

When Reasoning Revises Relevance: Analysing Score Trajectories in LLM Reasoning Traces

Louis Geiger

louis.geiger@email.uni-freiburg.de
University of Freiburg
Freiburg, Germany

Falk Scholer

falk.scholer@rmit.edu.au
RMIT University
Melbourne, Australia

Danula Hettiachchi

danula.hettiachchi@rmit.edu.au
RMIT University
Melbourne, Australia

Johanne R. Trippas

j.trippas@rmit.edu.au
RMIT University
Melbourne, Australia

Abstract

Large language models can generate reasoning traces when assigning relevance labels, but aggregate effectiveness measures do not reveal whether reasoning stabilises, corrects, or degrades individual judgements. Prior work shows that greater reasoning effort does not reliably improve agreement with human assessors, and small gains in binary accuracy may coincide with lower graded agreement, higher computational costs, and additional failure modes. In this work, we examine reasoning traces produced during graded relevance assessment and investigate whether trace length varies across interrogative queries (i.e., queries starting with *what*, *how*, *why*, *who*, *where*, and *when*). Our results show that the median reasoning trace length varies substantially across query categories, suggesting that the interrogative form of a question may indicate the level of reasoning effort a model will deploy. We then examine how relevance scores change as reasoning progresses, and show that the trajectories can contain multiple candidate relevance scores before the model determines a final judgement. These score sequences differ across models and reasoning configurations, showing a mix of consistent directional changes as well as reversals. On aggregate, across three model configurations and two datasets, the initial relevance score assigned by a model is more likely to be correct than the final assigned score after reasoning. However, a more detailed analysis reveals that these trends are not uniform across relevance levels: reasoning acts as a directional, label-dependent transformation rather than a uniformly beneficial or harmful refinement. Overall, this work demonstrates that final labels can obscure considerable variation in the observable reasoning process and highlight the need to evaluate reasoning behaviour separately from judgement accuracy to identify when additional reasoning is able to correct a judgement and when it removes an initially correct decision.

CCS Concepts

• **Information systems** → **Language models; Users and interactive retrieval.**

Keywords

Relevance judgements, Information Retrieval

1 Introduction and Related Work

Large language models (LLMs) offer a low-cost alternative or complement to human relevance assessment in information retrieval (IR). Given a query and a document or passage, an LLM can assign a binary or graded relevance label [2, 3, 15, 17]. However, these judgements are sensitive to prompt design, model choice, output constraints, and the requested level of reasoning effort [10]. Such sensitivity matters because labelling errors can affect test collections and alter conclusions about the effectiveness of retrieval systems. Recent work also identifies systematic overrating in LLM relevance assessment, particularly for non-relevant passages and graded labels [1, 20]. The representation of the information need also matters, as formalised topic descriptions can reduce positivity bias and improve agreement relative to query-only prompts [12]. However, these aspects have principally been studied by examining final labels.

One proposed way to improve LLM judgements is to ask the model to *reason* before assigning a label [4, 5, 7, 10]. Reasoning may help identify the information need, compare it with the passage, and distinguish partial from complete relevance. However, recent research found that greater reasoning effort did not consistently improve agreement with human assessors [10, 14]. Binary accuracy increased slightly in some settings, but graded agreement declined, token use increased by up to 15 times, and output-format failures became more frequent. They also observed cases of overthinking, in which a plausible initial assessment led to an incorrect final label.

Aggregate effectiveness measures indicate whether a reasoning configuration improves or harms overall relevance assessment, but they do not explain how the final label is reached. Two systems with similar final accuracy may exhibit different behaviours: one may produce a stable assessment, while the other may repeatedly revise its score or deviate from an initially correct label. Reasoning trace length is similarly ambiguous, because a longer trace may reflect useful comparison, unresolved ambiguity, repetition, or an ineffective revision. Examining the trace as an observable sequence is therefore necessary to distinguish the amount of generated reasoning from the behaviour and outcome of that reasoning.

