

HIERA: Hierarchical Multi-Agent Relevance Assessment for Content Discovery Systems

Pritom Saha Akash, Phanideep Gampa, Chao Shen, Ying Chen, Sheikh Muhammad Sarwar
Amazon

{sapritom, phanide, shencha, yingchm, smsarwar}@amazon.com

Abstract

Content discovery systems depend on relevance judgment for search quality evaluation, but human annotation faces inter-annotator disagreement and scaling costs. While Large Language Models show promise as automated assessors, current approaches rely on flat aggregation strategies: single-step prompting, voting ensembles, or uncoordinated multi-agent pipelines that aggregate independent outputs without integration. We propose HIERA, a hierarchical multi-agent relevance assessment framework with four specialized agents: a Relevance Judge, Query Analyzer, Item Analyzer, and Relation Analyzer. The Judge determines when specialist analysis is needed; the Relation Analyzer then coordinates query and item analyses with external knowledge to establish relevance relationships for final judgment. Ablation studies show that the same agents and external knowledge without hierarchical coordination degrade performance, confirming that the coordination structure itself accounts for the improvement. Evaluation across five datasets (EVS, MSRD, ESCI, WANDS, Home Depot) shows improvements over 11 baselines: 10.2% on Home Depot, 4.8% on ESCI, and up to 38% on EVS ($p < 0.05$). Hierarchical coordination yields 12.7% improvement over uncoordinated collaboration using identical agents.

1 Introduction

Content discovery systems help users navigate vast repositories across streaming platforms and e-commerce marketplaces. Their effectiveness depends on **relevance assessment**: determining whether content matches user intent. These assessments underpin performance metrics (NDCG, MAP) (Järvelin and Kekäläinen, 2002; Schütze et al., 2008) and training data for recommendation models (Sanderson et al., 2010; Faggioli et al., 2023), directly impacting user satisfaction and platform success.

However, obtaining reliable assessments faces

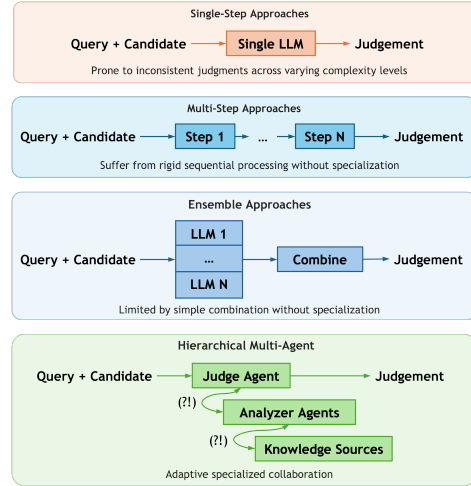


Figure 1: HIERA vs. other LLM-based approaches.

two challenges. First, relevance is inherently subjective, with annotators frequently disagreeing on labels (Voorhees, 1998; Bailey et al., 2008). Second, annotating thousands of query-item pairs across domains is economically infeasible (Thomas et al., 2024), forcing systems to operate with limited ground truth (Reddy et al., 2022; Deldjoo et al., 2020).

To address these challenges, Large Language Models (LLMs) have emerged as promising alternatives for automated relevance assessment (Faggioli et al., 2023; Thomas et al., 2024). These models can generate assessments at scale with reasonable human agreement (Soviero et al., 2024; Thomas et al., 2024). As shown in Figure 1, existing LLM-based approaches fall into three main paradigms. **Single-step approaches** use direct prompting to evaluate relevance in one step (Thomas et al., 2024; Soviero et al., 2024; Mehrdad et al., 2024), but compress all reasoning into a single inference, limiting their ability to decompose complex relevance relationships. **Multi-step approaches** employ sequential processing such as binary filtering followed by fine-grained classification (Schnabel et al., 2025), or progressive refinement through multiple passes (Hosseini et al., 2025), yet process information in

isolation across stages without integrating insights between components. **Ensemble approaches** combine outputs from multiple models through voting or blending (Rahmani et al., 2025; Qian et al., 2025), but lack coordination between models and specialized analysis of different relevance dimensions.

