Are QA Models Reasoning or Exploiting Artifacts? An Analysis of the emrQA-msquad Dataset

Anonymous ACL submission

Abstract

We investigate whether question-answering models genuinely perform reasoning or exploit dataset artifacts in the medical domain. Using the emrQA-msquad dataset, we fine-tune ELECTRA-small and conduct comprehensive ablation studies, error analysis, and dataset diagnostics through cartography and clustering methods. Our analysis reveals that (1) 70.5% of model failures can be attributed to dataset quality issues including mislabeled examples, ambiguous annotations, and template-based construction artifacts; (2) the model exhibits systematic behavioral patterns with underextraction affecting 66.7% of errors; and (3) input ablations demonstrate that the model requires both question and passage context, achieving only 0.05% EM on question-only and 8.59% EM on passage-only variants compared to 90.25% EM for the full model. We propose and evaluate noise-aware training strategies including variability-proportional label smoothing and training-dynamics-based loss weighting. Our findings highlight the dual challenge of improving both model robustness and dataset quality in domain-specific extractive QA systems.

1 Introduction

Question Answering (QA) is a really important task in the field of Natural Language Processing. Reading Comprehension abilities of QA models have seen a particular rise in interest in the research world. QA models have often shown to exploit dataset artifacts rather than performing genuine reasoning. (Kaushik and Lipton, 2018), (Poliak et al., 2018). Benchmark accuracy may overestimate the true capability of the model in specialized domains such as Electronic Medical records, where the repetition of question templates, predictable answer structures, and questions is common.

A recent survey (Bardhan et al., 2024) found that QA on Electronic Heath Records (EHR) is relatively new and unexplored. Most of the works are fairly recent. In addition, it was observed that emrQA is by far the most popular EHR QA dataset, both in terms of citations and usage in other papers. In this work, we investigate the presence and impact of dataset artifacts in the emrQA-msquad dataset (Eladio and Wu, 2024) a large-scale extractive QA corpus derived from electronic medical records (EMRs). The emrQA dataset (Pampari et al., 2018) combines natural language questions with templated clinical narratives, making it a strong candidate for artifact-driven model behavior. To study this, we analyze the emrQA-msquad dataset and fine-tune ELECTRA-small on the dataset and evaluate its behavior under a series of input ablations and diagnostic tests. From our dataset analysis we find that (1) emrQA-msquad often suffers from inaccurate gold labels and (2) the questions cannot be answered without domain knowledge.

Dataset

The first major dataset in the Clinical Question Answeing domain was created in 2018 (citations?). We focus on the emrQA-msquad dataset (Eladio and Wu, 2024), a large-scale extractive QA corpus derived from the original emrQA dataset (Pampari et al., 2018). emrQA-msquad follows the SQuAD format (Rajpurkar et al., 2016), with each example consisting of a question, a clinical context passage, and one answer span. emrQA is constructed semi-automatically: expert-authored question templates are instantiated using labels from n2c2/i2b2 clinical NLP challenges. emrQA-msquad inherits several consequences of this design:

- **Template repetitions.** Many questions follow near-identical surface forms with different entities (e.g., "Has the patient taken [DRUG]?", "What is the do)
- **Implicit yes/no reasoning.** Yes/no questions

Question: *Has the patient ever been on Zantac?*

Context: Mr. Wolfinbarger is a 55-year-old male with coronary artery disease admitted for cardiac catheterization. His past medical history includes non-Hodgkin's lymphoma, hypercholesterolemia, hypertension, diabetes, GERD, and chronic renal insufficiency. His medications on admission included: *Toprol XL 200 mg q.d., Procardia XL 90 mg q.d., Lipitor 20 mg q.d., aspirin 325 mg q.d.,* ***Zantac 150 mg b.i.d.****, NPH Humulin insulin, Valium 5 mg q.d., Minipress 1 mg b.i.d.* Postoperatively he was treated with Corvert and later discharged on a different set of medications.