In this paper, we analyse reasoning traces from a prior evaluation of graded LLM relevance judgements [10], across three model configurations and two test collections. First, we measure how trace length varies across individual queries and across queries beginning with *what*, *how*, *why*, *who*, *where*, and *when*. Second, we

extract sequences of explicitly expressed relevance scores, which we call *trajectories*, and characterise their direction, magnitude, and reversals. Finally, we compare the first expressed score and final trajectory score with human judgements, both overall and by model configuration and human relevance grade. This analysis contributes a trace-level account of when reasoning changes a relevance judgement. In particular, it separates variation in reasoning effort from revision behaviour and identifies which relevance grades account for improvements or losses in final accuracy.

2 Methodology

To study LLM reasoning traces when making relevance judgements, we analysed data from a prior evaluation of LLM relevance judgements made publicly available by Geiger et al. [10]. This covers two TREC collections (DL23 [6] and DL-HARD [13]) and two LLMs (*gpt-oss-120b* and *Gemini 2.5 Flash*). The evaluation varied the model and reasoning level while keeping the query, passage, relevance scale, and output instructions fixed across conditions. For each query-passage pair, the model generated a reasoning trace and a graded relevance label. Human labels from the benchmark collections provided the reference judgements.

For our analysis, we treated the reasoning traces (the text output produced by a model to explain its reasoning) as generated text, rather than as direct evidence of the models’ internal decision processes. While the experimental settings include four combinations (two LLMs, each at two reasoning levels), the data generated by *gemini-2.5-flash/0* (i.e., the no reasoning setting for that model) does not produce any reasoning traces by design. Therefore, the analysis in this paper focuses on three model configurations: *gemini-2.5-flash/500* (high reasoning setting), *gpt-oss-120b/low* (low reasoning setting), and *gpt-oss-120b/high* (high reasoning setting).

Unless otherwise noted, statistical significance is tested using the χ^2 test based on underlying response counts, with a standard threshold of $p < 0.05$ to determine significance.

2.1 Interrogative Query Classification and Trace Length

To study the impact of interrogative query type, we classified queries according to their first interrogative word: *what*, *how*, *why*, *who*, *where*, or *when* [16, 18, 19]. Before classification, we normalised case and removed leading punctuation and whitespace. For ambiguous or compound queries, we applied the classification rule solely to the first interrogative word. Queries that did not begin with one of these six terms were assigned to an *other* category. This approach provides a simple and reproducible classification scheme, without assuming that the interrogative type fully captures the underlying information need or query difficulty.

Reasoning trace length is measured by counting the characters in each generated trace, excluding the final answer when it could be reliably separated. We compared reasoning length across interrogative types, individual queries, models, and reasoning configurations.

2.2 Relevance Trajectory Measures

To construct *relevance trajectories*, we used pattern-matching rules to extract explicit relevance scores from each reasoning trace as they occur. The rules identified numeric scores and unambiguous textual

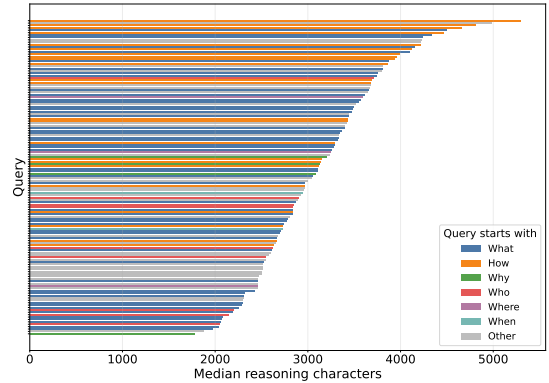


Figure 1: Median reasoning trajectory length per query, sorted by median length.

equivalents on the original ordinal relevance scale. We extracted the final relevance label separately.

We excluded numbers referring to passage positions, quantities, dates, probabilities, and other non-relevance information as well as negations (i.e. “not a score of 2”). To assess extraction quality, we manually inspected 100 randomly sampled traces, and found 5 false positives as well as 5 false negatives (*Precision* = 0.615, *Recall* = 0.615). We refined the rules further on the basis of this inspection before conducting the reported analysis. The trajectory analysis includes only traces containing at least two valid and distinct expressed scores. It is therefore a conditional analysis of score-revising traces, not an estimate of behaviour across all generated judgements. Comparisons between the first expressed score and final output label include only cases in which both quantities and the human label are available.

