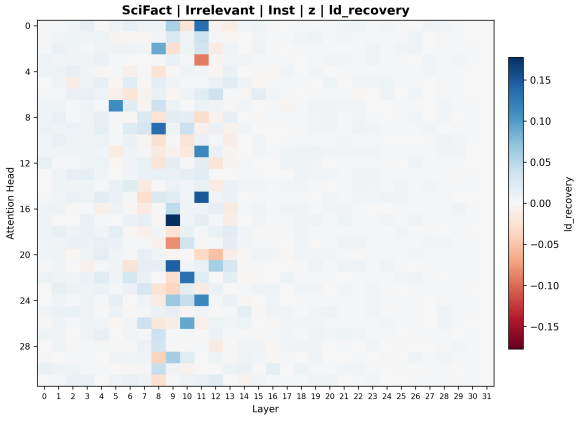
(b) Role-Adj



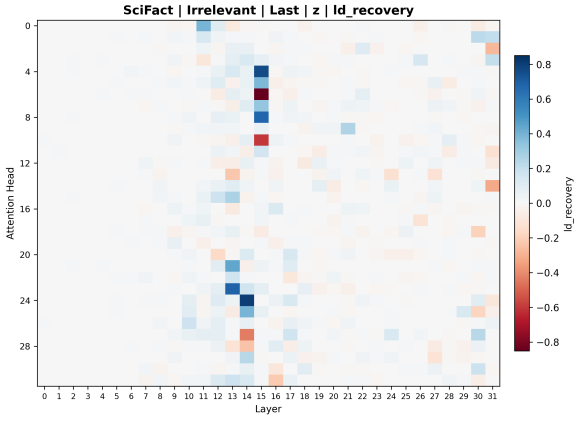
(c) Role-Adv



(d) Role-Adj-Adv



(e) Inst



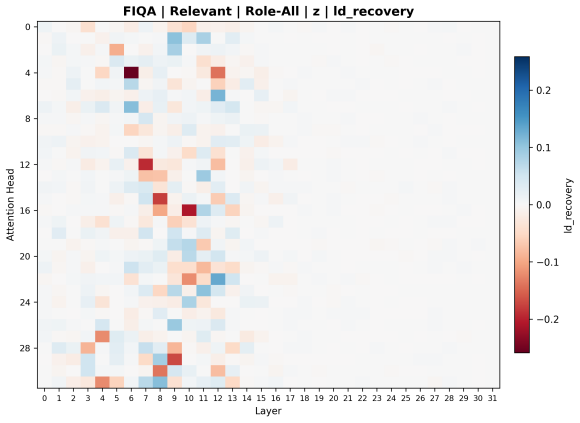
(f) Last

Figure 27: SciFact head-level LD recovery for **Irrelevant** role prompts.

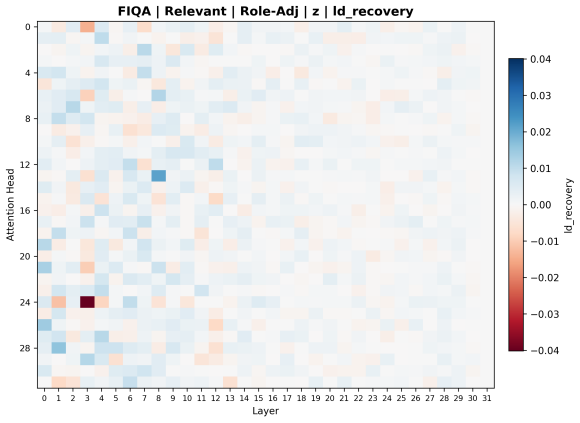
D.3 FIQA Head-Level Results

D.3.1 FIQA: Relevant

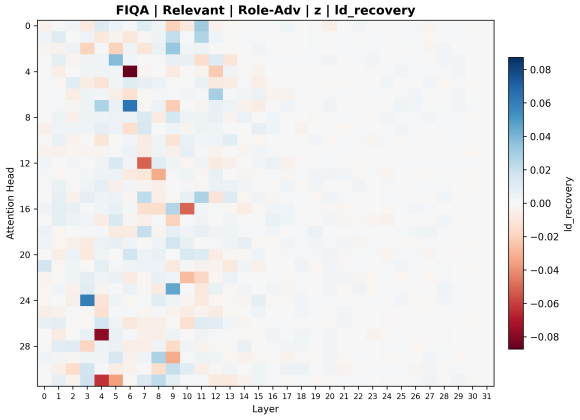
D.3.2 FIQA: Irrelevant



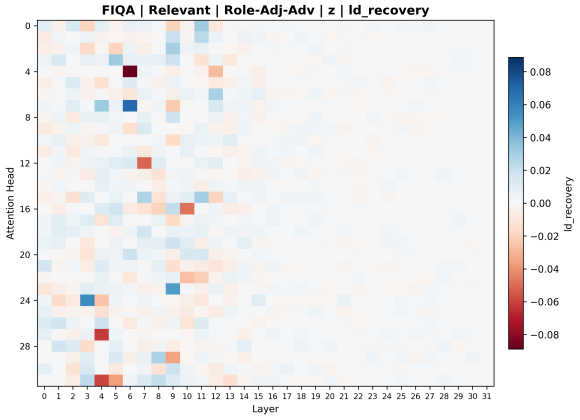
(a) Role-All



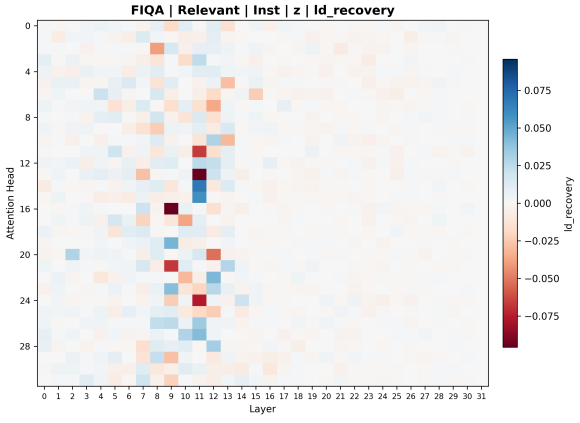
(b) Role-Adj



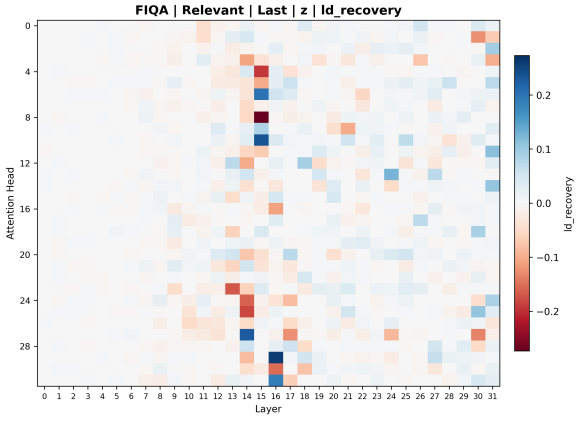
(c) Role-Adv



(d) Role-Adj-Adv

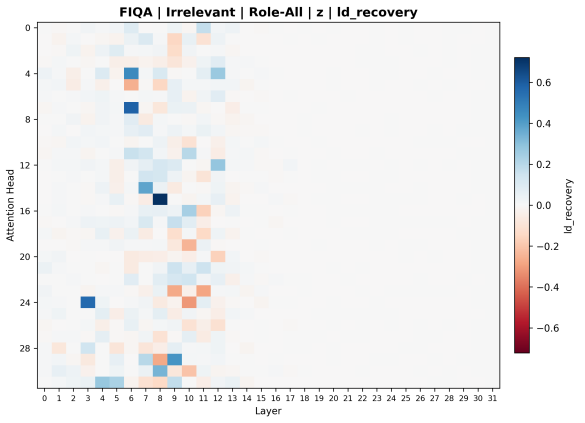


(e) Inst

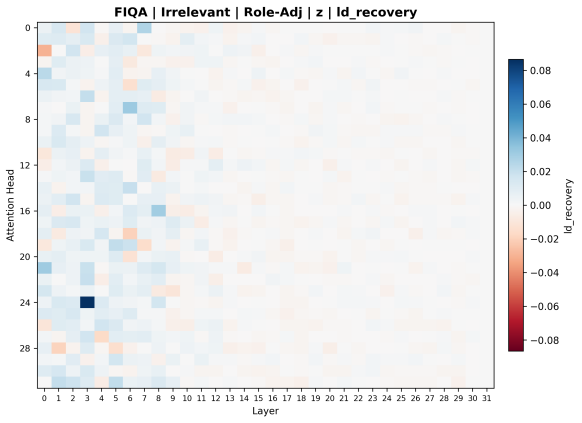


(f) Last

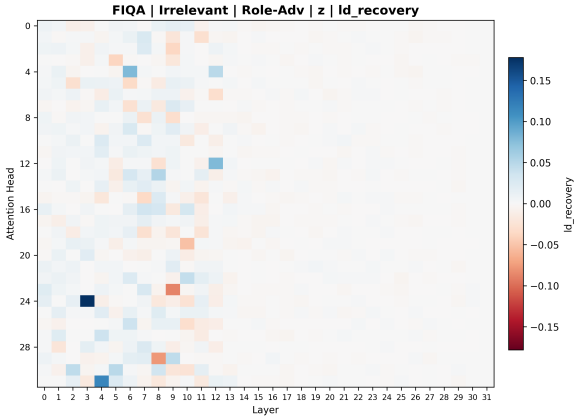
Figure 28: FIQA head-level LD recovery for **Relevant** role prompts.



