The Narrative Cliff: Hierarchical Contrast Sets Reveal Non-Monotonic Robustness in QA Models

Garrett Moran
Department of Computer Science, University of Texas at Austin
garrettmoran@utexas.edu

Abstract

Pre-trained Question Answering (QA) models frequently fail to transfer in-distribution competence to out-of-distribution settings. We investigate the model degradation curve across a curated, three-tiered contrast set, introducing increasing levels of qualitative shift in question and context: syntactic variation (L1), elaborate rephrasing (L2), and narrative domain shift (L3). Using Jaccard distance to quantify drift, our evaluation of ELECTRA-small reveals a non-monotonic robustness curve: performance drops on L1 (0.41 mean drift), peaks on L2 (0.67 mean drift), and falls significantly on L3 (0.73 mean drift). Crucially, a minimal lexical drift increase (Δ 0.06) between L2 and L3 corresponds with a sharp decline in performance, indicating that narrative framing impacts robustness far more than lexical overlap suggests. Regarding mitigation, we find full-parameter fine-tuning induces catastrophic forgetting. However, freezing the encoder preserves performance on L1 and L2, but collapses on L3, falling significantly below the baseline. This exposes a sharp "narrative cliff": the frozen model adapts to elaborate paraphrasing but collapses under narrative shift, demonstrating that lexical adaptability and domain generalization are distinct capabilities, and that optimizing for one can inadvertently degrade the other.

1 Introduction

The advent of Transformer-based architectures has revolutionized Natural Language Processing, with models like ELECTRA (Clark et al., 2020) surpassing human performance on reading comprehension benchmarks such as SQuAD (Rajpurkar et al., 2016). However, this “superhuman” performance is often fragile. Diagnostic research has consistently demonstrated that these models frequently rely on shallow heuristics, such as lexical overlap or hypothesis-only bias, rather than robust linguistic reasoning (Poliak et al., 2018; McCoy et al., 2019). When these statistical artifacts are removed or perturbed, model performance degrades significantly, exposing a critical gap between in-distribution benchmark scores and real-world reliability (Gan and Ng, 2019).

To diagnose this brittleness, researchers have proposed expert-curated stress tests such as contrast sets (Gardner et al., 2020) and paraphrase robustness benchmarks (Gan & Ng, 2019). While Gardner et al. originally emphasized perturbations that flip the label, the broader methodology involves creating minimally perturbed examples to probe local decision boundaries, often by testing for invariance under surface-form changes. Additionally, existing work typically frames distribution shift as a binary phenomenon: an example is either 'in-distribution' (Independent and Identically Distributed, or IID) or 'out-of-distribution' (OOD) (Koh et al., 2021). This monolithic treatment of OOD data often conflates benign elaborations with structural domain shifts, obscuring the specific point at which a model's representation breaks. Consequently, standard evaluations may fail to distinguish between robustness to lexical complexity versus robustness to narrative reframing, which is a distinction critical for diagnosing why mitigation strategies fail.

In this work, we systematically map this degradation by constructing a hierarchical contrast set based on SQuAD. We design three qualitative tiers of perturbation to isolate specific failure modes: L1 (Syntactic Variation), L2 (Elaborate

Rephrasing), and L3 (Narrative Shift). Unlike binary evaluations, we quantify the magnitude of these shifts using Jaccard distance, allowing us to decouple lexical drift from semantic difficulty and analyze robustness as a function of measurable distance.

Our experiments reveal two primary anomalies in model behavior. First, we observe a non-monotonic robustness curve: contrary to the assumption that increased drift linearly increases difficulty, the baseline model performs best on elaborate rephrasing (L2), surpassing its performance on the original data. Second, consistent with recent findings on the difficulty of OOD mitigation, we find that naive contrastive fine-tuning fails (Yuan et al., 2023). Full-parameter fine-tuning leads to catastrophic forgetting, while freezing the encoder creates an "illusion of safety," improving performance on lower tiers while failing to remedy the collapse on narrative shifts.

1.1 Contributions

Our primary contributions are as follows:

- We propose a hierarchical evaluation framework to characterize distribution shift, distinguishing between syntactic, elaborate, and narrative shifts.

- We provide empirical evidence of a disconnect between lexical drift and model robustness, showing that a minimal increase in drift ($\Delta 0.06$) corresponds with a disproportionate collapse in performance.