Despite these advances, existing LLM approaches face notable limitations when applied to content discovery systems. Traditional information retrieval matches explicit queries to documents based on topical relevance (Sanderson et al., 2010), while content discovery requires interpreting implicit user intent across multi-faceted items (Tsagkias et al., 2021). Consider the query “gray dresser” matched against “ajinkya 7 drawer dresser” with “metallic mercury finish.” The ground truth labels this as relevant, yet all existing approaches consistently classify it as only partially relevant. This systematic underestimation occurs because the case requires simultaneously interpreting color relationships (metallic mercury as gray), assessing style compatibility (ornate versus simple), and evaluating functional matching (storage needs). Single-step approaches compress these into one inference and miss the connections, multi-step pipelines eliminate the item during binary filtering before domain interpretation can occur, and ensemble methods combine multiple underestimations since no individual model captures the cross-dimensional relationships.

This points to three requirements unmet by current approaches. First, **specialized analysis**: interpreting color relationships, style compatibility, and functional matching each require distinct analytical expertise. Second, **structured coordination**: these analyses must inform each other, as color interpretation depends on item characterization, which depends on query intent. Third, **evidence integration**: final judgments must synthesize findings from multiple perspectives with external knowledge, rather than aggregating independent assessments.

To address these limitations, we propose **HIERA** (HIERarchical Relevance Assessment), a multi-agent framework that decomposes relevance assessment into specialized, coordinating components. As illustrated in Figure 1, HIERA uses a **three-level hierarchy** where a *Relevance Judge* determines when specialist analysis is needed and delegates to a *Relation Analyzer*, which consults a *Query Analyzer* for intent interpretation and an *Item Analyzer* for candidate characterization, then

synthesizes their outputs with external knowledge into a structured relevance argument returned to the Judge for final assessment.

Our contributions are threefold. First, we introduce HIERA, a hierarchical multi-agent framework for relevance assessment that enables specialized analysis, structured coordination, and evidence integration. Second, we demonstrate empirically that this hierarchical coordination outperforms flat aggregation (voting, blending, uncoordinated multi-agent) across five datasets (ESCI, WANDS, Home Depot, MSRD, and EVS), with improvements of 38% on EVS and 10.2% on Home Depot over the strongest baselines. Third, we demonstrate that providing agents with additional information or additional perspectives degrades performance without appropriate coordination structure, establishing that coordination topology determines whether scaling agents and knowledge sources helps or hurts.

2 Related Work

2.1 LLM-based Relevance Judgment

Single-Step Direct Assessment. Early LLM-based relevance judgment employed direct prompting where models evaluate query-document relevance in a single inference step. Faggioli et al. (Faggioli et al., 2023) established the theoretical foundation, with subsequent work demonstrating human-comparable accuracy (Thomas et al., 2024), consistent performance at scale (Upadhyay et al., 2024), and successful application to e-commerce (Soviero et al., 2024; Mehrdad et al., 2024). However, single-step approaches suffer from *reasoning compression*: they collapse all analytical reasoning into a single inference, preventing the decomposition needed to evaluate distinct aspects of complex relevance relationships.

Multi-Stage Sequential Processing. To address single-step limitations, researchers developed multi-stage pipelines that decompose relevance assessment into sequential components. Schnabel et al. (Schnabel et al., 2025) introduced binary filtering followed by fine-grained classification, showing improved accuracy over monolithic approaches. Farzi and Dietz (Farzi and Dietz, 2024, 2025) demonstrated how criteria-based evaluation stages enhance judgment accuracy, while Hosseini et al. (Hosseini et al., 2025) developed multimodal frameworks combining textual and visual analysis stages. These approaches achieve better performance through *analytical decomposition*, but suffer from *stage isolation*: stages either operate

independently or pass forward filtered outputs without the analytical reasoning behind them, limiting the integration needed when relevance depends on connecting query intent, item properties, and contextual knowledge.

Ensemble and Collaborative Methods. Recent work has explored collaborative approaches to overcome single-model limitations. Rahmani et al. (Rahmani et al., 2025) developed Judge-Blender, demonstrating that combining multiple LLM judges through ensemble methods consistently outperforms individual models by mitigating biases and improving robustness. Sachdev et al. (Sachdev et al., 2024) integrated retrieval-augmented generation for content discovery contexts. These collaborative methods address individual model limitations through *perspective diversity*, but suffer from *expertise uniformity*: they combine general-purpose models without specialized domain knowledge, and simple aggregation methods cannot synthesize specialized insights needed for complex relevance relationships.

These approaches lack mechanisms for coordinating specialized expertise, motivating hierarchical architectures that organize agents by analytical function with structured information flow.

