ELSER Relevance Scores Analysis

This document analyzes ELSER (Elastic Learned Sparse Encoder) relevance scores across different enrichment subtypes and query strategies. Unlike retrieval metrics (Recall, nDCG), relevance scores represent the **model's confidence** in the top-ranked document match.

Overview

What are ELSER Relevance Scores?

- Unbounded positive floats derived from the dot product of sparse vector representations
- Higher scores indicate stronger semantic match between query and document
- Typical range: 0-30, with scores >20 indicating strong matches
- Scores are relative and should be compared within the same index/domain

Query Strategies Analyzed:

- **Last Turn**: Uses only the current question (no conversation context)
- **Rewrite**: Uses LLM-rewritten question with full conversation context
- **Questions**: Uses original self-contained questions (No Agent Response)

Relevance Score Distribution Statistics

Since ELSER relevance scores are relative and unbounded, understanding their distribution across our dataset helps establish meaningful thresholds. The following statistics are calculated from **219 enrichment subtype-strategy combinations** across all domains and enrichment types.

Overall Distribution

- **Mean of all mean scores**: 20.52
- **Median of all mean scores**: 20.74
- **Standard deviation**: 2.17
- **Range**: 13.05 - 26.51

Percentile Breakdown

- **25th percentile**: 19.39 (bottom quartile)
- **50th percentile (median)**: 20.74 (middle)
- **75th percentile**: 22.00 (top quartile)
- **90th percentile**: 22.89 (top decile)

Score Threshold Analysis

- **Scores >20**: Represent **63.0%** of all enrichment subtype-strategy combinations (138/219)
- **Scores >22**: Represent approximately **25%** of combinations (top quartile)
- **Scores <19**: Represent approximately **25%** of combinations (bottom quartile)

Why Scores >20 Indicate Strong Matches

1. **Above median**: Scores above 20 exceed the median (20.74), placing them in the upper half of all observed scores
2. **Above mean**: Scores above 20 exceed the overall mean (20.52), indicating above-average semantic match quality
3. **Majority threshold**: Since 63% of combinations achieve scores >20, this represents the typical performance for well-formatted queries

4. **Statistical significance**: The 75th percentile (22.00) and 90th percentile (22.89) provide natural breakpoints for "very strong" and "exceptional" matches

Interpretation

- **Scores >20**: Strong semantic matches - model is confident in the top result (above median and mean)
- **Scores 15-20**: Moderate matches - some ambiguity or partial relevance (below median but within typical range)
- **Scores <15**: Weak matches - query may be too short, ambiguous, or lacks context (bottom ~10% of observed scores)