- We demonstrate that standard small-data fine-tuning strategies are insufficient for establishing robustness, with frozen-encoder adaptation effectively masking worst-case brittleness on narrative shifts.

2 Related Work

2.1 Benchmarks and Contrast Sets

The fragility of pre-trained models on IID data has been extensively documented. Early diagnostic work revealed that models frequently exploit dataset artifacts rather than learning the intended task. For instance, Poliak et al. (2018) demonstrated that NLI models could often predict labels using the hypothesis alone, bypassing the premise entirely. Similarly, McCoy et al. (2019) introduced the HANS benchmark, showing that models rely on syntactic heuristics like lexical overlap, failing when these heuristics are decoupled from the label.

To address these blind spots, the field moved toward systematic behavioral testing. Ribeiro et al. (2020) proposed CheckList, a matrix of linguistic capabilities to test models against specific failure modes. Most relevant to our work is the concept of "contrast sets" introduced by Gardner et al. (2020), which advocates for manually perturbing test instances to probe local decision boundaries. While these approaches effectively diagnose local brittleness, they typically view distribution shift through a binary lens. Our work builds on this by structuring contrast sets hierarchically to evaluate robustness across a spectrum of qualitative shifts, explicitly distinguishing between benign lexical elaboration and structural narrative shifts.

2.2 Domain Generalization and Mitigation Failures

Beyond local perturbations, significant research has focused on OOD generalization. Benchmarks like WILDS (Koh et al., 2021) have demonstrated that models suffer substantial performance drops on real-world distribution shifts, such as moving between distinct data sources or time periods. However, these "in-the-wild" benchmarks often conflate different types of difficulty. While data augmentation is a common strategy for mitigating simple lexical shifts (Feng et al., 2021), recent evaluations suggest these methods fail to transfer to the structural shifts found in enterprise environments. Yuan et al. (2023) introduced the BOSS benchmark, demonstrating that classic mitigation strategies often fail to offer significant improvements over vanilla fine-tuning on challenging OOD data.

Our work provides nuance to this finding. While we corroborate the difficulty of OOD mitigation, we explicitly simulate the hierarchical drift of organizational settings: where a fact remains constant but the narrative frame changes (e.g., from general description to corporate reporting). By isolating this "narrative shift" (L3) from benign elaboration (L2), we show that mitigation strategies can succeed at the lexical level while masking catastrophic failure at the domain level.

3 Methodology

3.1 Base Data

We utilize the SQuAD 1.1 dataset (Rajpurkar et al., 2016) as our source domain. We selected a diverse subset of 120 examples from the SQuAD training set to serve as the foundation for our hierarchy. We intentionally drew from the training partition rather than the validation set to isolate robustness from generalization. Since the base model has ostensibly "learned" these examples during pre-training, any failure on the perturbed variants represents a failure of robustness to surface-form shifts rather than a lack of knowledge regarding the underlying context. By strictly evaluating perturbations of examples the model has already encountered, we explicitly control for knowledge retrieval. This adapts the principle of Invariance Testing (INV) (Ribeiro et al., 2020) to a diagnostic setting where failure can be definitively attributed to sensitivity to surface form. This experimental design explicitly simulates the enterprise deployment scenario where models are fine-tuned on internal documentation (e.g., manuals, wikis) but must reliably answer user queries framed in dynamic, operational contexts that differ from the static source text.

3.2 The Hierarchy of Qualitative Shift

While contrast sets are an established tool for diagnosing model brittleness (Gardner et al., 2020), manual construction can be labor-intensive. To scale the creation of our hierarchy, we employed a semi-automated, generative data augmentation approach (Ribeiro et al., 2020; Feng et al., 2021). We utilized a Large Language Model to generate candidate perturbations for three distinct levels of distribution shift, followed by rigorous human verification to ensure the original answer span remained valid within the context.

- L1 (Syntactic Variation): This level represents a "light paraphrase." The model was prompted to restrict edits to synonym replacement, tense changes, and minor word order permutations, ensuring the semantic "template" of the question remained intact.

- L2 (Elaborate Rephrasing): This level tests robustness to descriptive complexity. The generation process focused on rewriting questions to be significantly longer and more descriptive, inserting extra qualifiers or contextual details while remaining anchored in the same "world" or scenario as the base question.