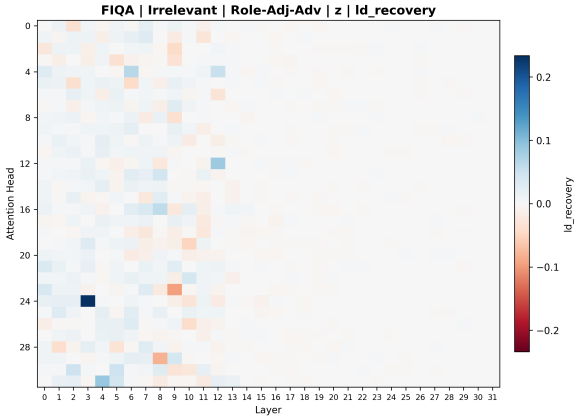
(a) Role-All



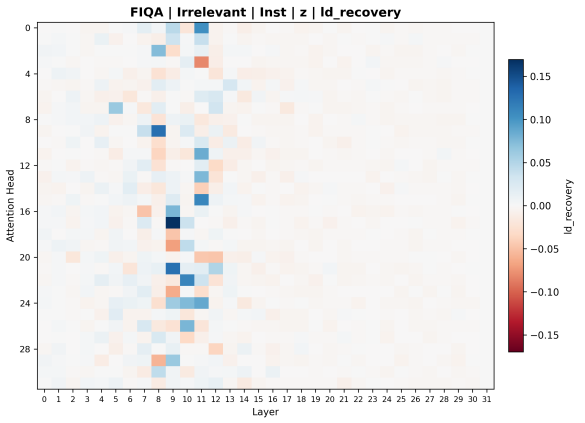
(b) Role-Adj



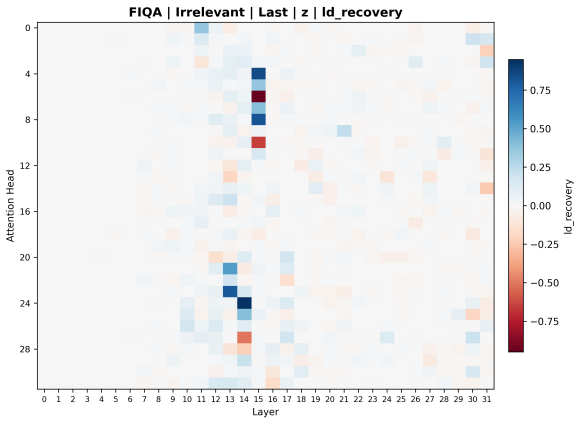
(c) Role-Adv



(d) Role-Adj-Adv



(e) Inst



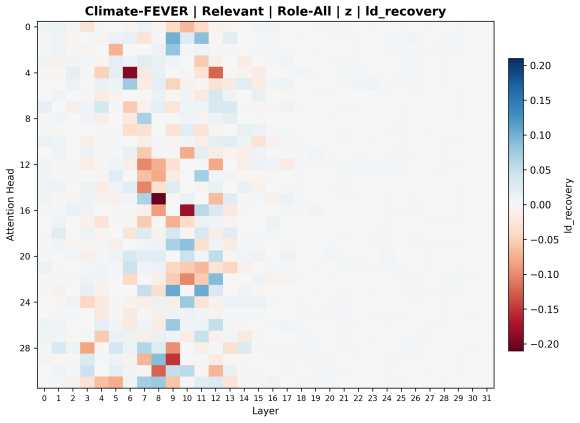
(f) Last

Figure 29: FIQA head-level LD recovery for **Irrelevant** role prompts.

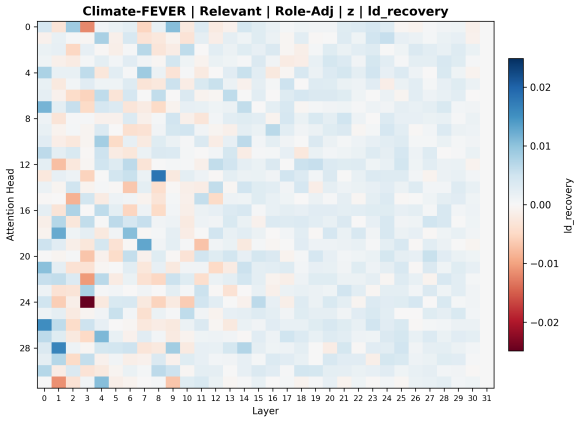
D.4 Climate-Fever Head-Level Results

D.4.1 Climate-Fever: Relevant

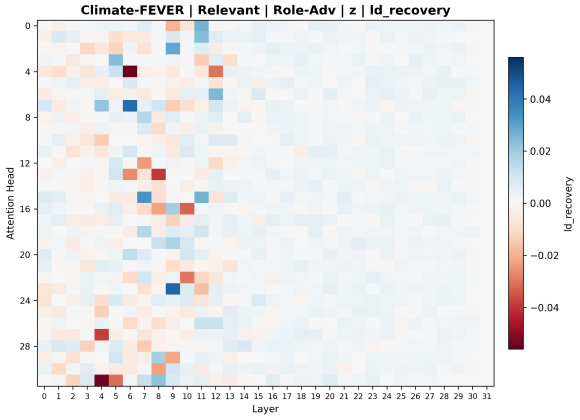
D.4.2 Climate-Fever: Irrelevant



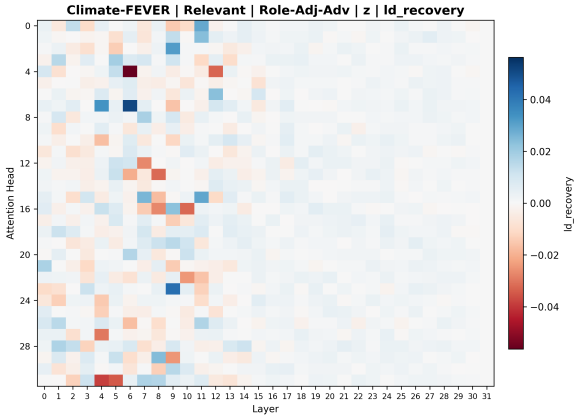
(a) Role-All



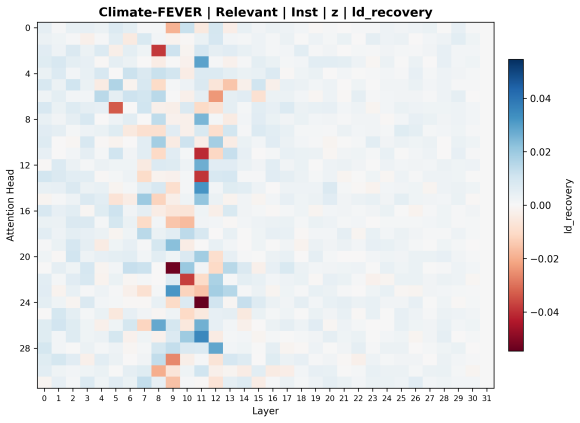
(b) Role-Adj



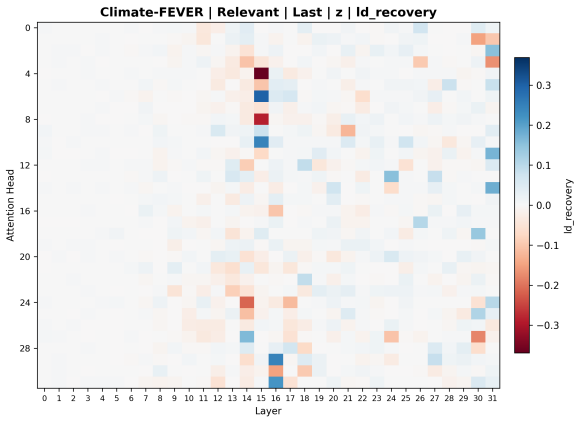
(c) Role-Adv



(d) Role-Adj-Adv

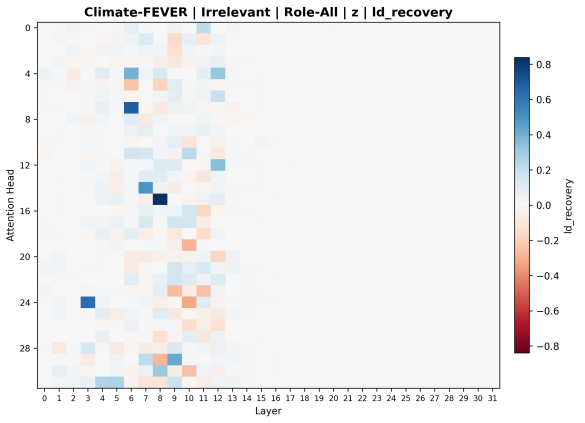


(e) Inst

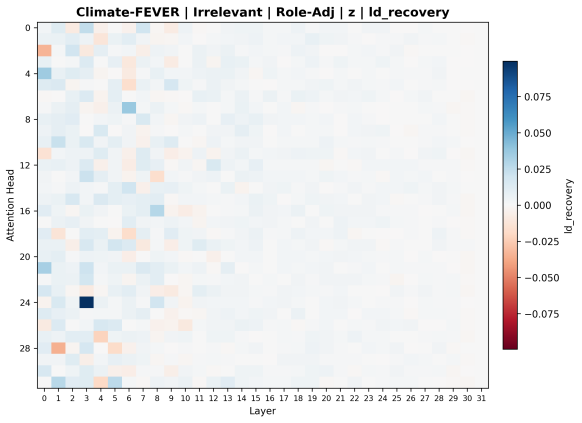


(f) Last

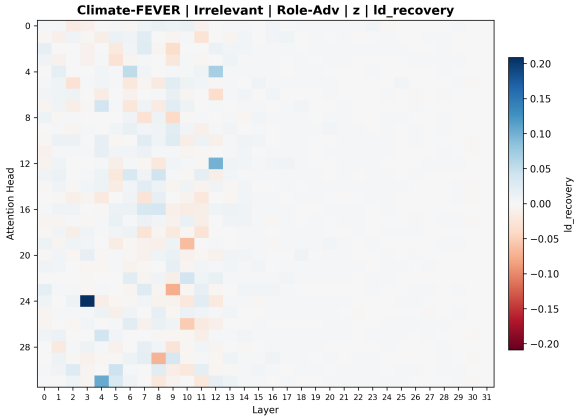
Figure 30: Climate-Fever head-level LD recovery for **Relevant** role prompts.