2.2 LLM Agents

Large Language Models have been enhanced with agent capabilities for complex reasoning through environmental interaction. Early single-agent systems (AutoGPT (AutoGPT-Team, 2023), ReAct (Yao et al., 2023), QueryAgent (Huang et al., 2024)) demonstrated autonomous problem-solving through tool access but lack collaborative mechanisms for problems requiring diverse expertise.

Multi-agent frameworks have emerged where specialized agents collaborate on complex tasks. MetaGPT (Hong et al., 2023) simulates software company structures with role-based collaboration, AutoGen (Wu et al., 2024) provides agent conversation orchestration, and AgentVerse (Chen et al., 2023) explores emergent behaviors in multi-agent systems. In evaluation contexts, Agent-as-a-Judge (Zhuge et al., 2024) employs agents for assessment tasks, while CollabEval (Qian et al., 2025) demonstrates how multi-agent collaboration enhances evaluation quality through peer-based consensus. However, these approaches target general reasoning rather than content discovery’s specific challenges. No existing approach provides analytical specialization tailored to relevance assess-

ment: distinct expertise for query interpretation, item characterization, and relationship analysis coordinated through hierarchical synthesis.

3 Task Definition

The *relevance judgment* task involves assessing the degree of relevance between user queries and search results in content discovery systems such as **streaming services** and **e-commerce platforms**. Formally, given a query $q \in \mathcal{Q}$ expressing user information needs and a search result $c \in \mathcal{C}$ with associated metadata $\mathcal{M}(c)$, the task is to determine a relevance score $r \in \mathcal{R}$ indicating the degree of match between the query and result.

Query-result pairs in content discovery require multi-dimensional reasoning rather than uniform matching. Cases involving explicit attribute matching, such as “*action movies 2023*” $\rightarrow$ “*Top Gun: Maverick*”, permit evaluation through direct criteria comparison. Cases requiring semantic interpretation, such as “*gray dresser*” $\rightarrow$ “*metallic mercury finish dresser*”, demand reasoning across multiple dimensions: recognizing implicit relationships like color equivalence, assessing style compatibility, and verifying functional matching. This multi-dimensional nature poses challenges for current automated approaches, which typically lack mechanisms for decomposing the assessment into specialized analyses and integrating their findings, leading to suboptimal accuracy on cases requiring coordinated reasoning across query intent, item properties, and contextual knowledge.

4 Methodology

4.1 Framework Overview

Figure 2 illustrates HIERA, which decomposes relevance assessment into specialized analytical roles that integrate query understanding, item characterization, and relationship analysis through hierarchical coordination rather than independent aggregation.

The framework operates through a *three-layer* architecture. The Decision Layer contains the *Relevance Judge Agent* ($\mathcal{J}$), which determines what evidence is required and orchestrates consultation. The Analysis Layer provides three specialists: the *Relation Analyzer* ($\mathcal{R}$) coordinates by consulting the *Query Analyzer* ($\mathcal{Q}$) for query intent and the *Item Analyzer* ($\mathcal{I}$) for candidate properties. The Information Layer enables external knowledge access ($\mathcal{K}$) for factual verification.

Hierarchical Coordination Architecture. Agents

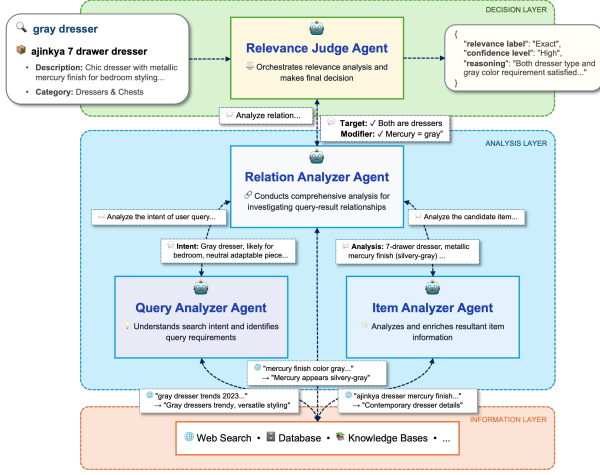


Figure 2: HIERA Framework

coordinate through a three-layer hierarchy where information flows between layers through tool-based consultation:

$$\text{Decision Layer : } \mathcal{J} \rightarrow \{\mathcal{R}\} \quad (1)$$

$$\text{Analysis Layer : } [\mathcal{R} \leftrightarrow \{\mathcal{Q}, \mathcal{I}\}] \rightarrow \mathcal{K} \quad (2)$$

$$\text{Information Layer : } \mathcal{K} \quad (3)$$

This hierarchical design enables structured evidence synthesis: each specialist’s output informs subsequent analytical steps, and the Relation Analyzer integrates findings into a coherent relevance argument before final judgment.