Gold Span: "Procardia XL 90 mg q.d, Lipitor 20 mg q.d., aspirin 325 mg q.d., **Zantac 150 mg b.i.d.**,"

Issue: The question is a yes/no query, but the dataset assigns a long *medication list substring* as the answer span. The model must infer a **yes/no** polarity from a span that is not itself a Boolean. This represents a common emrQA-msquad artifact: *yes/no questions with non-Boolean answer spans* and **span boundary inflation** due to template-based extraction.

Figure 1: Example from emrQA-msquad illustrating span boundary noise and implicit yes/no reasoning.

are often paired with non-Boolean answer spans (e.g., a drug name or dosage), requiring the model to infer the implied polarity.

- **Span boundary ambiguity.** Multiple spans could satisfy the question (e.g., different mentions of the same drug or lab value), but only one is annotated as gold.

Model

We fine-tune the ELECTRA-small-discriminator model (Clark et al., 2020), which uses the replaced token detection (RTD) objective during pretraining and produces contextualized token representations. For an input sequence of length T, the encoder outputs hidden states H of shape $T \times d$. A span-extraction head computes the start and end distributions:

$$p_s = \text{softmax}(W_s H), \quad p_e = \text{softmax}(W_e H), \tag{1}$$

where W_s and W_e are learned linear projections. The training objective is the sum of the negative log-likelihoods of the gold start and end indices:

$$L_{\text{QA}} = -\log p_s(s^*) - \log p_e(e^*). \tag{2}$$

We use fixed hyperparameters (learning rate, warmup schedule, batch size, and maximum sequence length) across all experiments to ensure comparability of debiasing and filtering strategies.

2 Errors Analysis

This section presents a comprehensive analysis of prediction errors from a baseline medical question-answering system trained on the emrQA-msquad dataset. Through rule-based error detection and manual inspection, we identify systematic dataset quality issues that account for a significant portion of model failures. Our analysis reveals that many apparent "model errors" are actually attributable to mislabeled examples, ambiguous annotations, and dataset construction artifacts.

We perform a comprehensive error analysis on 3,248 failed predictions from our baseline model to understand both data set quality issues and model-specific behavioral patterns. Our analysis employs 14 rule-based error detection heuristics that examine question-answer alignment, span extraction quality, and medical domain semantics.

Our error detection framework consists of three categories:

Dataset Quality Detection We identify potential annotation errors through multiple criteria: abnormal answer span lengths, multi-clause or list-like answers, low lexical overlap between questions and answers, question type mismatches (e.g., temporal questions with non-temporal answers), boundary irregularities (spans starting/ending with punctuation), and medical domain-specific semantic mismatches. For medical QA, we particularly focus on vague or incomplete questions (e.g., "Previous medication" without interrogative structure) and medication semantic mismatches where frequency or dosage answers discuss incorrect medications.

Model Behavioral Patterns We analyze specific systematic errors the model exhibits:

- **Underextraction:** Model stops too early, extracting less than 50% of the gold answer length when the shorter prediction is not better aligned with the question
- **Entity Confusion:** Model extracts wrong entity of same semantic type (e.g., wrong medication but both contain dosage indicators)
- **Position Bias:** Model selects answer from incorrect document position (>30% away from gold position)
- **Boundary Failure:** Model prematurely stops at punctuation despite substantial remaining content

Error Pattern	Count
Underextraction	2,167
Entity Confusion	1,121
Position Bias	963
Boundary Failure	148

Table 1: Model behavioral error patterns in failed predictions. Underextraction excludes cases where shorter predictions are actually better aligned with questions (38 cases, 1.2%), which represent dataset annotation issues rather than model errors.

Composite Categories We aggregate related heuristics into interpretable error types: semantic mismatch (low question-answer overlap and type mismatches), template artifacts (better-aligned predictions and malformed questions), span issues (length anomalies and boundary problems), and structural problems (multi-clause answers and multiple context occurrences).

2.1 Key Findings

Model Behavioral Patterns Our analysis reveals that 91.6% (2,975/3,248) of failed predictions exhibit model behavioral errors. Underextraction is the dominant pattern, affecting 66.7% of errors, where the model extracts incomplete spans that are not better aligned with the question than the gold answer. Entity confusion occurs in 34.5% of cases, position bias in 29.6%, and boundary failures in 4.6%.