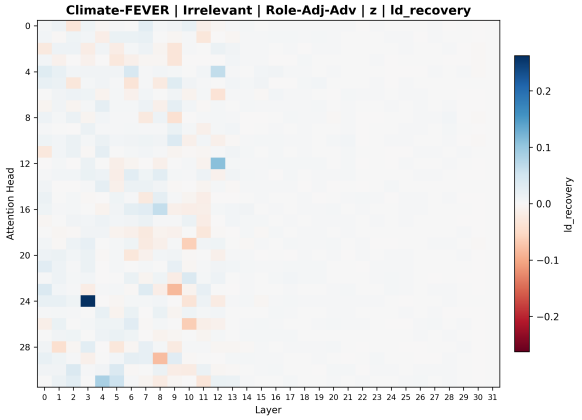
(a) Role-All



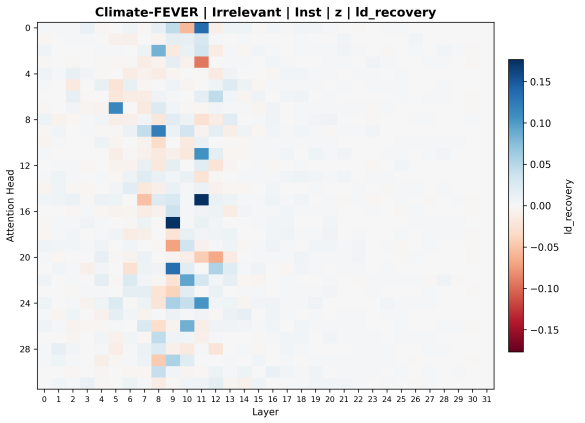
(b) Role-Adj



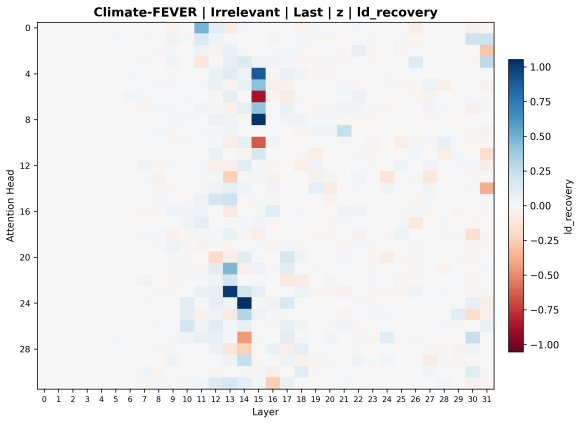
(c) Role-Adv



(d) Role-Adj-Adv



(e) Inst



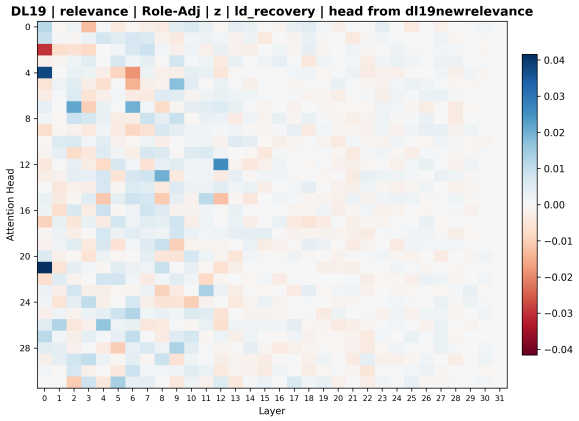
(f) Last

Figure 31: Climate-Fever head-level LD recovery for **Irrelevant** role prompts.

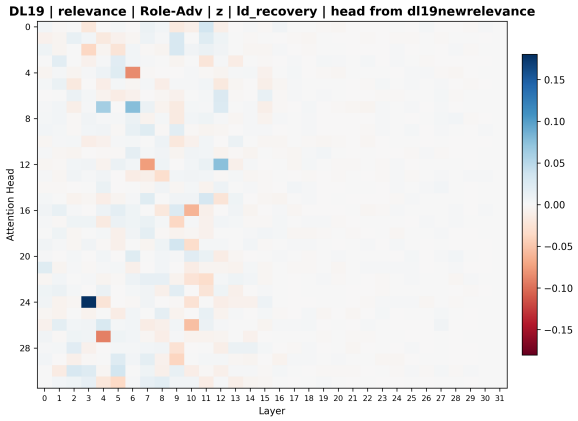
D.5 TREC-DL 2019 Head-Level Results

D.5.1 DL19: Relevant

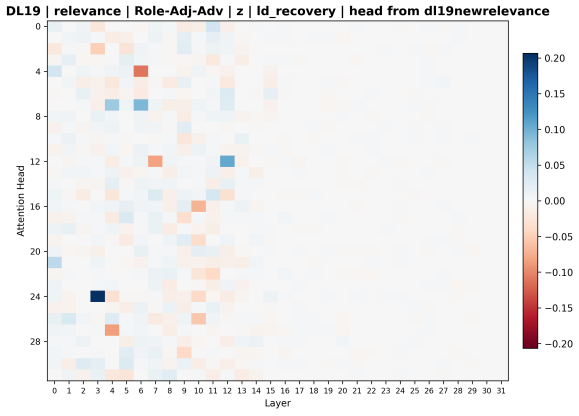
D.5.2 Trec DL 2019: Irrelevant



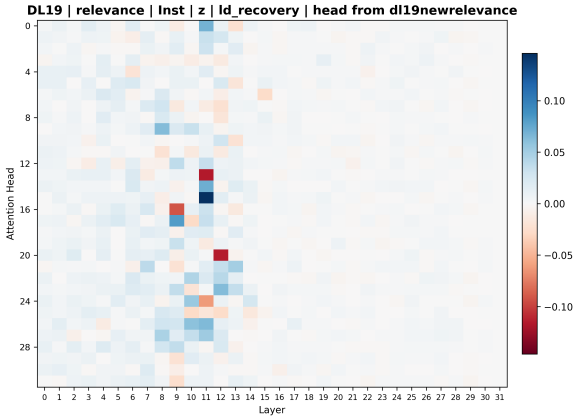
(a) Role-Adj



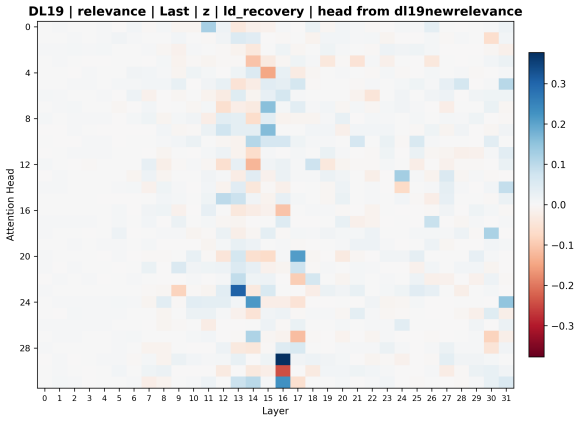
(b) Role-Adv



(c) Role-Adj-Adv

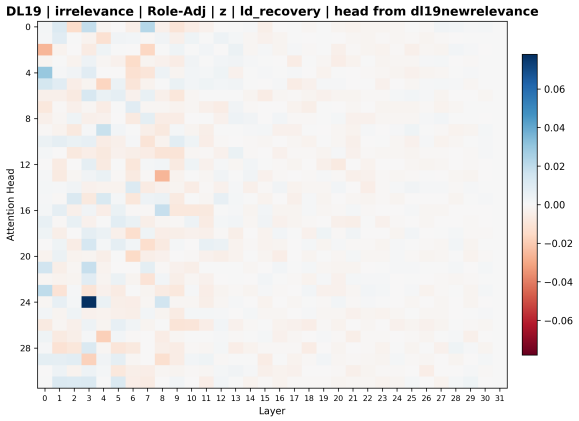


(d) Inst

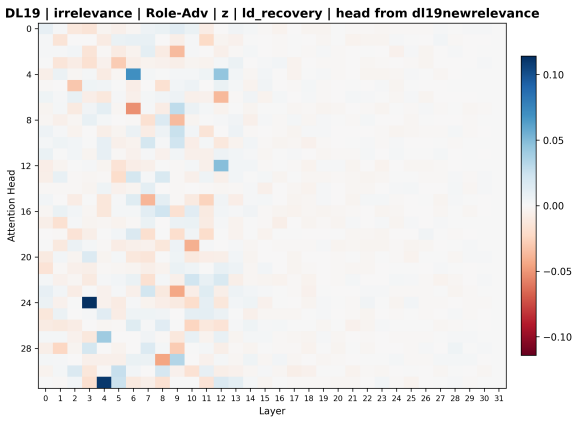


(e) Last

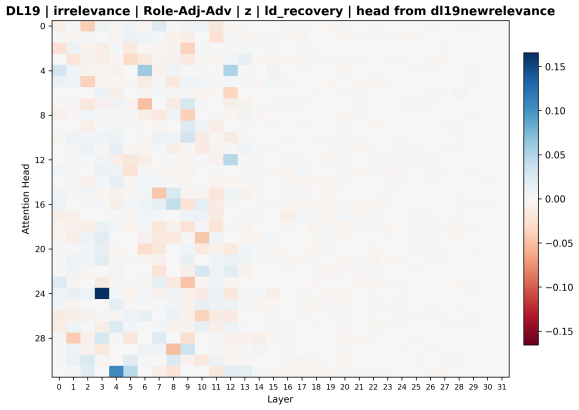
Figure 32: DL19 head-level LD recovery for **Relevant** role prompts.