3 Results

3.1 Reasoning Lengths Per Interrogative Word Query Type

We group reasoning traces by query and compute the median trace length (in characters) for each query. Figure 1 shows the resulting distribution for *gpt-oss-120b* under the high reasoning-effort setting. Per-query medians span approximately 1,750 to 5,100 characters, a near threefold range within a single model and budget setting. Because the requested reasoning setting is fixed, this variation shows that the configured budget does not determine the amount of reasoning produced; the query also affects how much of that budget is used.

Table 1 shows the median length of reasoning traces, grouped by leading interrogative words. Queries starting with “how” account for four of the five highest per-query medians, while “who” queries appear almost entirely in the lower third of the distribution. In addition, “what” queries form the largest category and occur throughout the distribution, with many in the upper-middle range. The median lengths differ significantly across the groups (Kruskal-Wallis test, $p < 0.0001$), indicating that the LLM dedicates substantially different levels of reasoning effort depending on the type of question.

Table 1: Median reasoning trace length in characters grouped by interrogative words.

Query starts with	Queries	Samples	Median
"how"	23	4,409	3,680.00
"where"	3	398	3,249.50
"why"	4	726	3,108.50
"what"	53	10,170	3,103.50
other	38	6105	2,906.75
"when"	2	393	2,834.75
"who"	9	928	2,620.50

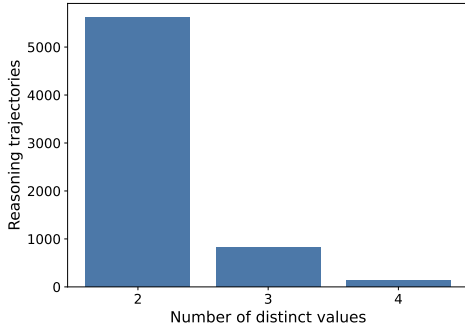


Figure 2: Number of distinct relevance levels per reasoning trajectory.

3.2 Relevance Changes Through Reasoning

We extract all candidate relevance scores expressed during a trace and retain traces containing at least two distinct values, yielding 6,600 score-revising traces. This restriction focuses the analysis on cases in which the model expresses a changed assessment; the reported proportions do not describe traces that retain a single score throughout. Among the selected traces, approximately 85% contain two distinct values, 13% contain three, and 2% contain all four relevance grades (Figure 2). Most score-revising traces therefore consider one alternative grade, although a smaller subset ranges over three or four grades.

Within the 6,600 score-revising trajectories, the first expressed score agrees with the human label in 40.2% of cases, whereas the final output label agrees in 26.4% (Figure 3), a decrease of 13.8 percentage points. This difference is statistically significant (χ^2 , $p < 0.001$). We study stratification by model configuration and relevance levels further below.

Figure 4 locates the source of this decline. In 1,930 traces, the first expressed score is correct, but the final output label is incorrect; in 1,022 traces, revision corrects an initially incorrect score. A further 723 traces are correct at both endpoints, and 2,925 are incorrect at both endpoints. Thus, when revision changes correctness, deterioration is 1.89 times as frequent as correction. Reasoning does not account for every error in this subset, but its net effect is to remove more correct scores than it recovers.

Directionality of Score Changes. Figure 5 shows the difference between the final output label and the first expressed score. Changes

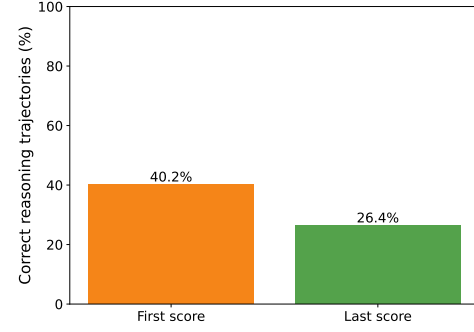


Figure 3: Percentage accuracy for the first versus the last relevance score in a reasoning trajectory.