- L3 (Narrative Shift): This level introduces a "regime change.” We re-contextualized both the question and the context paragraph into distinct domains (e.g., transforming a Wikipedia biography into a "corporate personnel file" or "compliance report"). While the answer string (e.g., "1992") remained invariant, its span index and surrounding tokens were shifted to align with the new narrative framing. This simulates true hierarchical domain drift, where the underlying fact remains constant but the document structure changes.

3.3 Quantifying Lexical Drift

To objectively measure the magnitude of these shifts, we utilized Jaccard distance as a metric for lexical drift. For a given pair of base question q_{base} and perturbed question q_{pert} the lexical drift is defined as:

$$d_{jac(q_{base},q_{pert})} = 1 - \frac{|V_{base} \cap V_{pert}|}{|V_{base} \cup V_{pert}|}$$

where V represents the set of unique, lowercased word tokens extracted via regex. This metric provides a scalar value representing the degree of lexical dissimilarity, allowing us to map model performance against a quantifiable axis of drift.

4 Experimental Setup

4.1 Data Partitions

To evaluate the impact of contrastive fine-tuning, we partitioned the 120 base examples into two distinct sets:

- Fine-tuning Set ($N = 80$): We aggregated the three perturbed counterparts (L1, L2, L3) for each of the 80 training examples into a single corpus of 240 examples. Notably, we excluded the original base questions from this set. This design rigorously tests "transfer" by forcing the model to adapt to the shifted distributions without being anchored by the original IID data during the fine-tuning phase.

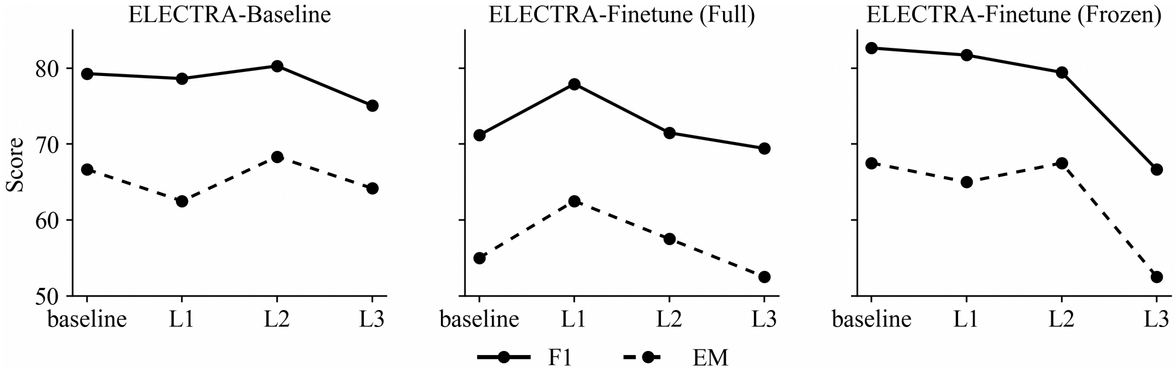

Figure 1: Performance (F1 and EM) across the hierarchical drift tiers. The Baseline model (left) exhibits non-monotonic robustness, peaking at L2 (elaborative rephrasing). In contrast, the Frozen Encoder (right) reveals a sharp "narrative cliff": while it successfully adapts to syntactic (L1) and elaborative (L2) shifts, it collapses catastrophically under narrative domain shift (L3), falling significantly below the baseline.

- Evaluation Set (N = 40): The remaining 40 base examples and their perturbations were held out. Evaluation is reported separately for each tier (Base, L1, L2, L3) to track the degradation curve on examples the model was *not* fine-tuned on.

4.2 Models

We evaluate the google/electra-small-discriminator model (Clark et al., 2020). We compare three configurations:

- ELECTRA-Baseline: The model trained on the full SQuAD dataset. This serves as the baseline to establish the model's competence on the 120 base examples and its baseline robustness to the contrast sets.

- ELECTRA-Finetune (Full): We fine-tune the baseline on the 240-example contrast set, updating all parameters. This represents a "pure" adaptation approach, where the model is updated solely on the failure modes and perturbations without rehearsing the source distribution (Gan and Ng, 2019).

- ELECTRA-Finetune (Frozen): We fine-tune the baseline on the 240-example contrast set, but freeze the encoder parameters, updating only the QA head. This strategy, often referred to as linear probing, is intended to prevent the distortion of pre-trained representations during adaptation to the target distribution (Niu et al., 2023).