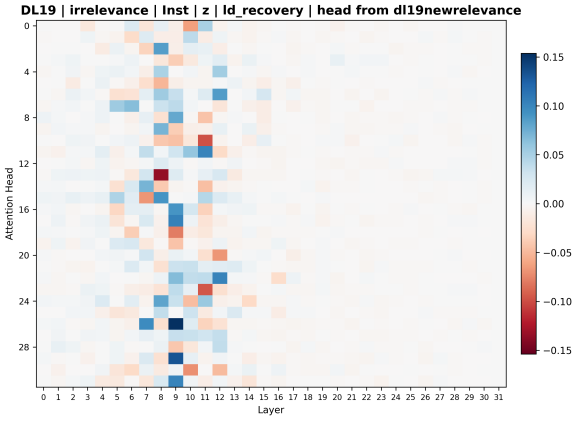
(a) Role-Adj



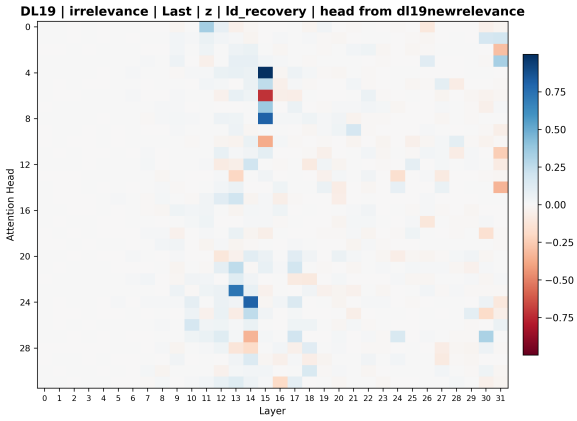
(b) Role-Adv



(c) Role-Adj-Adv



(d) Inst



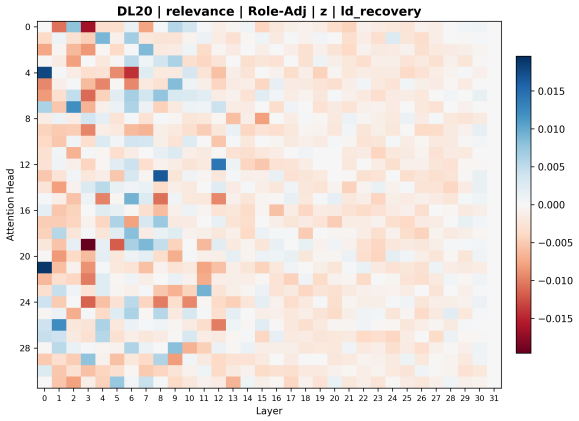
(e) Last

Figure 33: DL19 head-level LD recovery for **Irrelevant** role prompts.

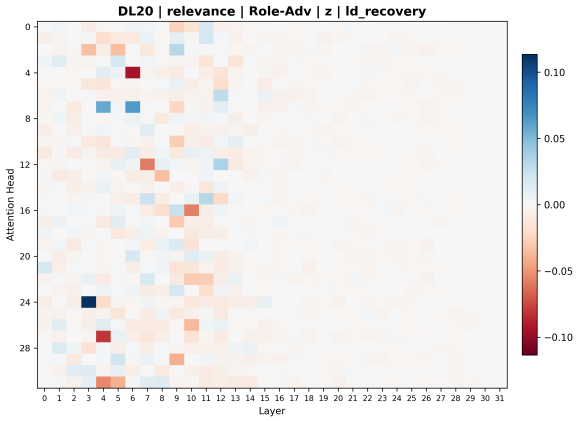
D.6 TREC-DL 2020 Head-Level Results

D.6.1 DL20: Relevant

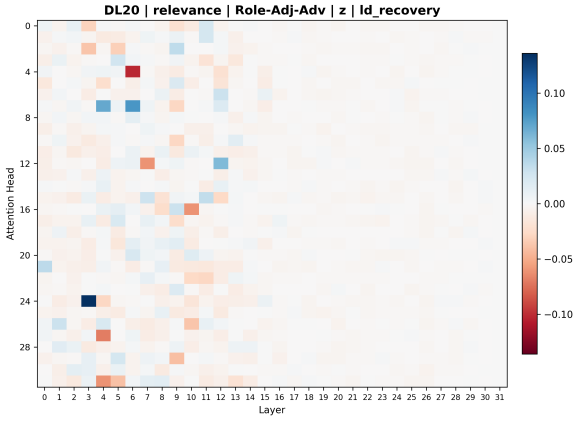
D.6.2 Trec DL 2020: Irrelevant



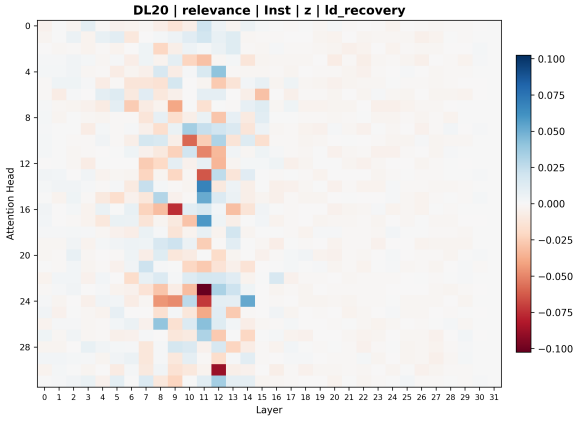
(a) Role-Adj



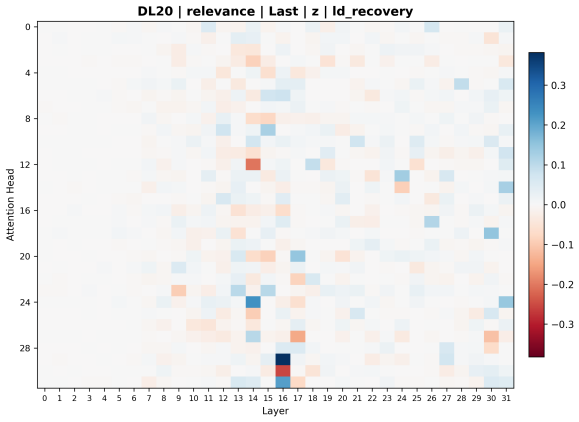
(b) Role-Adv



(c) Role-Adj-Adv

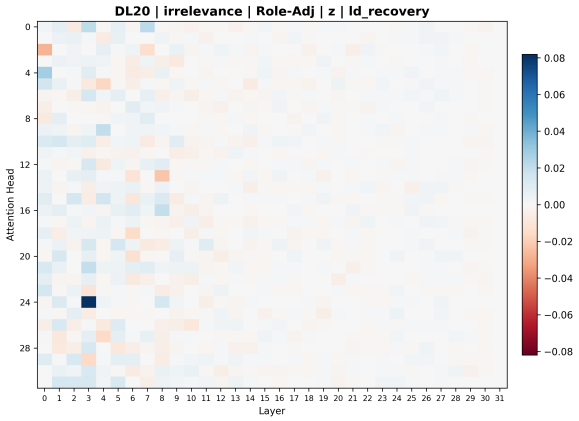


(d) Inst

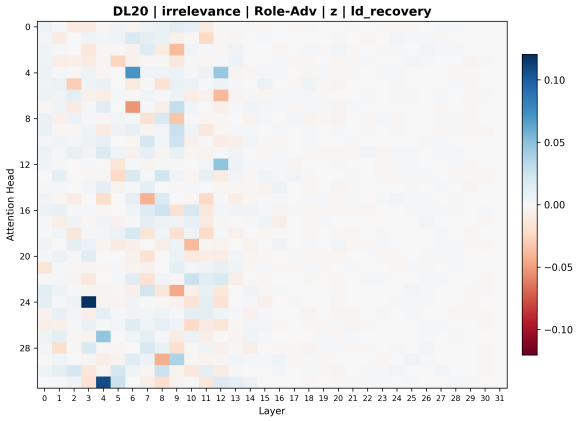


(e) Last

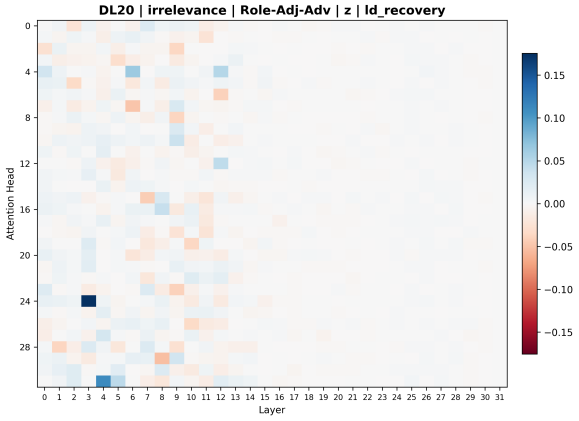
Figure 34: DL20 head-level LD recovery for **Relevant** role prompts.



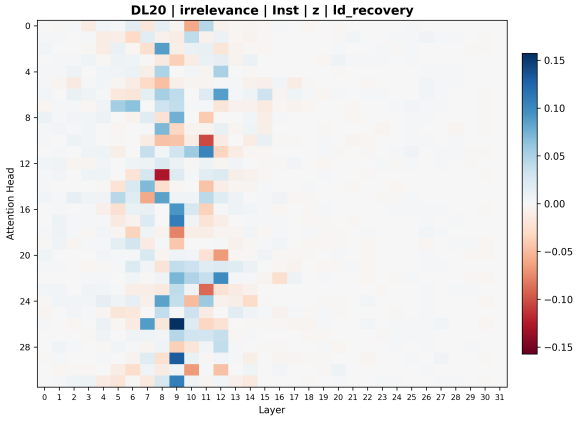
(a) Role-Adj



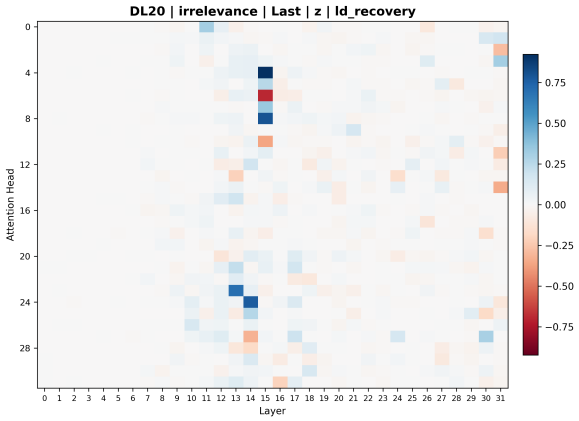
(b) Role-Adv



(c) Role-Adj-Adv



(d) Inst



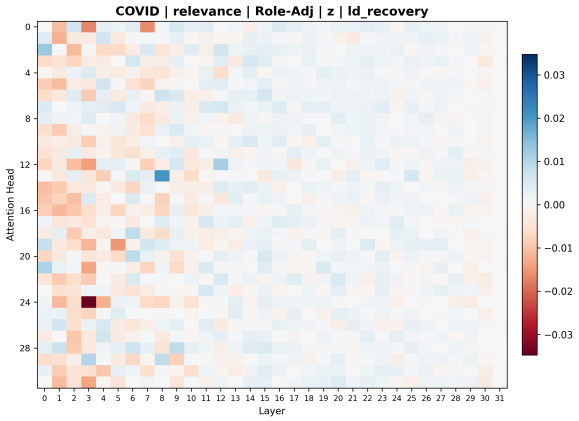
(e) Last

Figure 35: DL20 head-level LD recovery for **Irrelevant** role prompts.