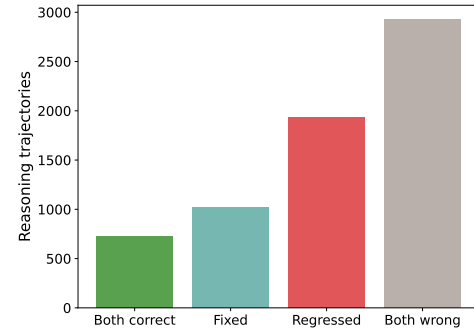


Figure 4: Comparing correctness across reasoning trajectories based on the first versus final relevance level assigned. “Fixed” indicates a change from an incorrect initial relevance assignment to a correct final relevance level, and “Regressed” the reverse (correct first assignment to incorrect final assignment).

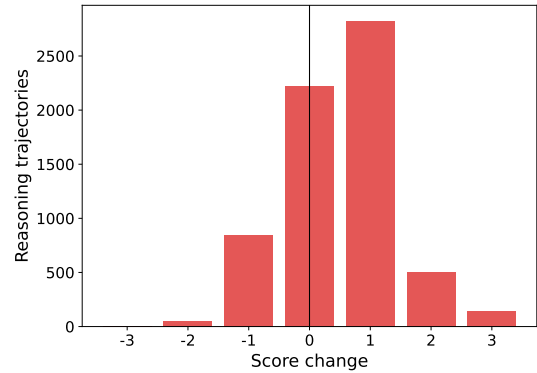


Figure 5: Frequency distribution of the difference between final and first assigned relevance scores. A positive change indicates an increased final score and a negative change indicates a decreased final score. Intermediate scores are not taken into account.

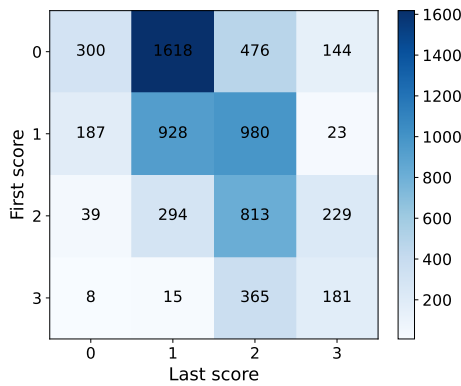


Figure 6: Distribution of changes between the first and final assigned relevance levels in reasoning trajectories.

are strongly skewed upwards: 3,470 of the 6,600 traces end at a higher relevance grade, compared with 908 that end at a lower grade. A one-grade increase is the most common outcome, occurring in 2,827 traces (42.8%). A zero net change is the second most common outcome, occurring in 2,222 traces (33.7%), while changes of two or three grades are uncommon. Among traces with different first and final values, 80% change by one grade.

A zero net change does not indicate that the expressed score remained constant. Every included trace contains at least two distinct scores, so these cases move away from the first score and later return to it. Figure 5 therefore describes endpoint differences, we next analyse the shape of the reasoning trajectories across the full sequence.

These changes are decomposed by relevance level in Figure 6, which shows that the upward shift is not a uniform movement towards maximum relevance. Instead, revisions concentrate the final labels on intermediate relevance levels. Across the lower boundary, 1,618 trajectories move from level 0 to level 1, compared with 187 in the opposite direction. Similarly, 980 move from level 1 to level 2, compared with 294 from level 2 to level 1. At the upper boundary, however, this direction reverses: 365 trajectories move from level 3 to level 2, while 229 move from level 2 to level 3. Levels 1 and 2 therefore account for 83.2% of final labels in the selected trajectories, compared with 52.9% of first expressed scores. Revision therefore appears to contract scores towards intermediate positive relevance, rather than simply increasing them.