Question Types - Mean Relevance Scores

Domain	Subtype	Count	%	Last Turn	Rewrite	Questions	Best Strategy
All	Troubleshooting	15	1.5%	19.72	20.58	**26.19**	Questions
	Non-Question	69	6.9%	17.61	20.45	**22.89**	Questions
	Opinion	80	8.0%	18.29	20.88	**22.70**	Questions
	Keyword	73	7.3%	16.65	20.45	**22.28**	Questions
	Factoid	255	25.4%	19.03	21.37	**22.10**	Questions
	Summarization	192	19.1%	18.33	20.82	**21.88**	Questions
	Explanation	148	14.7%	19.27	21.30	**21.33**	Questions
	Composite	49	4.9%	19.72	21.57	**21.60**	Questions
	Comparative	44	4.4%	19.85	**22.41**	21.75	Rewrite
	How-To	80	8.0%	18.74	**21.46**	20.83	Rewrite
CLAPNQ	Non-Question	12	4.3%	17.57	19.55	**21.90**	Questions
	Opinion	19	6.8%	17.10	20.04	**23.34**	Questions
	Keyword	17	6.1%	16.35	21.48	**22.76**	Questions
	Factoid	101	36.1%	20.12	22.44	**22.62**	Questions
	Summarization	59	21.1%	19.71	21.57	**22.48**	Questions
	Explanation	39	13.9%	20.06	**22.77**	21.53	Rewrite
	Composite	19	6.8%	20.59	**22.92**	21.96	Rewrite
	Comparative	7	2.5%	15.09	**20.98**	19.97	Rewrite
	How-To	7	2.5%	19.08	21.81	**23.00**	Questions
CLOUD	Troubleshooting	13	5.6%	19.41	20.23	**26.51**	Questions
	Non-Question	14	6.0%	18.01	18.97	**24.38**	Questions
	Opinion	8	3.4%	19.40	20.87	**24.30**	Questions
	Keyword	25	10.7%	17.49	20.02	**24.03**	Questions
	Factoid	43	18.4%	18.06	20.56	**22.21**	Questions
	Summarization	37	15.9%	16.79	21.32	**23.00**	Questions
	Explanation	35	15.0%	18.85	19.78	**21.60**	Questions
	Composite	12	5.2%	18.80	20.74	**22.01**	Questions
	Comparative	13	5.6%	17.18	19.24	**20.85**	Questions
	How-To	33	14.2%	17.91	19.90	**19.99**	Questions
FIQA	Non-Question	20	8.0%	19.63	22.49	**22.69**	Questions
	Opinion	42	16.9%	19.98	22.02	**22.38**	Questions
	Keyword	12	4.8%	16.46	21.00	**22.37**	Questions
	Factoid	41	16.5%	20.20	**22.94**	21.84	Rewrite
	Summarization	43	17.3%	19.34	21.11	**21.79**	Questions
	Explanation	44	17.7%	20.56	**22.66**	21.74	Rewrite

Domain	Subtype	Count	%	Last Turn	Rewrite	Questions	Best Strategy
	Composite	6	2.4%	20.13	21.51	**21.73**	Questions
	Comparative	22	8.8%	22.89	**24.71**	23.09	Rewrite
	How-To	19	7.6%	20.64	**22.88**	21.05	Rewrite
GOVT	Troubleshooting	2	0.8%	21.75	22.89	**24.15**	Questions
	Non-Question	23	9.5%	15.64	20.04	**22.66**	Questions
	Opinion	11	4.5%	13.05	18.03	**21.63**	Questions
	Keyword	19	7.8%	15.95	**19.73**	19.51	Rewrite
	Factoid	70	28.8%	17.35	19.42	**21.42**	Questions
	Summarization	53	21.8%	17.04	19.39	**20.51**	Questions
	Explanation	30	12.3%	16.85	19.16	**20.18**	Questions
	Composite	12	4.9%	19.08	20.29	**20.55**	Questions
	Comparative	2	0.8%	20.44	**22.62**	19.07	Rewrite
	How-To	21	8.6%	18.19	**22.51**	21.21	Rewrite

Total: 1005 tasks (tasks can have multiple question types)

Question Types - Relevance Score Patterns

Troubleshooting (15 queries)

Highest confidence overall with Questions strategy (26.19). Extremely strong semantic matches. Questions strategy significantly outperforms (+33% vs Last Turn), suggesting self-contained question format works exceptionally well for problem-solving queries.

Non-Question (69 queries)

Strong confidence with Questions strategy (22.89). Shows substantial improvement from Last Turn (17.61) to Questions (+30%), indicating that converting implicit statements into explicit questions dramatically improves match quality.

Opinion (80 queries)

High confidence with Questions strategy (22.70). Consistent pattern across domains. Benefits significantly from question formatting (+24% vs Last Turn), suggesting opinion queries need explicit framing for best semantic matching.

Keyword (73 queries)

Good confidence with Questions strategy (22.28). Despite **poor retrieval performance** (worst R@5), shows **high relevance scores**, indicating ELSER is confident but often wrong—a critical disconnect between confidence and accuracy.

Factoid (255 queries)

Solid confidence with Questions strategy (22.10). Largest category. Moderate improvement with rewriting (+12% vs Last Turn), suggesting factual queries benefit from context but not as dramatically as other types.