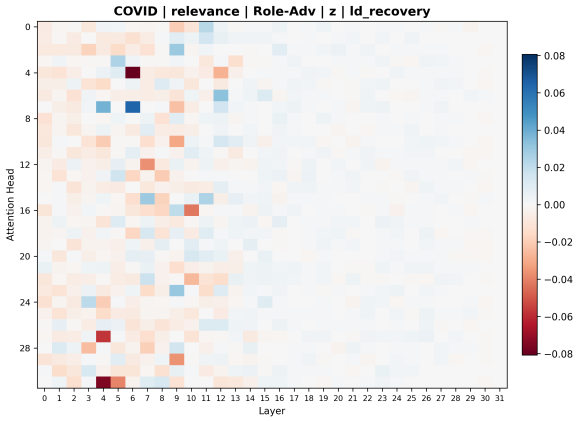
D.7 TREC-COVID Head-Level Results

D.7.1 TREC-COVID: Relevant

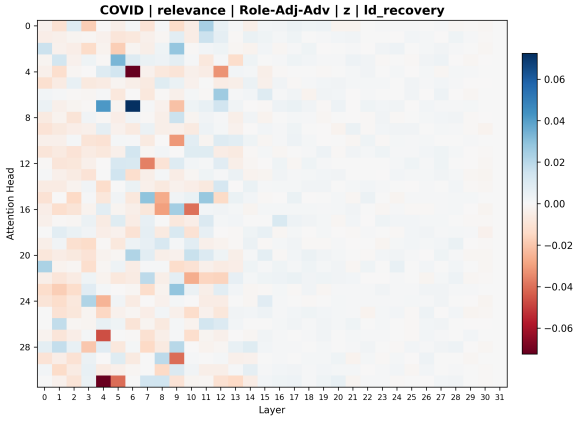
D.7.2 TREC-COVID: Irrelevant



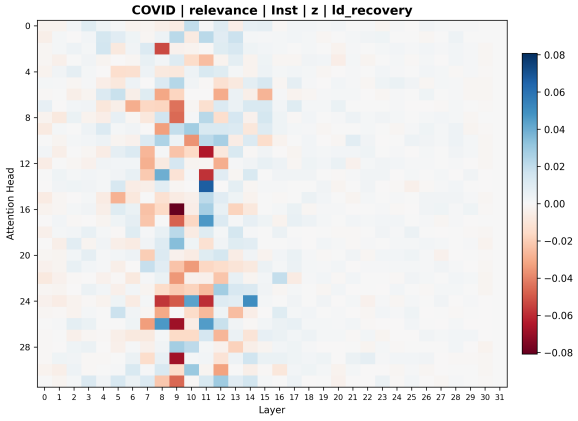
(a) Role-Adj



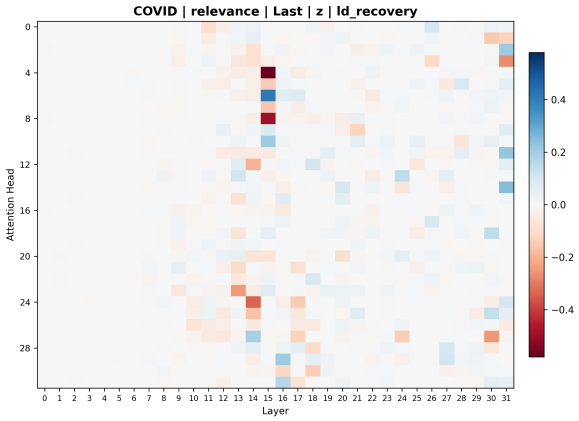
(b) Role-Adv



(c) Role-Adj-Adv

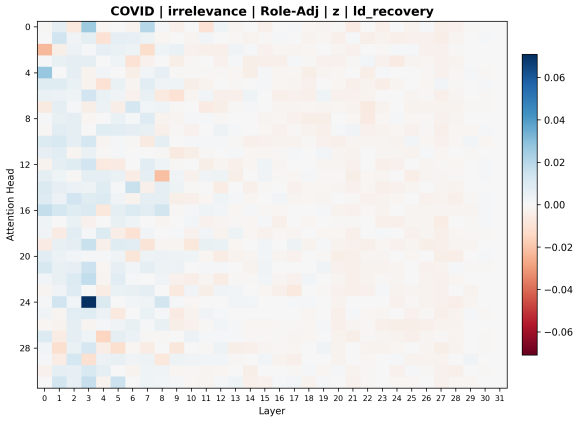


(d) Inst

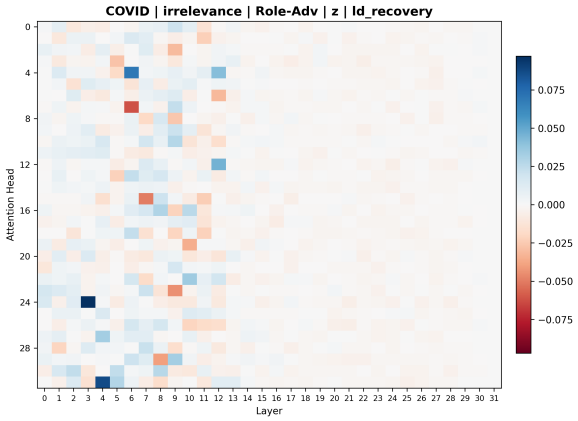


(e) Last

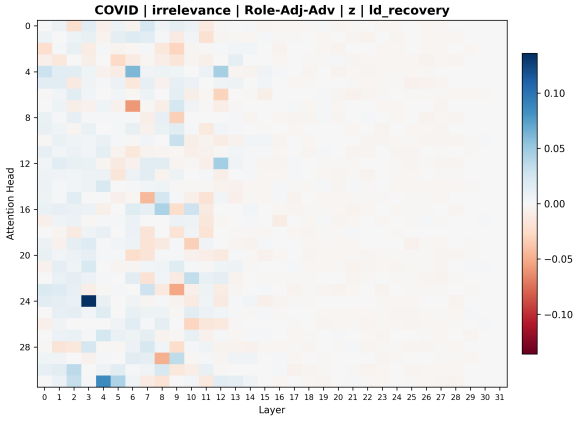
Figure 36: Covid head-level LD recovery for **Relevant** role prompts.



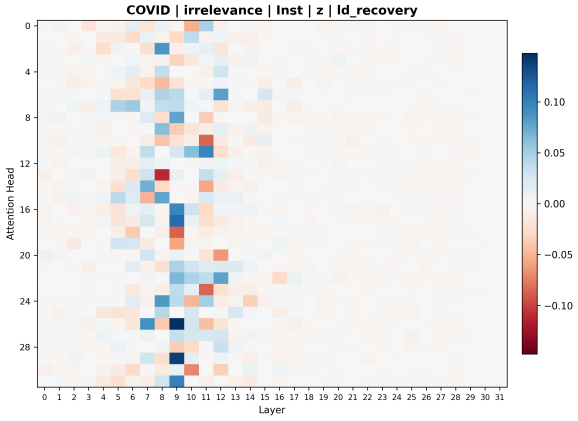
(a) Role-Adj



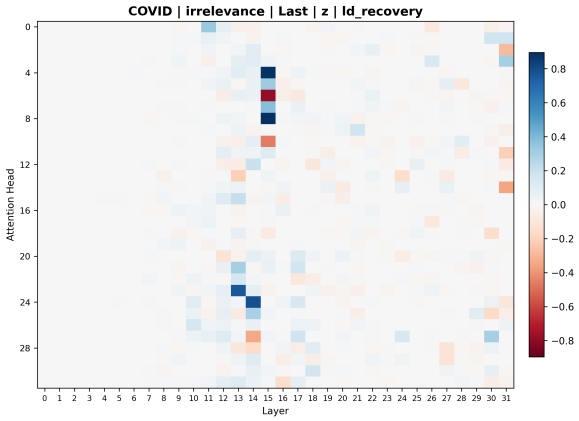
(b) Role-Adv



(c) Role-Adj-Adv



(d) Inst



(e) Last

Figure 37: Covid head-level LD recovery for **Irrelevant** role prompts.

Appendix E Additional Head-Selection and Ablation Tables

E.1 Additional Top-10 Helpful-Head Results

This appendix reports the additional top-10 helpful-head results for Climate-Fever, FIQA, and SciFact. The main text reports the corresponding TREC-DL 2019, TREC-DL 2020, and TREC-COVID results. The figures below provide supplementary evidence that the same target-position differences and relevance–irrelevance asymmetry also appear in the sampled datasets.