4.2 Framework Components

4.2.1 Decision Layer

Human experts decompose relevance assessment into distinct analytical steps: understanding user intent, characterizing candidate properties, and reasoning about their relationship. They consult specialized knowledge and synthesize findings before judgment, reflecting how structured coordination produces more reliable assessments than monolithic evaluation.

The *Relevance Judge Agent* ($\mathcal{J}$) embodies this pattern. $\mathcal{J}$ examines available query-candidate evidence, assesses which analytical dimensions require specialist input, and formulates guidance for $\mathcal{R}$ accordingly. When all dimensions are clear, $\mathcal{J}$ produces a direct judgment.

Multi-Dimensional Evidence Assessment. $\mathcal{J}$ evaluates three dimensions when assessing whether specialist consultation is needed. *Query Interpretability* evaluates whether user intent is explicitly stated or requires interpretation, as queries often contain implicit preferences beyond literal text. *Information Completeness* reflects whether available candidate information provides sufficient

evidence for confident judgment, as ambiguous product descriptions may demand clarification. *Relational Transparency* determines whether query-result semantic connections are immediately apparent, as cases requiring reasoning about relationships such as color equivalence or material compatibility prompt external verification.

These dimensions do not imply one another: a query may be clear while item information is incomplete, or vice versa. $\mathcal{J}$ engages $\mathcal{R}$ whenever any dimension presents insufficient evidence. Figure 2 illustrates this with “gray dresser” $\rightarrow$ “metallic mercury finish dresser”: $\mathcal{J}$ recognizes clear query interpretability but limited relational transparency, prompting consultation with $\mathcal{R}$, which integrates query intent (from $\mathcal{Q}$), item characterization (from $\mathcal{I}$), and external evidence (from $\mathcal{K}$) into a coherent relevance argument before $\mathcal{J}$ produces the final judgment.

4.2.2 Analysis Layer

When the Decision Layer identifies that consultation is needed, resolving gaps requires distinct expertise: extracting explicit requirements and implicit intent from queries, identifying stated and inferred attributes from candidates, and mapping semantic connections across dimensions. This separation reflects the single responsibility principle: agents perform more effectively when focused on specific tasks rather than handling diverse analytical challenges simultaneously.

The Analysis Layer addresses this through three specialized agents: the *Relation Analyzer Agent* ($\mathcal{R}$) performs connection analysis and coordinates specialist consultation, the *Query Analyzer Agent* ($\mathcal{Q}$) extracts explicit requirements and implicit intent, and the *Item Analyzer Agent* ($\mathcal{I}$) identifies both explicit and inferred candidate characteristics. The hierarchy is realized through $\mathcal{R}$, which coordinates specialist analyses and synthesizes their outputs into a structured relevance argument before returning evidence to $\mathcal{J}$.

Relation Analyzer Agent ($\mathcal{R}$). $\mathcal{R}$ analyzes the semantic connections between query and candidate while coordinating specialist consultation. Unlike peer-based multi-agent approaches that aggregate independent agent outputs, $\mathcal{R}$ acts as a hierarchical coordinator: it consults $\mathcal{Q}$ when query intent is uncertain, $\mathcal{I}$ when candidate information is insufficient, and $\mathcal{K}$ for factual verification, then synthesizes these findings into a structured relevance argument. For the “gray dresser” case, $\mathcal{R}$ identi-

fies that the color relationship requires both query intent analysis and item characteristic verification: it consults $\mathcal{Q}$ to clarify the implicit preference for furniture, $\mathcal{I}$ to verify that “metallic mercury finish” exhibits gray-toned visual properties, and $\mathcal{K}$ to confirm the color equivalence in furniture contexts. The resulting argument is returned to $\mathcal{J}$ for final judgment.

Query Analyzer Agent ($\mathcal{Q}$). $\mathcal{Q}$ focuses on query understanding, activated when $\mathcal{R}$ identifies insufficient query interpretability. It decomposes queries into explicit requirements (stated constraints) and implicit intent (unstated preferences for style or functionality). When contextual understanding is needed, $\mathcal{Q}$ accesses $\mathcal{K}$ for clarification, returning findings to $\mathcal{R}$.