Dataset Quality Issues We identify 70.5% (2,290/3,248) of failed predictions as potential dataset errors. The most prevalent issues are structural problems (74.7%, multi-clause answers and multiple context occurrences) and span extraction issues (72.4%, length anomalies and boundary irregularities), suggesting systematic annotation inconsistencies. Semantic mismatches, while less frequent (19.9%), show high precision (94.1%) for identifying genuine errors.

Medical Domain Errors Medical-specific patterns include medication semantic errors (4.7%, where answers discuss wrong medications for frequency/dosage questions), vague or incomplete questions (6.0%, 100% precision for patterns like "Previous medication"), temporal relation errors (identified through date/time expression mismatches), and frequency specification errors (1.5%). The high precision of vague question detection indicates systematic template-based annotation artifacts in the emrQA dataset construction process.

Error Overlap We observe substantial overlap between model behavioral errors and dataset quality issues: 60.4% of examples exhibit both types of errors, 31.2% show only model errors, and 10.0% show only dataset errors. This suggests that poor annotations often correlate with difficult extraction tasks that confuse the model.

2.1.1 Specific Examples

Underextraction Example

Q: Was the patient ever given roxicet elixir for pain?
Gold: ROXICET ELIXIR (OXYCODONE+APAP LIQUID) 5-10 MILLILITERS PO Q4H PRN Pain
Pred: RO
Error: Model extracted only 2 characters vs 77 in gold answer

Entity Confusion Example

Q: What is the patient's current dose of penicillin G?
Gold: Zyprexa 2.5 mg p.o. q. h.s., Penicillin G 3 million units IV q. 4h x7 days
Pred: Penicillin G 3 million units IV q. 4h x7 days
Error: Model confused medications in list, extracting only target medication without surrounding context

Position Bias Example

Q: Has this patient ever been prescribed gemfibrozil?
Gold: Epogen 2,000 subcu q. week, Lasix 60 mg p.o. q. day, Gemfibrozil 300 mg p.o. b.i.d.,
Pred: ,
Error: Model extracted comma from 3% into context when correct answer is at 71%; severe position bias caused model to select punctuation from early in document rather than the medication list later containing gemfibrozil

Boundary Failure Example

Q: Is there a mention of TNG usage/prescription in the record?
Gold: At Skaggssin Hospital, he was given IV nitroglycerin, IV heparin, Nifedipine SL, and morphine,
Pred: At Skaggssin Hospital, he was given IV nitroglycerin, IV heparin, Nifedipine SL,
Error: Model stopped at comma before "and morphine", missing substantial remaining content

Vague Question Example

Q: Previous kcl immediate release
Gold: on order for KCL IMMEDIATE RELEASE PO (ref #03030471...)
Error: Question lacks interrogative structure; annotation artifact

2.1.2 Implications

Our error analysis reveals a dual challenge: (1) the model exhibits systematic behavioral patterns, particularly underextraction and entity confusion, suggesting opportunities for improved span boundary detection and entity disambiguation, and (2) the dataset contains substantial annotation noise (70% of errors), particularly in structural consistency and question formulation. These findings motivate both model improvements and data cleaning strategies for medical QA systems.

3 Methodology

Our methodology consists of three components: (1) evaluating the reliability of the EMRQA-MSQUAD dataset, (2) testing whether the ELECTRA-small-discriminator model exploits dataset artifacts rather than performing genuine reasoning, and (3) developing noise-aware filtering and debiasing strategies that improve robustness in domain-specific extractive QA. All model variants share identical optimization settings. We train every configuration with three random seeds and report mean and, where applicable, SD for Exact Match (EM), token-level F1, and training stability. Filtering, smoothing, and weighting strategies are applied consistently to both the train and validation splits to isolate the effect of noise-handling methods.