4.3 Implementation Details

We utilized the Hugging Face Transformers library (Wolf et al., 2020). Data was tokenized with a maximum sequence length of 512 and a document stride of 128. For fine-tuning, we trained for 3 epochs with a batch size of 8 and a learning rate of 3e-5.[1] Performance is reported using Exact Match (EM) and F1 score, following the standard SQuAD evaluation protocols (Rajpurkar et al., 2016).

5 Results

5.1 Quantifying Lexical Drift

We first analyze the magnitude of distribution shift introduced by our hierarchical contrast set. Figure 2 illustrates the distribution of Jaccard drift scores across the three perturbation levels.[2]

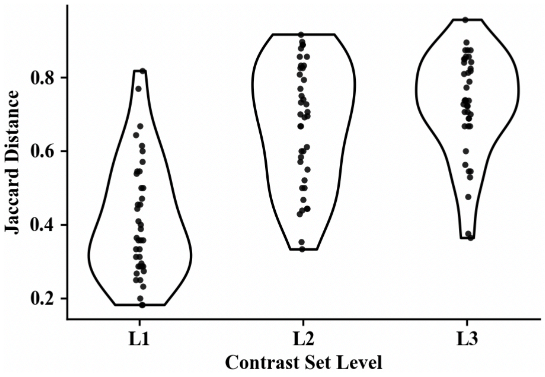

Figure 2: Distribution of Jaccard lexical distance scores across the three perturbation tiers. Note the significant overlap between L2 and L3, indicating that the L3 performance collapse is driven by framing rather than lexical drift.

[1] Full hyperparameters are detailed in Appendix A.

[2] Jaccard distribution statistics are provided in Appendix B.

- L1 (Syntactic Variation): We observe a mean drift of 0.41, confirming that even "light" paraphrasing introduces substantial lexical variation relative to the base.

- L2 (Elaborate Rephrasing): The mean drift increases significantly to 0.67, reflecting the introduction of descriptive clauses and qualifiers.

- L3 (Narrative Shift): The mean drift peaks at 0.73, driven by the extensive substitution of source-domain terms with out-of-distribution narrative markers.

Notably, the increase in drift from L2 to L3 (Δ 0.06) is minimal compared to the jump from L1 to L2 (Δ 0.26). This suggests that while L2 and L3 are lexically similar in magnitude, they differ fundamentally in qualitative framing, a distinction that lexical metrics fail to capture but, as shown below, strongly impacts model performance.

	EM	**F1**
ELECTRA-Baseline		
baseline	66.67	79.28
L1	62.50	78.63
L2	68.33	80.29
L3	64.17	75.08
ELECTRA-Finetune (Full)		
baseline	55.00	71.18
L1	62.50	77.92
L2	57.50	71.48
L3	52.50	69.42
ELECTRA-Finetune (Frozen)		
baseline	67.50	82.66
L1	65.00	81.72
L2	67.50	79.45
L3	52.50	66.64

Table 1: Evaluation results (EM/F1) on the baseline, L1, L2 and L3 contrast sets by model.

5.2 Baseline Robustness: The Non-Monotonic Curve

We evaluated the pre-trained ELECTRA-Base model on the 40 held-out evaluation examples. The results (see Figure 1) reveal a counter-intuitive, non-monotonic relationship between drift and performance.

- L1 Sensitivity: Performance drops from the in-distribution baseline 66.67 EM to 62.50 EM on L1 (Table 1). This indicates that simple synonym replacements and syntactic shuffling are sufficient to disrupt the model's shallow heuristics.

- The L2 "Sweet Spot": Surprisingly, performance peaks on L2, achieving 68.33 EM and 80.29 F1, surpassing the original baseline performance. This suggests that the "elaborate rephrasing" at this level, despite high lexical drift, may introduce semantic anchors or keywords that clarify the question for the model.

- L3 Collapse: This robustness evaporates at L3. Despite a minimal increase in lexical drift compared to L2, performance degrades to 64.17 EM and 75.08 F1. This drop highlights the model's specific sensitivity to narrative reframing, separate from pure lexical distance.

5.3 Failure of Naive Mitigation

We investigated whether fine-tuning on the 240-example contrast set could mitigate these failures. We compared full-parameter fine-tuning against a frozen-encoder approach.