Summarization (192 queries)

Good confidence with Questions strategy (21.88). Second-largest category. Consistent ~20% improvement from Last Turn to Questions, indicating summary requests need explicit question framing.

Explanation (148 queries)

Moderate-to-good confidence with Questions strategy (21.33). Smallest improvement gap between Rewrite (21.30) and Questions (21.33), suggesting explanatory queries are relatively robust to formatting.

Composite (49 queries)

Good confidence with Questions strategy (21.60). Very close scores between Rewrite (21.57) and Questions, indicating multi-part questions maintain quality across strategies.

Comparative (44 queries)

Only category where Rewrite wins (22.41 vs Questions 21.75). Unique pattern suggests comparative questions benefit more from context integration than explicit question formatting. Aligns with retrieval performance where Rewrite also excelled.

How-To (80 queries)

Second category where Rewrite wins (21.46 vs Questions 20.83). Procedural questions benefit from contextual rewriting, likely because steps and procedures require conversation continuity.

Multi-Turn Types - Mean Relevance Scores

Domain	Subtype	Count	%	Last Turn	Rewrite	Questions	Best Strategy
All	Follow-up	574	73.9%	18.47	21.06	**21.67**	Questions
	N/A	102	13.1%	21.22	21.22	**21.22**	Questions
	Clarification	101	13.0%	17.28	21.29	**23.15**	Questions
CLAPNQ	Follow-up	154	74.0%	19.16	21.74	**22.16**	Questions
	N/A	28	13.5%	21.94	21.94	**21.94**	Questions
	Clarification	26	12.5%	19.49	**23.18**	22.97	Rewrite
CLOUD	Follow-up	135	71.8%	17.89	20.19	**22.05**	Questions
	N/A	25	13.3%	**19.80**	19.80	19.80	Last Turn
	Clarification	28	14.9%	16.72	20.59	**25.22**	Questions
FIQA	Follow-up	129	71.7%	19.92	**22.40**	21.82	Rewrite
	N/A	24	13.3%	**23.67**	23.67	23.67	Last Turn
	Clarification	27	15.0%	18.64	**22.32**	21.77	Rewrite
GOVT	Follow-up	156	77.6%	17.11	20.04	**20.74**	Questions
	N/A	25	12.4%	**19.48**	19.48	19.48	Last Turn
	Clarification	20	9.9%	13.37	18.46	**22.36**	Questions

Total: 777 tasks

Multi-Turn Types - Relevance Score Patterns

Follow-up (574 queries)

Largest multi-turn category. Moderate confidence with Questions strategy (21.67). Shows consistent ~17% improvement from Last Turn to Questions. Context clearly helps, but questions format provides additional boost.

N/A (102 queries)

Good confidence, **identical across all strategies** (21.22). These single-turn queries within multi-turn conversations don't benefit from any strategy—they're self-contained and context-independent.

Clarification (101 queries)

Highest confidence with Questions strategy (23.15). Shows **dramatic improvement** from Last Turn (17.28) to Questions (+34%), the largest gain of any subtype. Clarification questions desperately need explicit formatting and context to achieve good semantic matches.

Answerability Types - Mean Relevance Scores

Domain	Subtype	Count	%	Last Turn	Rewrite	Questions	Best Strategy
All	ANSWERABLE	709	91.2%	18.90	21.27	**21.85**	Questions
	PARTIAL	68	8.8%	16.39	19.51	**21.30**	Questions
CLAPNQ	ANSWERABLE	192	92.3%	19.85	22.08	**22.25**	Questions
	PARTIAL	16	7.7%	16.20	20.36	**21.97**	Questions
CLOUD	ANSWERABLE	177	94.2%	17.97	20.29	**22.32**	Questions
	PARTIAL	11	5.8%	17.94	18.68	**20.74**	Questions
FIQA	ANSWERABLE	163	90.6%	20.51	**22.67**	22.00	Rewrite
	PARTIAL	17	9.4%	17.57	21.48	**22.64**	Questions
GOVT	ANSWERABLE	177	88.1%	17.31	20.07	**20.82**	Questions
	PARTIAL	24	11.9%	14.98	17.93	**20.16**	Questions

Total: 777 tasks

Answerability Types - Relevance Score Patterns

Answerable (709 queries)

Strong baseline confidence with Questions strategy (21.85). Represents the "standard" query quality. Moderate improvement (~16%) from Last Turn to Questions.