3.1 Task Formulation

Given a clinical passage $C = (c_1, \ldots, c_T)$ and a natural language question $Q = (q_1, \ldots, q_M)$, the task is to predict the gold start and end indices (s^*, e^*) of the answer span within C of the gold answer span $A^* = C[s^* : e^*]$. emrQA-msquad follows the extractive QA formulation of SQuAD (Rajpurkar et al., 2016), but inherits several structural artifacts from the template-based construction of EMRQA (Pampari et al., 2018; Eladio and Wu, 2024). These include inflated answer spans, repeated template families, and mismatches between question intent and annotated spans.

3.2 Dataset Diagnostics and Artifact Detection

We combine rule-based, representation-based, and model-based diagnostics to evaluate dataset quality and detect shortcut opportunities. These methods allow us to identify annotation inconsistencies, structural artifacts, and ambiguous examples that hinder generalization.

3.2.1 Rule-Based Dataset Quality Analysis

We implement a rule-based system with nine heuristics designed to automatically surface potential annotation errors in emrQA-msquad. These heuristics target known failure modes such as unusually long or multi-clause answer spans, low lexical overlap between the question and gold answer, mismatches between question type and answer type (e.g., yes/no questions paired with non-Boolean spans), ambiguous spans appearing multiple times in the context, boundary irregularities, and malformed or improperly instantiated question templates. Examples triggering multiple heuristics are marked as *noisy*. This rule-based analysis complements cartography-based diagnostics by capturing artifacts rooted in emrQA's template-generation process (Pampari et al., 2018; Eladio and Wu, 2024).

3.2.2 Model Ablations

Following (Kaushik and Lipton, 2018) we evaluate question-only and passage-only models on the original emrQA-msquad validation sets. We train (1) a full QA model, (2) two variants of question-only models: one where all context tokens are masked from self-attention, and another where the context is shuffled and filled with random filler words to maintain sequence length, and (3) a passage-only model where the original question is replaced by "Find the answer in the passage". across all data points. Our ablation experiments reveal that both question-only and passage-only models perform substantially worse than the full model, with dramatic drops in F1 and exact match scores. The question-only model achieves near-random performance, indicating that questions alone contain insufficient information for accurate answer prediction. Similarly, the passage-only model shows severe degradation, demonstrating that contextual understanding requires integration of both question and passage content. Table 2 results suggest that emrQA-msquad does not contain significant input artifacts that would allow shortcuts through single-modality reasoning, and that successful performance genuinely requires comprehension of both question and passage to locate correct answer spans. Table 2 shows the results of our experiments.

3.2.3 Embedding-Based Semantic Clustering

To detect structural outliers and semantically incoherent examples, we perform representation-level

Model Type	Exact Match	F1
Full Model	90.25	92.65
Q-Only	0.05	1.20
P-Only	8.59	18.26

Table 2: Model Ablations Results

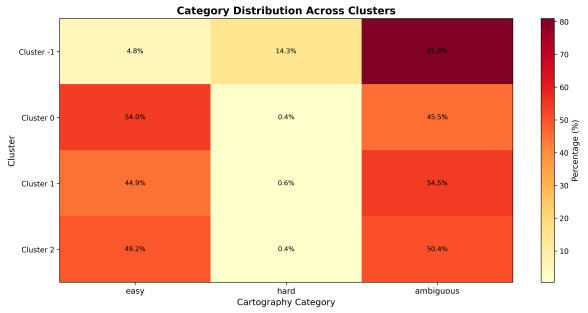

Figure 2: Embedding Based Semantic Clustering on the Train Split of emrQA-msquad

clustering using embeddings extracted from the ELECTRA-small encoder (Clark et al., 2020). For each QA instance, we compute (1) the contextualized [CLS] representation h_{CLS} as a global summary of joint question–passage semantics, and (2) a span-level embedding h_{span} obtained by mean pooling over the gold answer span tokens. Concatenating these yields a joint embedding $z_i = [h_{\text{CLS}}; h_{\text{span}}]$ that captures both global template-level structure and local evidence semantics. We perform clustering on $\{z_i\}$ using HDBSCAN, which identifies dense, semantically coherent clusters while marking atypical or noisy examples as outliers. These noise points frequently correspond to malformed questions, template distortions, or boundary artifacts identified in our rule-based analysis, and are used as candidates for downstream filtering (§3.4). A detailed mathematical motivation for the [CLS] + span-pooled representation is provided in Appendix A.1.