Catastrophic forgetting (Full Fine-tune approach): Naive, full-parameter fine-tuning proved detrimental. As detailed in Table 1, performance on the original baseline examples plummeted from 66.67 EM to 55.00 EM. This degradation extended to the contrast sets, with L3 performance dropping to 52.50 EM. This confirms that small-scale fine-tuning on OOD data induces catastrophic forgetting of the source domain features.

The “Illusion of Safety” (Frozen Encoder approach): Freezing the encoder yielded a deceptive robustness profile (Figure 1, right chart).

- In-Distribution Success: The model increased or maintained performance on the baseline (67.50 EM) and L1/L2 variants.

- Hidden Brittleness: However, the model collapsed on L3, dropping to 52.50 EM (an 11.67 point drop relative to the un-fine-tuned baseline).

This result is particularly concerning: the frozen-encoder model appears robust on standard metrics (Baseline/L1/L2) but effectively masks its fragility to narrative shifts until the regime change occurs on L3.

6 Qualitative Analysis

To better understand the non-monotonic robustness curve observed in Section 5, we conduct a qualitative analysis of specific high-drift examples.[3] We examine why the model successfully demonstrates lexical adaptability on elaborate rephrasing (L2) but fails on domain shifts (L3), even when applied to the same underlying fact.

6.1 Lexical Adaptability in L2

Our quantitative results showed that the baseline model performed best on L2 examples, despite their high lexical drift (mean 0.67). Examination suggests that L2 perturbations often act as "semantic anchors." By adding descriptive qualifiers, these questions increase the lexical distance, but simultaneously reinforce the connection to the answer span. Consider this L2 example (Drift: 0.917):

Base: *Where do you find silicate?*

L2: *In the geoscience unit's internal training slide on serpentinization, where is the relevant silicate described as being located within the mineral structure?*

Despite the introduction of complex vocabulary ("serpentinization", “geoscience unit”), the core entity ("silicate") remains the syntactic focus and the model correctly retrieves the answer. The elaborate phrasing likely provides unique token overlaps (e.g., "mineral structure") that exist in the context paragraph, narrowing the search space. This confirms that the model possesses high lexical adaptability: it is robust to "noise" as long as the underlying semantic domain remains consistent. This illustrates why high lexical drift does not necessarily degrade performance: the added tokens are domain-consistent, acting as signal rather than noise.

6.2 Narrative Shifts as Domain Mismatch in L3

In contrast, L3 examples introduce a "narrative shift" that decouples the question from the original context's genre. Even when lexical drift is similar to L2, the framing causes failure. Consider the L3 variant of the same base question (Drift: 0.875):

Base: *Where do you find silicate?*

L3: *Where did the lab report say silicate impurities were embedded inside the furnace lining?*

Here, the narrative frame shifts from geology to industrial manufacturing ("lab report," "furnace lining"). Notably, this L3 example has lower lexical drift (0.875) than the L2 example (0.917), yet it introduces a fundamental domain mismatch. The model must now resolve the query within a specialized "manufacturing" frame despite being pre-trained on general "scientific" Wikipedia contexts.

This highlights the "Narrative Cliff" relevant to enterprise deployment: just as a model trained on public data struggles to interpret a proprietary "lab report," it fails here because the *document structure* (a report finding) conflicts with the *knowledge source* (an encyclopedia entry). The difficulty stems not from the quantity of new words, but from the introduction of out-of-distribution distractors (e.g., the "lab report" framing) that break the model's feature mapping.

7 Discussion

7.1 Robustness is Non-Monotonic

A prevailing assumption in robustness research is that greater distribution shift equates to greater difficulty. Our findings challenge this linearity. The non-monotonic performance curve, exhibiting a local peak or hill-shaped pattern at high-drift L2, suggests that lexical drift (as quantified by Jaccard distance) is not a uniform stressor.

We posit that elaborative drift (L2) fundamentally differs from distractive drift (L1/L3). By expanding the question with domain-consistent descriptors, L2 perturbations likely introduce "semantic anchors" (redundant keywords that intersect with the passage context) effectively narrowing the search space for the model. In contrast, L3 perturbations, despite having similar lexical distance to L2, remove these anchors and replace them with out-of-domain distractors (e.g., "corporate committee"). This indicates that robustness evaluations must distinguish between *informative* shifts (which may aid the model) and *adversarial* shifts (which test invariance), rather than relying solely on surface-level quantitative metrics.