Partial (68 queries)

Lower confidence overall with Questions strategy (21.30). Shows **larger improvement gap** from Last Turn (16.39) to Questions (+30% vs 16% for Answerable). Partial queries struggle more with context but benefit dramatically from proper formatting.

Insight 1: Query Type as a Strong Confidence Indicator

Analysis reveals that query type is the most reliable predictor of which strategy performs best.

Query Type	Best Strategy	Confidence	Key Observation
Troubleshooting	Last Turn	19.72	Specific error codes/logs benefit from exact matching; Rewrite may dilute critical tokens
Keyword	Questions	22.28	Sparse queries are overconfident but inaccurate; context expansion is essential
Comparative	Rewrite	22.41	Comparison framing requires explicit context integration
How-To	Rewrite	21.46	Procedural queries depend on conversational continuity
Factoid	Questions	22.10	Explicit question format yields clearest semantic signal
Explanation	Questions	21.33	Direct framing improves retrieval accuracy
Summarization	Questions	21.88	Explicit summary requests retrieve better
Opinion	Questions	22.70	Subjective queries need clear framing
Composite	Questions	21.60	Multi-part queries are already self-contained
Non-Question	Questions	22.89	Implicit statements need explicit conversion

---

Insight 2: Multi-Turn Context Modifies Retrieval Behavior

Conversation position affects how much reformulation helps.

Multi-Turn Type	Best Strategy	Confidence Gain vs Last Turn	Key Observation
N/A (First Turn)	Any	0%	Self-contained; strategy is irrelevant
Follow-up	Questions	+17%	Benefits from context-enriched explicit format
Clarification	Questions	+34%	Highest sensitivity to reformulation; largest improvement but still weakest retrieval

---

Insight 3: Domain Shifts Strategy Preferences

Different domains favor different strategies, suggesting domain-specific optimization is valuable.

Domain	Strategy Preference	Key Characteristics
CLAPNQ	Balanced	**Rewrite** preferred for complex types (Explanation, Composite, Comparative, Clarification), **Questions** preferred for simpler types
CLOUD	Strong Questions	**Questions** strategy achieves highest scores across all types, technical documentation aligns well with explicit question formats
FIQA	Rewrite-heavy	**Rewrite** wins for most types (Factoid, Explanation, Comparative, How-To, Follow-up, Clarification), financial context benefits from LLM rewriting
GOVT	Questions (with caution)	**Questions** preferred but lowest confidence scores overall (15–20 range), policy language may be ambiguous or domain-mismatched

Pattern: FIQA is the most Rewrite-preferring domain. CLOUD is the most Questions-preferring. GOVT shows weak confidence across the board, indicating potential domain mismatch.

Insight 4: Confidence-Performance Alignment Varies

Relevance scores are not uniformly reliable predictors of retrieval accuracy.

What is Alignment?

Alignment measures how well ELSER's confidence (relevance score) correlates with actual retrieval performance (R@5, R@10).

How is Alignment Calculated?

- **Aligned**: Confidence score directionally matches retrieval performance. Higher confidence corresponds to higher or similar R@5/R@10 compared to other strategies for the same query type.
- **Not aligned**: Confidence score contradicts retrieval performance. Higher confidence but lower R@5/R@10, or confidence increases while performance decreases.
- **Mix**: Mixed signals—confidence and performance move in the same direction but with notable discrepancies.