Figure 7 compares the shape, in terms of direction and stability, of score sequences across model configurations, grouping relevance trajectories into: monotonically increasing, monotonically decreasing, 1 reversal, 2 reversals, or more than 2 reversals. For *gemini-2.5-flash/500* and *gpt-oss-120b/low*, monotonic increases account for approximately 55.6% and 51.8% of traces, while monotonic decreases account for approximately 37.6% and 32.6%. Reversals occur in approximately 6.8% and 15.6% of traces, respectively. The pattern differs for *gpt-oss-120b/high*, where more than 55% of score-revising traces are non-monotonic, and 13% contain more than two reversals. The expressed assessments under the high reasoning configuration for GPT OSS 120b are therefore substantially less stable than those

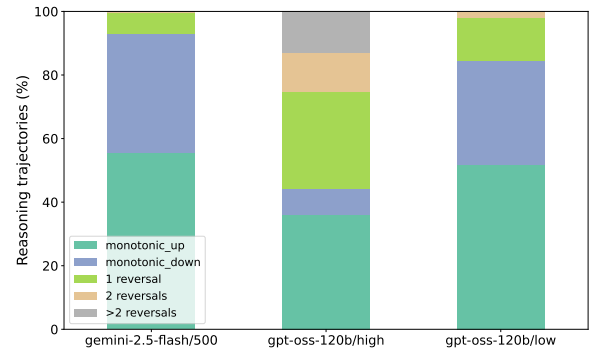


Figure 7: Direction of score changes grouped by model and reasoning level.

in the low reasoning configuration for the same model, as well as for the high reasoning configuration of Gemini 2.5 Flash. Note that these reversals describe variation in the generated relevance trace, i.e. the textual output produced by the model to explain its reasoning. Therefore, they should not be treated as direct observations of internal model uncertainty.

First Score Versus Last Score Variations. We now focus on the start and end of relevance trajectories, since these are the key points that represent an initial relevance assessment, and the final level assigned by the model. First-score accuracy varies substantially across configurations, whereas final-label accuracy converges (see Figure 8). For *gemini-2.5-flash/500*, accuracy falls from 34.6% to 27.3%. The largest decline occurs for *gpt-oss-120b/high*, from 40.7% to 26.4%, while *gpt-oss-120b/low* increases from 25.5% to 27.7%. First-score accuracy therefore spans 15.2 percentage points across the three configurations, compared with 1.3 points for final-label accuracy. In particular, the high-reasoning configuration begins with the most accurate expressed scores but loses this advantage during revision. The difference between *gpt-oss-120b/high* and *gpt-oss-120b/low* is statistically significant (χ^2 , $p < 0.05$), the other pairwise differences are not.

Table 2 shows that the aggregate decline is not uniform across human relevance levels. For *gpt-oss-120b/high*, level 0 passages account for 3,298 of 6,254 selected trajectories, and their accuracy falls from 49.3% for the first expressed score to 13.2% for the final output label, a decrease of 36.1 percentage points. In contrast, level 2 accuracy rises from 32.0% to 58.4%, an increase of 26.4 points; grades 1 and 3 show smaller gains. This grade-specific pattern is consistent with the upward skew in Figure 5. Upward revision can correct an underestimated relevant passage, but it also converts many correct non-relevant assessments into false positives. The effect of reasoning on aggregate accuracy therefore depends not only on the model configuration, but also on the prevalence of different relevance grades.

The Gemini and low-reasoning samples also lose accuracy on level 0 passages, although their per-grade counts are small and the remaining changes are mixed. These results motivate a selective strategy in which an early score is retained for clear non-relevant cases, and additional reasoning is reserved for borderline or potentially relevant passages. Evaluating such a policy requires a

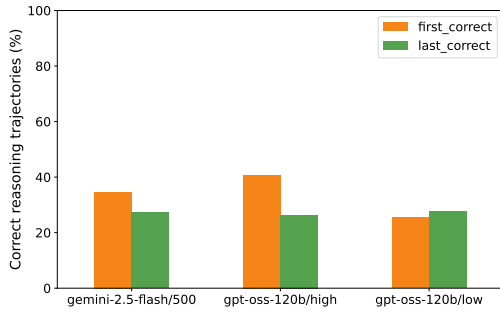


Figure 8: Percentage accuracy for the first versus the last relevance score in a reasoning trajectory, grouped by model and reasoning level.

separate experiment because the present analysis is observational and conditioned on traces that already contain score revisions.