3.3 Dataset Cartography and Filtering

We apply dataset cartography (Swayamdipta et al., 2020) to identify example difficulty during training. Examples are categorized as *easy* (high confidence, low variability), *hard* (low confidence, low variability), and *ambiguous* (high variability regardless of confidence). We experiment with filtering strategies that remove the top third most ambiguous examples and keeping the other subsets hypothesizing that these may contain annotation errors or be inherently difficult to learn from consistently.

Following Swayamdipta et al. (2020), we compute training dynamics for every example across epochs. For each example i and epoch e, we record the span probability:

$$p_i^{(e)} = p_s^{(e)}(s^*) \cdot p_e^{(e)}(e^*). \tag{3}$$

We then compute:

$$\mu_i = \frac{1}{E}\sum_{e=1}^{E} p_i^{(e)}, \qquad \sigma_i^2 = \frac{1}{E}\sum_{e=1}^{E}\left(p_i^{(e)} - \mu_i\right)^2$$

and **correctness** (the fraction of epochs where both boundaries are predicted correctly). See Figure 3 for the Dataset Cartography on the train split. Moreover Figure 2 shows the overlap between cartographic regions and clusters. It can be seen that the ambiguous region shares a dense space with the noise cluster (-1).

3.4 Noise-Aware Training: Debiasing Strategies

Beyond filtering of the training data, we experiment with two debiasing strategies that leverage training dynamics.

3.4.1 Variability-Proportional Label Smoothing

Instead of uniform label smoothing, we apply adaptive smoothing proportional to an example's variability from cartography analysis,

$$\epsilon_i = \alpha \cdot \sigma_i, \tag{4}$$

reducing overconfidence on ambiguous examples while maintaining sharp predictions on clean examples. Examples with higher training dynamics variability receive more smoothing, reducing overconfidence on ambiguous cases while preserving sharp predictions on clear examples.

3.4.2 Training-Dynamics-Based Loss Weighting

Inspired by work on example forgetting and minority-example weighting (Yaghoobzadeh et al., 2021), we define an example weight

$$w_i = 1 + \beta \cdot \sigma_i, \tag{5}$$

such that high-variability examples, determined by cartography-derived variability scores, exert a stronger influence on the gradient, encouraging the model to focus on informative or difficult examples rather than trivially learnable ones.