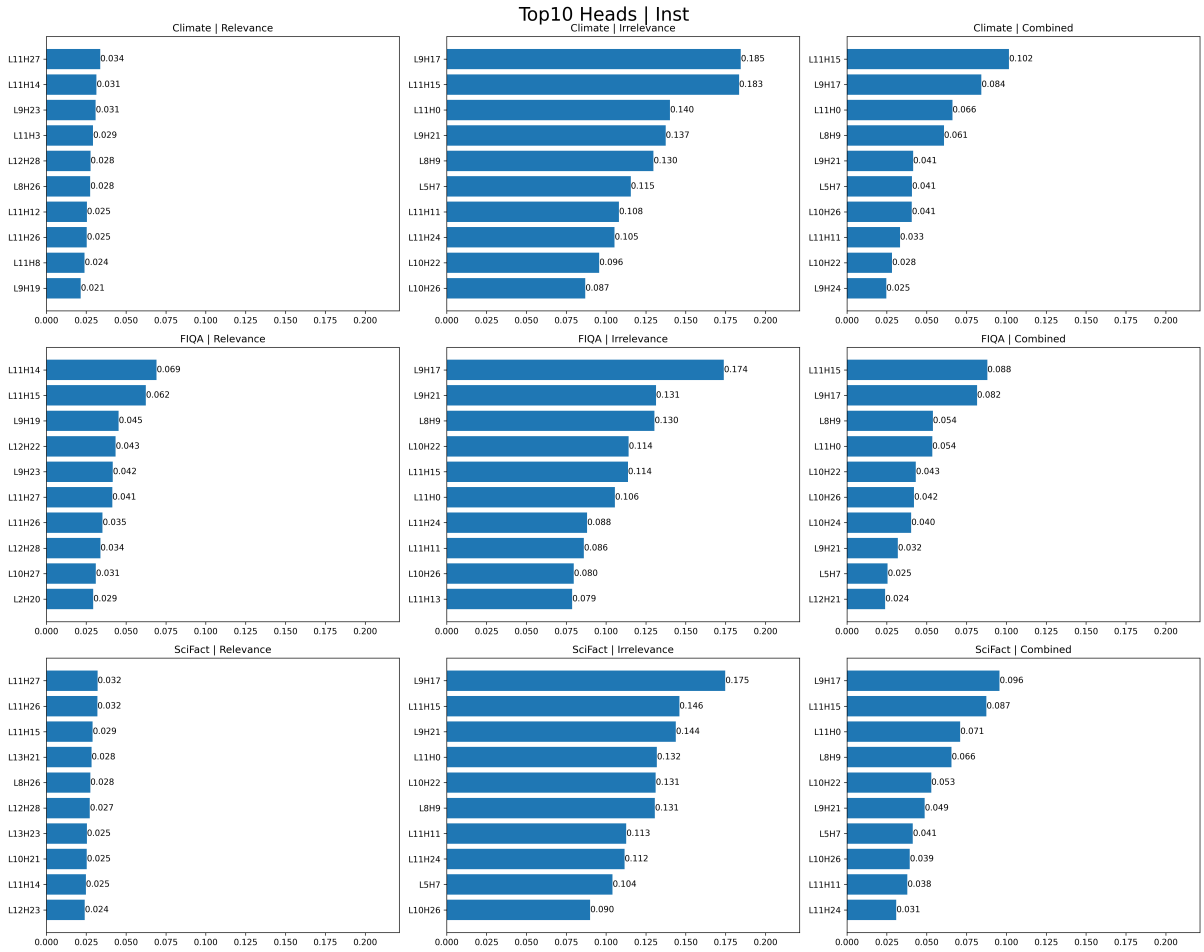


Figure 38: Top-10 helpful heads for instruction-level patching on Climate-Fever, FIQA, and SciFact.

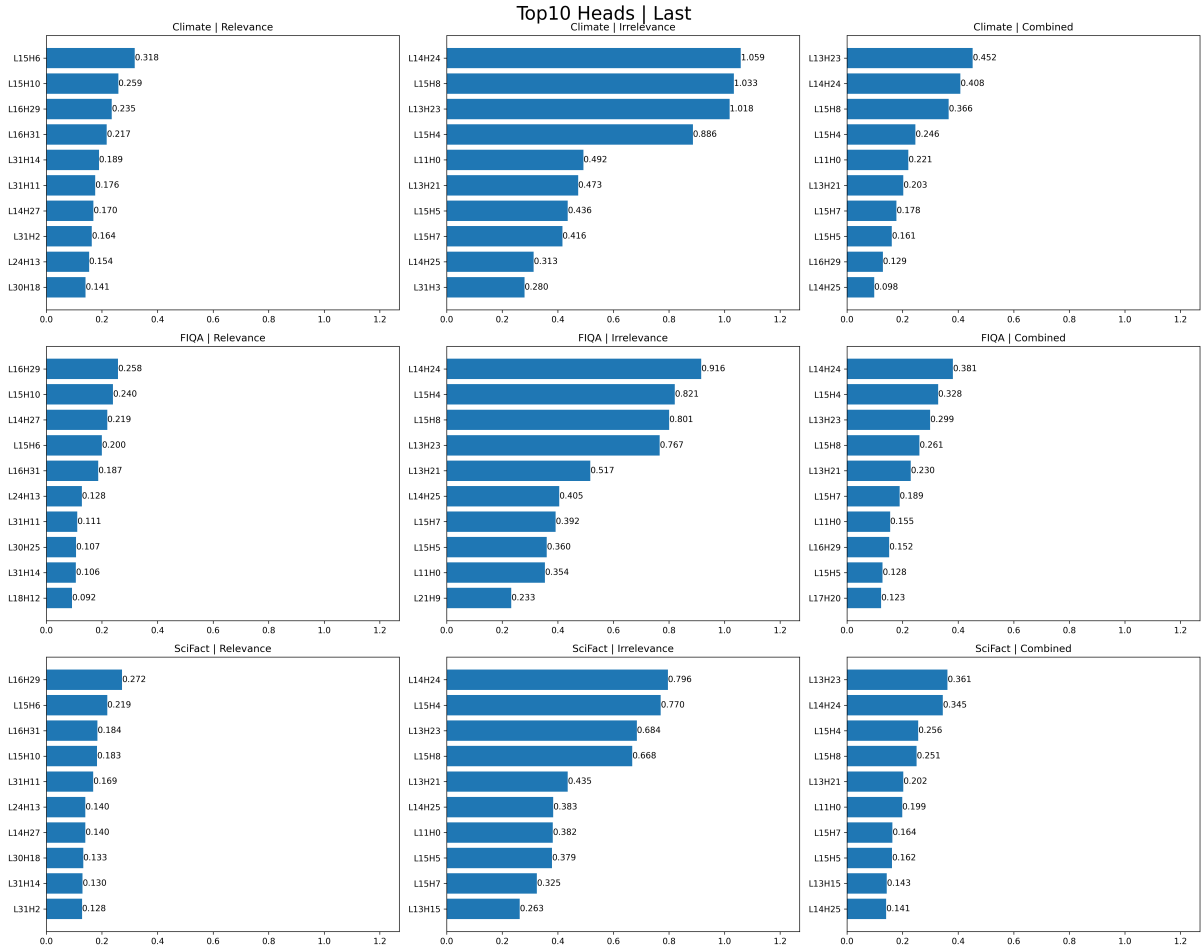


Figure 39: Top-10 helpful heads for last-token patching on Climate-Fever, FIQA, and SciFact.

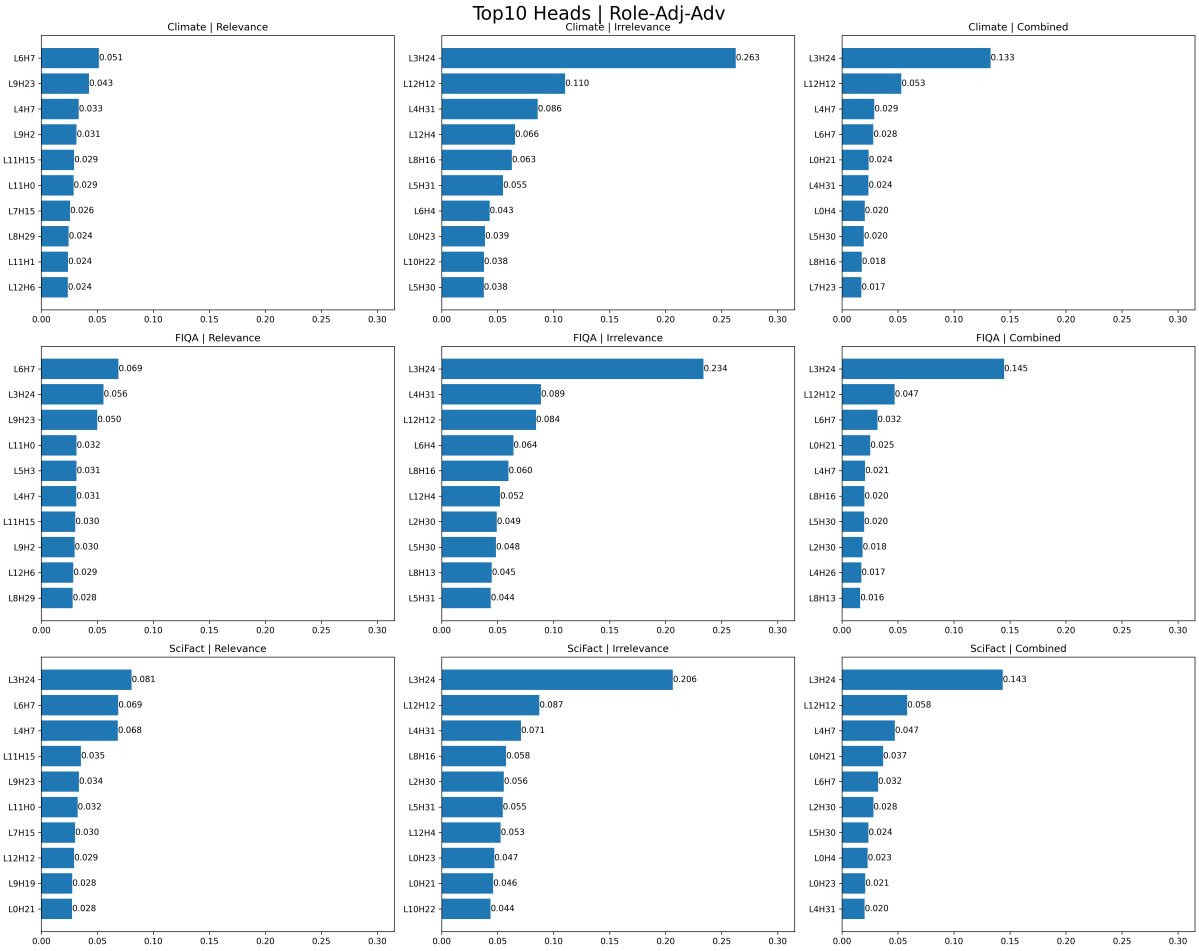


Figure 40: Top-10 helpful heads for joint adjective-adverb role-phrase patching on Climate-Fever, FIQA, and SciFact.