[3] See Appendix C for a broader set of randomly selected examples across all tiers.

7.2 The Dangers of Naive Fine-tuning

Our mitigation experiments yield a critical negative result: naive fine-tuning on small contrast sets is not a "free lunch.”

Catastrophic Forgetting: The full-parameter fine-tuning run demonstrates that optimizing for a narrow slice of "hard" examples (240 contrast pairs) rapidly overwrites the model's general linguistic features, leading to a performance collapse on the original distribution. This severe degradation is a hallmark of catastrophic forgetting (Kirkpatrick et al., 2017) and corroborates recent findings by Yuan et al. (2023), who reported that classic mitigation strategies often fail to improve upon vanilla fine-tuning in OOD settings.

The Illusion of Safety: The frozen-encoder approach presents a more subtle danger. By preserving the encoder, the model retains its baseline performance and adapts well to syntactic (L1) and elaborative (L2) shifts. However, this stability is deceptive; the head-only adaptation fails completely when the narrative frame breaks at L3. For practitioners, this implies that "light" adaptation methods can mask a model's underlying brittleness, creating a false sense of security that vanishes only when specific "regime changes" occur in deployment.

7.3 Beyond Binary Evaluation

The distinct failure modes at L3 underscore the limitations of monolithic OOD evaluation frameworks (Koh et al., 2021). A model evaluated solely on L1 (paraphrases) or L2 (elaborations) would appear robust, yet it would fail catastrophically in an enterprise setting requiring L3 (narrative) invariance. To accurately characterize reliability, evaluation benchmarks must be hierarchical, explicitly isolating "narrative framing" as a distinct dimension of distribution shift separate from lexical overlap.

7.4 Relevance to Large Language Models

While our experiments utilize ELECTRA-small to isolate mechanistic failures, the phenomena we observe—non-monotonic robustness and fine-tuning instability—mirror critical challenges in contemporary Large Language Models (LLMs). Recent work has shown that parameter-efficient fine-tuning (e.g., LoRA) on aligned LLMs often induces catastrophic forgetting of safety guardrails, even when the fine-tuning data is benign (Qi et al., 2024; Hsu et al., 2024). This directly parallels the "illusion of safety" we observed in our frozen-encoder experiments, where superficial metrics (L1/L2 or safety on standard prompts) remain high while deep robustness (L3 or safety under adversarial framing) collapses.

Similarly, the non-monotonic performance we observe under elaborative paraphrasing (L2) aligns with "U-shaped" performance curves seen in LLM context processing. Liu et al. (2024) demonstrated that LLMs often fail to retrieve information located in the middle of long contexts ("Lost in the Middle"), a phenomenon where added verbosity, similar to our L2 perturbations, can unpredictably aid or hinder retrieval depending on its position and framing. By characterizing these failures in a controlled encoder environment, we provide a transparent view of the structural brittleness that persists in larger, opaque generative architectures.

8 Limitations and Future Work

While our findings provide distinct insights into the nature of hierarchical drift, several limitations regarding scale and scope should be noted.

First, our experimental scale was constrained by the resource-intensive nature of verifying contrast sets. We limited our investigation to 120 base examples and a single model architecture (ELECTRA-small). While consistent with pilot studies in robustness literature (Gardner et al., 2020), future work should validate these trends across larger datasets and more recent architectures (e.g., DeBERTa-v3 (He et al., 2021) or generative LLMs) to confirm if the "L2 peak" and "L3 collapse" persist at scale.

Second, by drawing base examples from the source training distribution, our study strictly evaluates *invariance* (robustness to surface perturbations of known contexts) rather than *generalization* to unseen passages. While this isolates the variable of lexical framing, it remains an open question whether the "semantic anchoring" effect of L2 persists when the model must attend to novel contexts it has never encoded before.

Third, our reliance on Jaccard distance as a primary metric for lexical drift is a simplification. While effective for measuring surface-level overlap, it does not capture semantic embedding distance or syntactic complexity. A more nuanced metric would perhaps leverage embedding-based distance (e.g., cosine similarity of [CLS] tokens) as that might better disentangle why L2's "elaborative" drift aids performance while L3's "narrative" drift harms it.