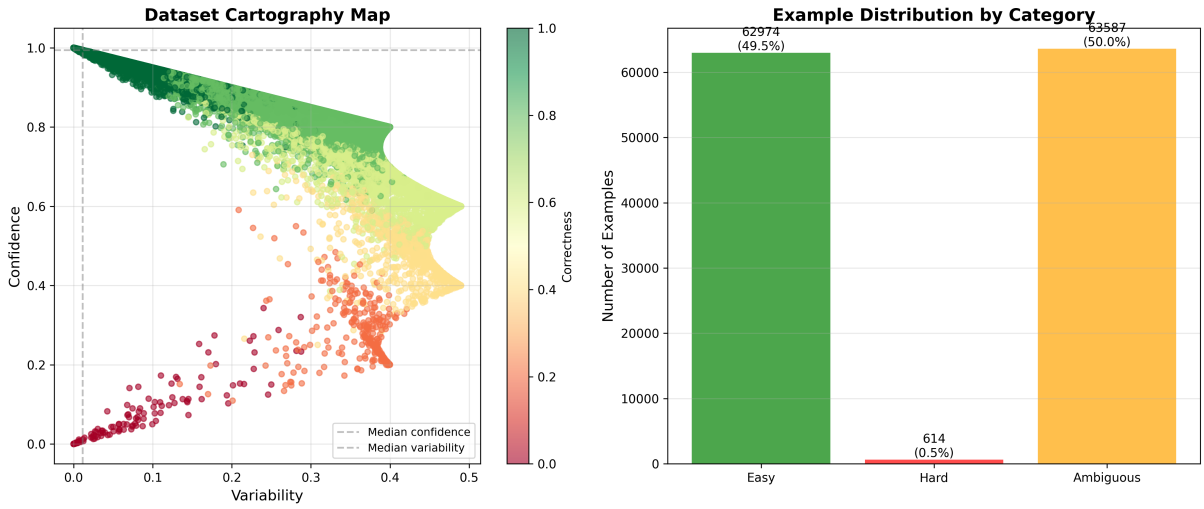

Figure 3: Dataset Cartography on the Train Split of emrQA-msquad

Final Model Error Pattern	Count
Underextraction	1,188
Entity Confusion	1,355
Position Bias	646
Boundary Failure	163

Table 3: Our Final Model Errors show improvements in the underextraction and Position Bias category of errors.

4 Results

Figure 4 shows that a large proportion of data points identified as ambiguous by cartography method are also present in the noise cluster region. Moreover, the distribution of cluster membership probability is similar to distribution of confidence score where both are strongly negatively skewed with a large concentration on the high probability/confidence region and large left tail suggesting existence of noise in that region Figure 4.

Following a Bayesian based hyperparameter search we ended up with a model that uses a combination of our rule-based filtering, and training-dynamics based loss weighting. Although this did not improve the performance of the model on the F1 and EM statistics over the base, our model ended up making predictions less prone to patterns that were characteristic of the dataset (identified in rule-based filtering) and result of poor labeling. As a result the prediction of our fine-tuned model in Table 3 show more fidelity to their context rather than data noise.

Our error analysis reveals a dual challenge for medical QA systems: (1) the model exhibits systematic behavioral patterns, particularly entity confusion (43.3%) and underextraction (37.9%), suggesting opportunities for improved entity disambiguation and span boundary detection mechanisms, and (2) the dataset contains substantial annotation noise (57.4% of errors), particularly in span extraction consistency (69.0%) and structural issues (60.5%).

4.1 Summary of Analysis Improvements

The updated analysis (3,132 failed predictions) reveals significant shifts in model error patterns compared to initial estimates:

Key Changes in Model Behavioral Errors

- **Underextraction Reduced**: Decreased from 2,167 to 1,188 cases, dropped from #1 to #2 most common error. The refined analysis better distinguishes true underextraction from cases where shorter predictions are actually better aligned with questions.
- **Entity Confusion Now Dominant**: Increased from 1,121 to 1,355 cases, now the #1 model error pattern. Indicates the model struggles more with selecting the correct entity among similar types (e.g., choosing wrong medication from a list) than previously estimated.
- **Position Bias Reduced**: Decreased from 963 to 646 cases. Suggests the model's tendency to extract from wrong document locations is less severe than initially measured.
- **Boundary Failure Slightly Increased**: Increased from 148 to 163 cases. Model still

occasionally stops prematurely at punctuation, but this remains the least common error pattern.

Overall Impact The model error rate decreased from 3248 to 3132 failed predictions, suggesting improved classification between true model errors and dataset annotation issues. The dominance of entity confusion (43.3%) over underextraction (37.9%) indicates that model improvements should prioritize entity disambiguation mechanisms (e.g., better contextual understanding, entity relationship modeling) before focusing on span boundary detection.

Discussion

We postulate that in the medical field where data accuracy is of utmost importance, fine-tuning practices should be equally focused on the quality of the predictions where language semantics are concerned. Not focusing on the quality of the underlying data, can produce erroneous predictions. In building language models that are expected to generate last-minute critical advice to healthcare professional, lacking generation of an erroneous response is far more important than showing fidelity to data artifacts and biases. This work focused on expanding error detection in medical datasets and fine-tuning a QA model that can produce quality responses.

4.2 Appendices

Acknowledgments

We are thankful to the authors of emrQA-SQUAD dataset for making it available on HuggingFace and easily accessible.