E.2 Mean-ablation results for top-20 candidate heads

The main text reports zero ablation because directly removing the selected head outputs produces the clearest causal contrast with the intact model. For completeness, Figure 41 reports the corresponding mean-ablation results.

In this intervention, the selected head outputs are replaced by their reference mean activations rather than by zero. The overall qualitative structure is similar to the zero-ablation results, but several effects are weaker because mean replacement preserves part of the typical activation magnitude and can therefore reduce the contrast between the clean and intervened runs. The mean-ablation figure is included as a robustness analysis rather than as the primary basis for the mechanistic interpretation.

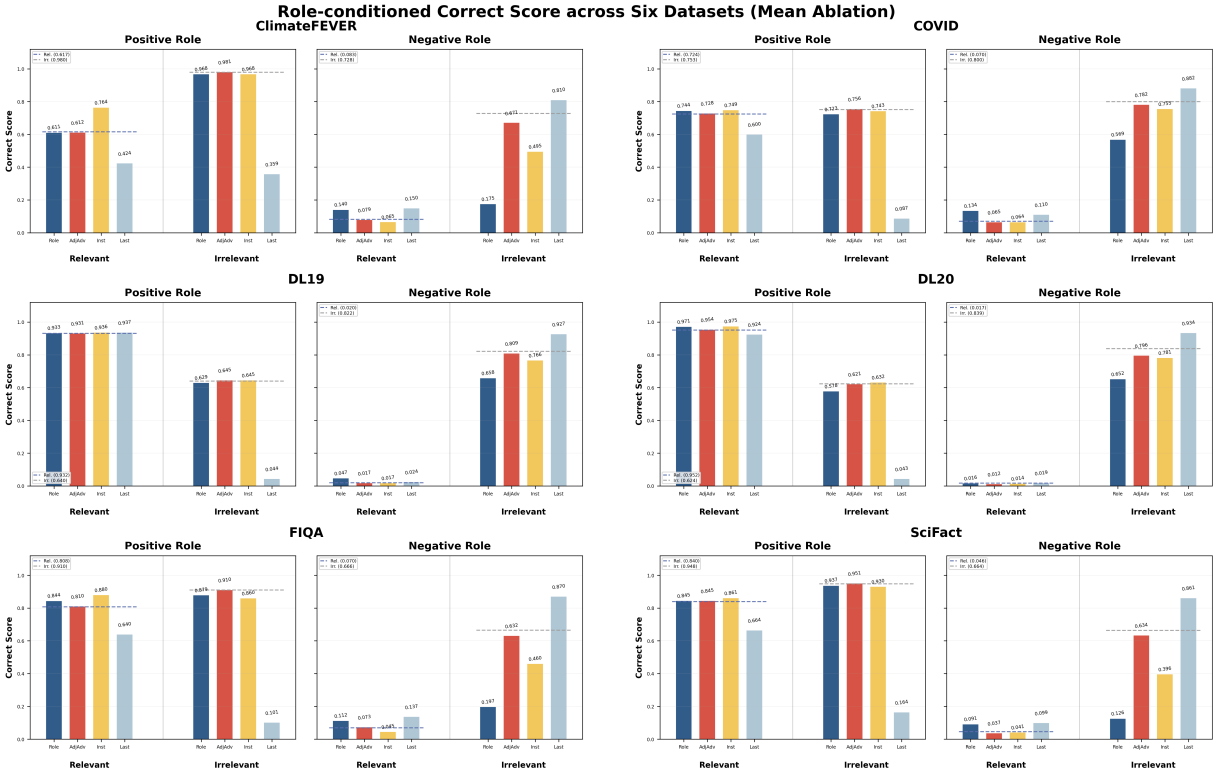


Figure 41: Role-conditioned correct score after mean ablation of the top-20 candidate heads across six datasets. Results are separated by positive and negative role prompts and by relevant and irrelevant documents. Dashed lines indicate the corresponding clean-model scores.

E.3 Complete Top-20 Head Rankings from Activation Patching

This appendix reports the complete Top-20 head rankings used for the head-selection and group-ablation experiments in Section 6.2. Heads are ranked separately for each dataset, relevance condition, and target position. For compactness, each dataset is shown with relevance and irrelevance rankings side by side.

Table 14: Top-20 head rankings for DL19.

DL19 relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L6H7	L3H24	L11H15	L16H29
2	L3H24	L12H12	L9H17	L13H23
3	L12H22	L6H7	L11H14	L14H24
4	L12H6	L4H7	L11H0	L16H31
5	L7H14	L0H21	L12H23	L17H20
6	L9H26	L0H4	L10H26	L15H9
7	L8H15	L11H0	L11H26	L15H7
8	L12H23	L11H15	L8H9	L31H24
9	L0H21	L1H26	L10H24	L24H13
10	L0H4	L5H17	L11H27	L30H18
11	L11H0	L7H15	L13H21	L31H5
12	L11H15	L12H6	L8H27	L14H27
13	L11H26	L5H30	L10H22	L14H10
14	L5H17	L4H26	L9H12	L31H14
15	L7H15	L2H30	L9H28	L14H31
16	L12H12	L12H7	L13H23	L12H15
17	L5H30	L5H29	L12H22	L13H22
18	L4H26	L8H31	L10H27	L13H15
19	L2H30	L9H19	L7H21	L13H31
20	L12H7	L9H19	L11H17	L15H13

DL19 irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L3H24	L3H24	L9H26	L15H4
2	L4H31	L4H31	L9H29	L14H24
3	L6H4	L6H4	L9H17	L15H8
4	L12H12	L12H4	L12H22	L13H23
5	L12H4	L12H12	L11H11	L15H7
6	L5H31	L5H31	L9H31	L11H0
7	L8H16	L8H16	L7H26	L31H3
8	L9H29	L9H29	L8H15	L30H27
9	L10H22	L0H4	L9H16	L13H21
10	L0H4	L9H10	L12H6	L14H25
11	L9H10	L3H28	L8H2	L15H5
12	L3H28	L10H22	L8H24	L14H12
13	L9H7	L9H7	L9H8	L13H15
14	L8H15	L8H15	L8H9	L17H27
15	L8H30	L8H30	L7H14	L31H1
16	L2H30	L2H30	L9H22	L17H21
17	L4H27	L4H27	L6H7	L30H1
18	L0H23	L0H23	L10H11	L21H9
19	L7H0	L7H0	L11H24	L24H27
20	L9H9	L9H9	L5H7	L18H30

Table 15: Top-20 head rankings for DL20.

DL20 relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L3H24	L3H24	L11H14	L16H29
2	L6H7	L6H7	L11H17	L14H24
3	L4H7	L4H7	L14H24	L16H31
4	L12H12	L12H12	L11H15	L30H18
5	L9H2	L0H21	L11H26	L17H20
6	L11H15	L9H2	L12H4	L31H24
7	L12H6	L11H15	L8H26	L24H13
8	L9H16	L9H16	L12H31	L31H14
9	L7H15	L12H6	L10H9	L15H9
10	L5H29	L1H26	L12H23	L14H27
11	L9H23	L7H15	L11H27	L26H17
12	L11H1	L5H3	L12H10	L15H23
13	L11H0	L11H0	L8H15	L18H12
14	L5H3	L5H29	L10H24	L28H5
15	L0H21	L11H1	L12H13	L15H6
16	L6H20	L6H20	L11H9	L13H23
17	L7H22	L5H17	L7H14	L16H6
18	L9H19	L4H26	L13H23	L31H11
19	L1H28	L9H5	L7H21	L12H9
20	L8H31	L9H23	L9H19	L21H10

DL20 irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L3H24	L3H24	L9H26	L15H4
2	L4H31	L4H31	L9H29	L15H8
3	L6H4	L6H4	L9H17	L14H24
4	L12H12	L12H4	L9H31	L13H23
5	L4H27	L5H31	L11H11	L15H7
6	L12H4	L12H12	L12H22	L11H0
7	L9H29	L8H16	L9H16	L31H3
8	L9H7	L9H10	L7H26	L30H27
9	L9H9	L9H29	L8H2	L14H25
10	L8H30	L0H4	L12H6	L15H5
11	L5H31	L4H27	L8H24	L13H21
12	L10H22	L8H15	L8H15	L17H27
13	L9H10	L9H7	L9H8	L13H15
14	L8H16	L9H9	L9H22	L14H12
15	L7H23	L7H23	L8H9	L31H1
16	L8H13	L7H0	L7H14	L17H21
17	L5H30	L8H30	L6H7	L21H9
18	L2H30	L3H28	L11H24	L24H27
19	L7H10	L7H28	L10H11	L10H26
20	L10H16	L3H19	L5H7	L30H1

Table 16: Top-20 head rankings for TREC-COVID.

TREC-COVID relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L6H7	L6H7	L11H14	L15H6
2	L4H7	L4H7	L14H24	L31H14
3	L5H3	L5H3	L8H26	L15H10
4	L7H15	L7H15	L11H17	L31H11
5	L9H23	L9H23	L11H26	L16H29
6	L9H2	L9H2	L10H24	L14H27
7	L11H15	L11H15	L8H13	L31H2
8	L12H6	L12H6	L9H19	L16H31
9	L9H16	L9H16	L10H30	L24H13
10	L11H0	L11H0	L12H31	L30H18
11	L0H21	L3H24	L12H23	L30H25
12	L3H24	L6H20	L10H9	L31H24
13	L6H20	L9H28	L12H10	L27H28
14	L9H28	L7H22	L11H16	L27H29
15	L7H22	L8H29	L8H10	L18H12
16	L8H29	L1H26	L9H5	L26H0
17	L1H26	L1H28	L11H8	L13H13
18	L1H28	L10H20	L9H1	L28H5
19	L10H20	L9H8	L9H18	L26H17
20	L9H8	L11H31	L10H22	L20H14

TREC-COVID irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L3H24	L3H24	L9H26	L15H4
2	L4H31	L4H31	L9H29	L15H8
3	L6H4	L6H4	L9H17	L14H24
4	L12H12	L12H12	L9H31	L13H23
5	L12H4	L12H4	L11H11	L15H7
6	L8H16	L8H16	L9H16	L15H5
7	L5H31	L5H31	L7H26	L11H0
8	L2H30	L2H30	L8H2	L30H27
9	L9H29	L9H29	L8H24	L14H25
10	L10H22	L10H22	L12H6	L31H3
11	L2H31	L2H31	L12H22	L13H21
12	L9H10	L9H10	L9H8	L13H15
13	L0H4	L0H4	L8H15	L14H12
14	L10H16	L10H16	L9H22	L21H9
15	L7H23	L7H23	L8H9	L17H27
16	L8H15	L8H15	L10H11	L31H1
17	L4H27	L4H27	L6H7	L30H1
18	L9H7	L9H7	L11H24	L17H21
19	L7H0	L7H0	L11H0	L18H30
20	L3H28	L3H28	L9H29	L10H26

Table 17: Top-20 head rankings for SciFact.