Finally, our mitigation strategies were limited to standard contrastive fine-tuning. We observed catastrophic forgetting, a known issue in small-data adaptation. Future research should explore advanced regularization techniques, such as Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) or Replay Buffers (Chaudhry et al., 2019), to better balance the trade-off between acquiring robustness to narrative shifts and maintaining in-distribution competence.

9 Conclusion

In this work, we challenged the binary view of distribution shift by evaluating QA robustness through a hierarchical framework. By constructing a three-tiered contrast set ranging from syntactic variation to narrative reframing, we demonstrated that lexical adaptability and domain generalization are distinct capabilities. Our experiments revealed a non-monotonic performance curve where ELECTRA-small effectively leverages elaborative rephrasing (L2) to improve over the baseline, yet hits a "narrative cliff" when exposed to the regime changes of L3.

Furthermore, our investigation into mitigation strategies highlighted the risks of naive adaptation. We found that standard fine-tuning on contrast sets leads to catastrophic forgetting, while freezing the encoder provides only an illusory safety, masking the model's fundamental brittleness to narrative shifts. These findings suggest that true robustness cannot be achieved by simply augmenting training data with "hard" lexical examples. Instead, future evaluation frameworks, particularly for enterprise deployment, must explicitly isolate "narrative framing" as a distinct dimension of generalization, ensuring that models are robust not just to *how* a question is phrased, but to *where* it is conceptually situated.

References

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019. On Tiny Episodic Memories in Continual Learning. *arXiv preprint arXiv:1902.10486*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.

Chia-Yi Hsu, Yu-Lin Tsai, Chih-Hsun Lin, Pin-Yu Chen, and Chia-Mu Yu. 2024. Safe LoRA: The Silver Lining of Reducing Safety Risks when Finetuning Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian

Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

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

Lifan Yuan, Linyi Yang, Leyang Cui, Wenyuan Zhang, and Yue Zhang. 2023. Revisiting Out-of-distribution Robustness in NLP: Benchmark, Analysis, and LLMs Evaluations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Appendices

A. Training Hyperparameters

We utilized the google/electra-small-discriminator model for all experiments. Fine-tuning was performed on a T4 GPU via Google Colab.

Hyperparameter	**Value**
Base Model	google/electra-small-discriminator
Batch Size	8
Learning Rate	3e-5
Epochs	3
Max Sequence Length	512
Document Stride	128
Optimizer	AdamW
Weight Decay	0.01
Warmup Steps	0

Table A1: Hyperparameter configuration for full and frozen fine-tuning runs.

B. Jaccard Drift Distribution

We observed the following distribution statistics for Jaccard drift across the 40 evaluation examples per level.

Statistic	L1 (Syntactic)	L2 (Elaborate)	L3 (Narrative)
Mean	0.4083	0.6724	0.7251
Std Dev	0.1570	0.1647	0.1376
Minimum	0.1818	0.3333	0.3636
Maximum	0.8182	0.9167	0.9565

Table B1: Jaccard drift statistics by perturbation level.

C. Example Contrast Sets

To demonstrate the qualitative distinctions between our perturbation levels, we provide additional examples from the hierarchical contrast sets below.

Level	Drift	Question
Base		**Why have DDTs been banned in some areas?**
L1	0.64	Why have DDT and similar pesticides been banned in many countries?
L2	0.85	In the internal briefing prepared for the global malaria-control program, what key concerns are cited as the reasons some regions in our organization have banned DDT?
L3	0.96	Why did the corporate risk committee decide to ban the legacy degreasing solvent from all production units?
Base		**Why were sanctions place on Liberian timber exports?**
L1	0.50	Why were UN sanctions imposed on Liberia's timber exports?
L2	0.90	In the compliance team's sanctions case study on Liberia, what specific concern about timber-export revenues is documented as the reason for the 2003 restrictions?
L3	0.90	Why did the executive committee freeze the mining division's timber contracts with the subcontractor?
Base		**Where can people using iPods on planes view the device's interface?**
L1	0.67	On airplanes, where can passengers view their iPod's library and interface?
L2	0.79	According to the airline's in-flight product brief, where on the aircraft can passengers mirror their iPod interface once the seat connector is available?
L3	0.88	Where did the airline route passengers' device interfaces so they could watch content more comfortably?

Table C1: Randomly selected examples of hierarchical drift. Note: For L2 and L3, the corresponding context paragraphs were also perturbed to align with the descriptive or narrative shifts shown here.