4 Discussion and Conclusions

This study conducted a detailed examination of LLM reasoning in relevance assessment, identifying patterns and behaviours obscured by aggregate judgement accuracy analyses in previous work. Analysis of reasoning trace length showed substantial and significant variation across interrogative query categories, even when the model and reasoning settings were fixed, suggesting that this could be a useful indicator of the model’s reasoning effort. More importantly, the sequences of relevance levels expressed within traces, have systematic trajectories. Among 6,600 traces containing at least two distinct scores, upward revisions are almost four times as frequent as downward revisions, and regressions from a correct to an incorrect relevance level are nearly twice as frequent as corrections. Consequently, the final output label is less accurate than the first expressed score in this selected subset.

More detailed trace-level analysis identified how that loss arises. For *gpt-oss-120b/high*, upward revision substantially improves accuracy for human level 2 passages, while sharply reducing accuracy for the more frequent level 0 passages. Reasoning, therefore, acts as a level-dependent transformation rather than a uniform refinement of an initial assessment. This distinction matters for relevance evaluation because an aggregate gain or loss can change with the grade distribution, even when the underlying revision behaviour remains the same. Reporting additional diagnostic information, such as trajectory-level analysis, also aligns with recommendations in previous work that high agreement among LLM judges is not sufficient evidence of validity [8, 9].

These findings suggest two practical changes to the use of reasoning-based LLM relevance judges. First, the evaluation should report relevance trajectory measures alongside final-level effectiveness, including first-to-final changes, revision direction, reversals, and accuracy by human relevance grade.

Second, additional reasoning should be treated as a decision whose value may depend on the current assessment, rather than as a fixed requirement for every query-passage pair. A selective procedure might leverage a non-reasoning approach, such as a simple LLM judge or a topic-specific classifier [11], to first identify

clearly non-relevant passages, and invoke further comparison with reasoning-enabled judges for a more fine-grained assessment of relevant cases. The present results motivate this design, but testing whether an early score can be reliably identified or whether selective reasoning improves end-to-end assessment requires further experiments and may be useful for future work.

The analysis has several limitations. It uses two English passage-level collections, a single prompting setup, and three model-reasoning configurations, two of which contribute relatively few score-revising traces. The reasoning traces are generated explanations, and need not faithfully represent internal computation. Score extraction relies on pattern matching, and some false positives may remain despite manual inspection. Most importantly, the trajectory results are conditional on traces containing at least two distinct expressed scores; they do not show that reasoning reduces accuracy across all judgements, nor do they provide a causal comparison with a no-reasoning condition. The first score expressed might also not actually represent the model’s initial guess. Broader experiments should validate extraction on held-out data, report the frequency and performance of single-score traces, and test stopping or selective-reasoning policies directly. The immediate implication is nevertheless clear: assessing only the final label conceals whether reasoning corrected a relevance judgement, or displaced an answer that was already right.