SciFact relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L12H22	L3H24	L11H27	L16H29
2	L3H24	L6H7	L11H26	L15H6
3	L9H1	L4H7	L11H15	L15H10
4	L9H26	L11H15	L13H21	L16H31
5	L8H31	L9H23	L12H28	L31H11
6	L11H13	L11H0	L8H26	L14H27
7	L12H6	L7H15	L11H14	L24H13
8	L11H1	L12H12	L10H21	L30H18
9	L6H7	L9H19	L12H23	L31H14
10	L8H29	L0H21	L13H23	L31H2
11	L7H31	L9H2	L9H23	L31H24
12	L9H2	L8H31	L11H12	L15H9
13	L7H15	L5H3	L9H17	L26H17
14	L4H7	L9H16	L7H31	L18H12
15	L9H19	L12H6	L10H18	L27H29
16	L7H28	L1H28	L10H24	L27H28
17	L11H23	L11H1	L9H28	L30H25
18	L7H8	L6H20	L7H18	L31H5
19	L7H23	L5H17	L14H24	L27H13
20	L3H30	L5H29	L9H26	L17H20

SciFact irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L8H15	L3H24	L9H17	L14H24
2	L6H7	L12H12	L9H21	L15H4
3	L3H24	L4H31	L11H15	L13H23
4	L7H14	L8H16	L10H22	L15H8
5	L9H29	L2H30	L11H0	L13H21
6	L6H4	L5H31	L8H9	L15H5
7	L12H12	L12H4	L11H11	L14H25
8	L8H30	L0H23	L11H24	L11H0
9	L12H4	L0H21	L5H7	L15H7
10	L5H31	L10H22	L10H26	L13H15
11	L10H11	L8H13	L8H2	L21H9
12	L4H31	L7H16	L9H24	L30H27
13	L10H16	L6H4	L9H29	L30H1
14	L7H29	L5H30	L12H21	L14H29
15	L12H6	L6H26	L9H0	L31H1
16	L9H17	L2H31	L9H16	L31H3
17	L9H22	L0H4	L10H24	L30H30
18	L9H31	L4H14	L7H26	L17H27
19	L11H21	L9H7	L8H30	L10H26
20	L11H0	L9H5	L12H6	L12H15

Table 18: Top-20 head rankings for FIQA.

FIQA relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L12H22	L8H13	L11H14	L16H29
2	L12H6	L1H28	L11H15	L15H10
3	L6H7	L0H26	L9H19	L14H27
4	L8H31	L0H21	L12H22	L15H6
5	L11H23	L8H16	L9H23	L16H31
6	L9H1	L3H29	L11H27	L24H13
7	L9H26	L4H27	L11H26	L30H25
8	L11H13	L0H19	L12H28	L31H11
9	L8H29	L2H7	L10H27	L31H14
10	L9H2	L7H2	L2H20	L18H12
11	L11H1	L4H1	L12H23	L30H18
12	L10H24	L12H12	L12H10	L31H2
13	L11H16	L6H24	L13H21	L31H24
14	L10H19	L4H23	L8H26	L27H13
15	L9H23	L7H4	L11H12	L15H9
16	L6H5	L1H8	L9H26	L13H12
17	L10H20	L1H18	L11H3	L17H20
18	L7H31	L6H12	L8H29	L31H5
19	L9H19	L6H31	L12H12	L26H17
20	L7H28	L7H17	L12H13	L25H10

FIQA irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L8H15	L3H24	L9H17	L14H24
2	L6H7	L6H7	L9H21	L15H4
3	L3H24	L0H21	L8H9	L15H8
4	L6H4	L8H16	L11H15	L13H23
5	L9H29	L7H0	L10H22	L13H21
6	L7H14	L0H4	L11H0	L14H25
7	L8H30	L5H31	L11H24	L15H7
8	L12H12	L5H19	L11H11	L15H5
9	L4H31	L3H6	L10H26	L11H0
10	L12H4	L1H31	L10H24	L21H9
11	L10H16	L6H14	L11H13	L13H15
12	L5H31	L3H22	L9H16	L30H27
13	L10H11	L3H21	L8H2	L14H29
14	L7H29	L3H13	L5H7	L30H1
15	L9H31	L6H19	L9H24	L31H3
16	L11H0	L2H31	L9H29	L17H27
17	L9H17	L7H8	L12H21	L13H22
18	L6H11	L2H2	L9H0	L17H20
19	L9H21	L3H27	L8H30	L10H26
20	L11H21	L3H0	L7H9	L31H1

Table 19: Top-20 head rankings for Climate-FEVER.

Climate-FEVER relevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L9H23	L8H13	L11H27	L15H6
2	L11H23	L1H28	L11H14	L15H10
3	L9H1	L0H26	L9H23	L16H29
4	L12H22	L7H19	L11H3	L16H31
5	L8H29	L1H18	L8H26	L31H14
6	L11H1	L0H7	L12H28	L31H11
7	L10H19	L4H27	L11H26	L14H27
8	L9H2	L9H0	L11H12	L31H2
9	L9H26	L4H31	L11H8	L24H13
10	L10H24	L6H18	L7H15	L30H18
11	L8H31	L0H21	L10H21	L30H25
12	L6H5	L2H0	L9H19	L26H17
13	L7H31	L7H4	L12H23	L31H24
14	L11H13	L9H28	L10H27	L18H12
15	L7H8	L4H10	L11H15	L18H22
16	L9H19	L4H1	L8H10	L27H29
17	L7H15	L3H23	L11H20	L27H13
18	L7H28	L0H4	L12H22	L31H5
19	L11H16	L2H16	L7H18	L20H14
20	L10H30	L14H28	L12H10	L28H5

Climate-FEVER irrelevance

Rank	Role-All	Role-AdjAdv	Inst-All	Last
1	L8H15	L3H24	L9H17	L14H24
2	L6H7	L6H7	L11H15	L15H8
3	L3H24	L0H4	L11H0	L13H23
4	L7H14	L0H21	L9H21	L15H4
5	L9H29	L8H16	L8H9	L11H0
6	L6H4	L1H31	L5H7	L13H21
7	L12H12	L5H31	L11H11	L15H5
8	L12H4	L3H0	L11H24	L15H7
9	L8H30	L1H10	L10H22	L14H25
10	L5H31	L3H13	L10H26	L31H3
11	L4H31	L3H21	L8H2	L30H27
12	L10H11	L7H0	L9H24	L14H29
13	L11H0	L6H14	L12H21	L21H9
14	L7H29	L8H24	L9H29	L13H15
15	L12H6	L2H2	L8H27	L31H1
16	L9H31	L5H19	L9H0	L30H1
17	L6H11	L3H22	L9H16	L13H22
18	L9H17	L0H23	L12H6	L10H26
19	L10H17	L3H28	L7H9	L14H23
20	L9H22	L9H5	L8H30	L13H31

References

[1] Chen, C., Merullo, J., and Eickhoff, C. (2024). Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1401–1410.

[2] Craswell, N., Mitra, B., Yilmaz, E., and Campos, D. (2021). Overview of the TREC 2020 deep learning track. *arXiv preprint arXiv:2102.07662*.

[3] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., and Voorhees, E. M. (2020). Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

[4] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. (2020). CLIMATE-FEVER: A dataset for verification of real-world climate claims. arXiv preprint arXiv:2012.00614.

[5] Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

[6] Heimersheim, S. and Nanda, N. (2024). How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

[7] Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., and Dong, X. (2024). Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 4099–4113.

[8] Liu, Q., Mao, J., and Wen, J.-R. (2025). How do large language models understand relevance? a mechanistic interpretability perspective. *arXiv preprint arXiv:2504.07898*.

[9] Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., and Balahur, A. (2018). WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

[10] Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*.

[11] Nanda, N. and Bloom, J. (2022). Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

[12] Nogueira, R., Jiang, Z., and Lin, J. (2020). Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

[13] Polyakov, G., Chen, C., and Eickhoff, C. (2025). Towards best practices of axiomatic activation patching in information retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2972–2976.

[14] Pradeep, R., Sharifymoghaddam, S., and Lin, J. (2023). RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

[15] Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., and Bendersky, M. (2024). Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

[16] Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L. L., and Hersh, W. R. (2021). Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *arXiv preprint arXiv:2104.09632*.

[17] Sun, S., Zhuang, S., Wang, S., and Zuccon, G. (2025). An investigation of prompt variations for zero-shot LLM-based rankers. In *Proceedings of the 47th European Conference on Information Retrieval (ECIR 2025)*.

[18] Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[19] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

[20] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[21] Wang, Y., Qi, J., Chen, C., Eustratiadis, P., and Verberne, S. (2025). How role-play shapes relevance judgment in zero-shot LLM rankers. *arXiv preprint arXiv:2510.17535*.

[22] Zhang, F. and Nanda, N. (2023). Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

[23] Zhuang, H., Qin, Z., Hui, K., Wu, J., Yan, L., Wang, X., and Bendersky, M. (2024a). Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[24] Zhuang, S., Zhuang, H., Koopman, B., and Zuccon, G. (2024b). A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.