For example, if Rewrite has higher confidence than Last Turn but lower R@5, that indicates overconfidence (not aligned). If higher confidence corresponds to higher R@5, that's aligned.

Query Type	Strategy	Confidence	R@5	R@10	Alignment
Troubleshooting	Last Turn	19.72	0.706	0.706	Aligned
	Rewrite	20.58	0.611	0.739	Aligned
	Questions	**26.19**	**0.333**	**0.467**	**Not aligned**
Non-Question	Last Turn	17.61	0.386	0.466	Aligned
	Rewrite	20.45	0.425	0.548	Aligned
	Questions	**22.89**	**0.210**	**0.289**	**Not aligned**
Opinion	Last Turn	18.29	0.408	0.486	Aligned
	Rewrite	20.88	0.462	0.578	Aligned
	Questions	**22.70**	**0.253**	**0.312**	**Not aligned**
Keyword	Last Turn	16.65	0.432	0.545	Aligned
	Rewrite	**20.45**	**0.395**	**0.546**	**Mixed**
	Questions	**22.28**	**0.161**	**0.206**	**Not aligned**
Factoid	Last Turn	19.03	0.414	0.515	Aligned
	Rewrite	21.37	0.465	0.599	Aligned
	Questions	**22.10**	**0.280**	**0.369**	**Not aligned**
Summarization	Last Turn	18.33	0.416	0.530	Aligned
	Rewrite	20.82	0.455	0.587	Aligned
	Questions	**21.88**	**0.259**	**0.325**	**Not aligned**
Explanation	Last Turn	19.27	0.451	0.575	Aligned
	Rewrite	21.30	0.492	0.616	Aligned
	Questions	**21.33**	**0.236**	**0.341**	**Not aligned**
Composite	Last Turn	19.72	0.505	0.646	Aligned
	Rewrite	21.57	0.573	0.750	Aligned
	Questions	**21.60**	**0.314**	**0.458**	**Not aligned**
Comparative	Last Turn	19.85	0.461	0.559	Aligned
	Rewrite	22.41	0.527	0.626	Aligned
	Questions	**21.75**	**0.340**	**0.435**	**Not aligned**
How-To	Last Turn	18.74	0.442	0.567	Aligned
	Rewrite	21.46	0.468	0.619	Aligned
	Questions	**20.83**	**0.311**	**0.394**	**Not aligned**

Pattern:

- **Last Turn and Rewrite strategies** generally show well-calibrated confidence-performance alignment across most query types
- **Questions strategy** exhibits overconfidence—high relevance scores (20-26) but lower retrieval performance (R@5: 0.16-0.34), indicating the model is confident but frequently wrong when using self-contained question formats
- **Keyword queries** show the "confidence paradox" most clearly in Questions strategy (22.28 confidence vs 0.161 R@5)
- **Troubleshooting** with Last Turn shows exceptional alignment (19.72 confidence, 0.706 R@5), suggesting exact matching works best for error codes/logs

---

Part 2: Building an Oracle Routing Strategy

Based on the insights above, an oracle routing strategy should leverage query type, multi-turn context, domain, and confidence calibration to select the optimal strategy for each query.

Routing Features

Feature	Detection Method	Routing Impact
Query Type	Classifier (10 classes)	Primary strategy selector
Multi-Turn Type	Position detection + intent classifier	Modifies strategy; flags high-risk queries
Domain	Known at query time	Applies domain-specific overrides
Token Count	Simple count	Flags sparse queries for forced Rewrite
Confidence Score	ELSER output	Triggers fallback when misaligned
Error/Log Patterns	Regex or pattern matching	Routes Troubleshooting to Last Turn

---

Data Source

Analysis generated by `scripts/discovery/analyze_relevance_scores.py` using:

- Task enrichments from `cleaned_data/tasks/`
- ELSER retrieval results from `scripts/baselines/retrieval_scripts/elser/results/`
- Detailed statistics available in `scripts/discovery/enrichment_analysis_results/relevance_scores_*.csv`