References

- [1] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2026. On the Use of LLMs for Relevance Labelling. *ACM Trans. Inf. Syst.* 44, 4 (2026), 31 pages.
- [2] Negar Arabzadeh and Charles LA Clarke. 2025. Benchmarking LLM-based relevance judgment methods. In *Proc. ACM SIGIR*.
- [3] Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation. In *Proc. ACM SIGIR*.
- [4] Minsuk Chang, Arjun Srinivasan, and Srishti Palani. 2026. Criticality: Scaffolding Decision-Making with Interactive Critical Thinking and Evidence-Based Reasoning Traces. In *Proc. Int. Conf. Intelligent User Interfaces*. 783–803.
- [5] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. Reasoning Models Don’t Always Say What They Think. *arXiv preprint arXiv:2505.05410* (2025).
- [6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2025. Overview of the TREC 2023 deep learning track. *arXiv preprint arXiv:2507.08890* (2025).
- [7] Mouly Dewan, Jiqun Liu, and Chirag Shah. 2026. TRUE: A reproducible framework for LLM-driven relevance judgment in information retrieval. In *Proc. ACM WSDM*. 1130–1134.
- [8] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)* (Padua, Italy). 218–229.
- [9] Maik Fröbe, Andrew Parry, Ferdinand Schlatt, Sean MacAvaney, Benno Stein, Martin Potthast, and Matthias Hagen. 2025. Large Language Model Relevance Assessors Agree With One Another More Than With Human Assessors. In *Proc. ACM SIGIR*. 2858–2863.
- [10] Louis Geiger, Danula Hettiachchi, Falk Scholer, and Johanne R. Trippas. 2026. Reasoning with Large Language Models for Relevance Judgements. In *Proc. ACM ICTIR*. ACM, 105–110. doi:10.1145/3805713.3820403
- [11] Lukas Gienapp, Martin Potthast, Andrew Yates, Harrison Scells, and Eugene Yang. 2026. Topic-Specific Classifiers are Better Relevance Judges than Prompted LLMs. In *Proc. ACM SIGIR* (Australia). 415–425.
- [12] Juri Keller, Maik Fröbe, Björn Engelmann, Fabian Haak, Timo Breuer, Birger Larsen, and Philipp Schaer. 2026. Formalized Information Needs Improve Large-Language-Model Relevance Judgments. In *Proc. ACM SIGIR*. 778–789.
- [13] Iain Mackie, Jeffrey Dalton, and Andrew Yates. 2021. How deep is your learning: The DL-HARD annotated deep learning dataset. In *Proc. ACM SIGIR*. 2335–2341.
- [14] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehraj Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer,

Table 2: Accuracy of first and last score in reasoning trajectories, by model configuration and human relevance grade. Change is measured in percentage points.

Configuration	Human grade	Trajectories	First accuracy	Last accuracy	Change
GPT-OSS/high	0	3,298	49.3%	13.2%	−36.1
	1	1,443	35.8%	39.9%	+4.1
	2	851	32.0%	58.4%	+26.4
	3	662	19.9%	21.5%	+1.5
GPT-OSS/high overall		6,254	40.7%	26.4%	−14.3
Gemini/500	0	96	29.2%	18.8%	−10.4
	1	53	49.1%	35.8%	−13.2
	2	34	38.2%	38.2%	0.0
	3	22	18.2%	27.3%	+9.1
Gemini/500 overall		205	34.6%	27.3%	−7.3
GPT-OSS/low	0	53	24.5%	15.1%	−9.4
	1	30	16.7%	30.0%	+13.3
	2	31	45.2%	48.4%	+3.2
	3	27	14.8%	25.9%	+11.1
GPT-OSS/low overall		141	25.5%	27.7%	+2.1

- Xing Han Lù, et al. 2025. DeepSeek-R1 Thoughtology: Let’s think about LLM Reasoning. *arXiv preprint arXiv:2504.07128* (2025).
- [15] Julian A Schnabel, Johanne R Trippas, Falk Scholer, and Danula Hettiachchi. 2025. Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment. In *Companion Proc. ACM WebConf*.
- [16] Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, et al. 2015. Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval. In *FIRE workshops*, Vol. 1587, 19–25.
- [17] Rikiya Takehi, Ellen M. Voorhees, Tetsuya Sakai, and Ian Soboroff. 2025. LLM-Assisted Relevance Assessments: When Should We Ask LLMs for Help?. In *Proc. ACM SIGIR*.
- [18] Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. 2024. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 2703–2707. doi:10.1145/3626772.3657914
- [19] Johanne R. Trippas, Luke Gallagher, and Joel Mackenzie. 2024. Re-evaluating the Command-and-Control Paradigm in Conversational Search Interactions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM ’24)*. ACM, 2260–2270. doi:10.1145/3627673.3679588
- [20] Chuting Yu, Hang Li, Guido Zuccon, Joel Mackenzie, and Teerapong Leelanupab. 2026. When LLM Judges Inflate Scores: Exploring Overrating in Relevance Assessment. In *Proc. ACM SIGIR*. 4322–4327.