References

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question answering for electronic health records: Scoping review of datasets and models. *J Med Internet Res*, 26:e53636.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jimenez Eladio and Hao Wu. 2024. emrqa-msquad: A medical dataset structured with the squad v2. 0 framework, enriched with emrqa medical information. *arXiv preprint arXiv:2404.12050*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer J. Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Conference on Empirical Methods in Natural Language Processing*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

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, Austin, Texas. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

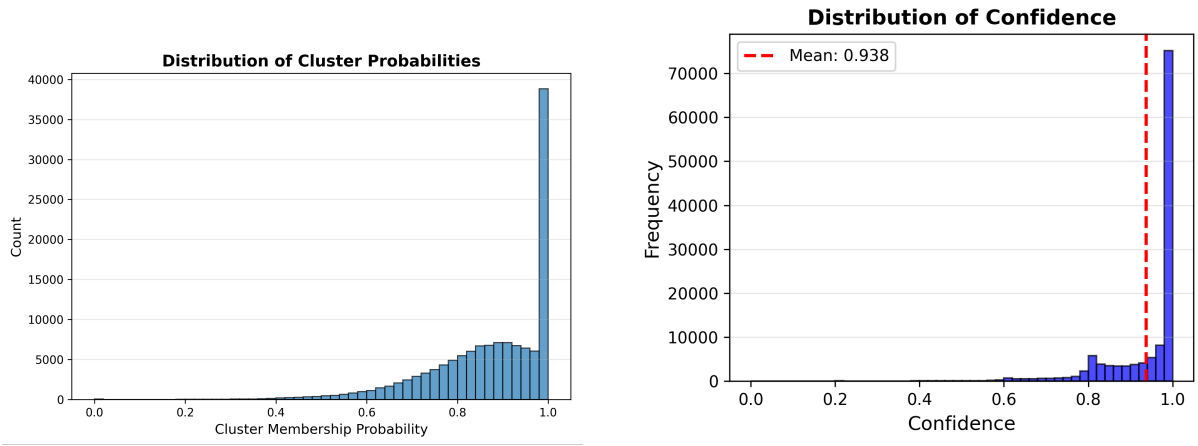

Figure 4: Cluster Probabilities and Cluster Confidence

A Appendix

A.1 [CLS] + Span Pooling

Let $x = (Q, C)$ denote a question–context pair from emrQA-msquad, and let $H(x) = (h_0, h_1, \ldots, h_T) \in R^{(T+1)\times d}$ be the final-layer hidden states of the ELECTRA-small encoder. We use h_0 to denote the contextualized representation of the special `[CLS]` token.

Global sequence summary. In a Transformer encoder, each layer updates the `[CLS]` representation via

$$h_0^{(\ell+1)} = \sum_{t=0}^{T} \alpha_{0t}^{(\ell)} W_V^{(\ell)} h_t^{(\ell)}, \tag{6}$$

where $\alpha_{0t}^{(\ell)}$ are self-attention weights from `[CLS]` to token t. Thus $h_{\text{CLS}} \equiv h_0$ forms a learned, attention-weighted mixture of all token representations, capturing global question–passage semantics. This matches the discriminator architecture described in Clark et al. (2020).

Local evidence summary. Given the gold answer span (s^*, e^*), we compute a pooled embedding

$$h_{\text{span}} = \frac{1}{e^* - s^* + 1} \sum_{t=s^*}^{e^*} h_t, \tag{7}$$

which is a uniform-attention summary over the evidence-bearing token region. This produces an embedding that captures the semantics of the answer span while remaining invariant to token-level permutations.

Joint representation. Concatenating the global and local summaries yields

$$z(x) = [h_{CLS}; h_{span}] \in R^{2d}, \tag{8}$$

which increases representational capacity compared to either component alone. A linear scoring function over $z(x)$ decomposes naturally as

$$w^\top z(x) = w_{\text{CLS}}^\top h_{\text{CLS}} + w_{\text{span}}^\top h_{\text{span}}, \tag{9}$$

separating contributions from global template-level information and localized evidence semantics. This allows clustering algorithms such as HDBSCAN to detect both structural divergences (via h_{CLS}) and span-specific anomalies (via h_{span}).