Item Analyzer Agent ($\mathcal{I}$). $\mathcal{I}$ focuses on candidate characterization, activated when $\mathcal{R}$ identifies insufficient candidate information. It extracts explicit characteristics (directly stated features) and inferred properties (implicit attributes such as color appearance or style category). When verification is needed, $\mathcal{I}$ accesses $\mathcal{K}$, returning findings to $\mathcal{R}$.

Coordination Protocol. When $\mathcal{J}$ determines that confident assessment requires additional evidence, it delegates to $\mathcal{R}$ with guidance on identified gaps. $\mathcal{R}$ performs its own connection analysis and, based on the gaps encountered, consults specialists: $\mathcal{Q}$ for query interpretation, $\mathcal{I}$ for candidate characterization, and $\mathcal{K}$ for factual verification. Critically, this is integrative rather than independent: $\mathcal{R}$ consults specialists based on identified gaps and synthesizes their findings into a single structured argument before returning it to $\mathcal{J}$.

This coordination pattern follows:

$$\text{Gaps} = \mathcal{R}(q, c, \text{guidance}_{\mathcal{J}}) \quad (4)$$

$$\text{Active} = \{s \in \{\mathcal{Q}, \mathcal{I}, \mathcal{K}\} : s \text{ addresses Gaps}\} \quad (5)$$

$$\text{Result} = \text{integrate}\left(\bigcup_{s \in \text{Active}} s(\text{Gaps})\right) \quad (6)$$

This hierarchical structure differs from peer-based multi-agent approaches where agents operate independently and outputs are aggregated post-hoc. Because all specialist findings pass through $\mathcal{R}$ before reaching $\mathcal{J}$, the final relevance argument reflects how query intent, item properties, and external evidence relate to each other rather than treating them as independent signals.

4.2.3 Information Layer

The Information Layer ($\mathcal{K}$) extends HIERA’s analytical capabilities beyond parametric knowledge when Analysis Layer agents identify information gaps. $\mathcal{K}$ provides external knowledge access to support agent analysis when internal reasoning is insufficient for establishing relevance relationships.

Knowledge Integration. Analysis Layer agents ($\mathcal{R}, \mathcal{Q}, \mathcal{I}$) access $\mathcal{K}$ when they identify information gaps that cannot be resolved through internal reasoning. $\mathcal{K}$ provides factual grounding needed to establish semantic connections that require domain-specific verification. The layer supports diverse knowledge sources including structured databases and web search to accommodate domain-specific requirements.

Structured Knowledge Access. Agents formulate targeted queries to $\mathcal{K}$ based on specific gaps identified during their analysis. Retrieved information is returned to the requesting agent, which integrates it into its ongoing analytical process before passing findings up through the hierarchy. This design ensures that external knowledge is contextualized within the agent’s analytical frame rather than appended as raw supplementary text.

5 Experimental Setup

5.1 Datasets

We evaluate HIERA across five content discovery datasets spanning *entertainment search* (EVS, MSRD) and *product search* (ESCI, WANDS, Home Depot). For public datasets, we construct evaluation subsets of approximately 1,000 query-item pairs through random sampling with fixed seed (42), ensuring equal representation of each relevance level. **EVS** is a proprietary entertainment video search dataset (300 query-title pairs, 3-level scale) with rich structured metadata and domain-specific database access. **MSRD**¹ contains binary relevance annotations for movie search covering actors, directors, genres, themes, and cultural references. **ESCI** (Reddy et al., 2022) provides e-commerce query-product pairs across diverse categories with 4-level taxonomy (Exact, Substitute, Complement, Irrelevant). **WANDS** (Chen et al., 2022) covers furniture and home decor search with 3-level relevance (Exact Match, Partial Match, Irrelevant). **Home Depot**² features home improvement product search with 3-level relevance across

¹<https://github.com/metarank/msrd>

²<https://www.kaggle.com/c/home-depot-product-search-relevance/data>

diverse query types including product names and technical specifications.

5.2 Baselines

We compare against 11 methods across three categories. **Single-step:** (1) **Zero-Shot** (Faggioli et al., 2023) uses direct prompting without examples; (2) **Multi-Criteria** (Mehrdad et al., 2024) evaluates multiple relevance dimensions through structured prompting; (3) **DNA-Prompt** (Thomas et al., 2024) uses Descriptive-Narrative-Aspects structured reasoning. **Multi-step:** (4) **Binary-Graded** (Rahmani et al., 2025) applies binary filtering then fine-grained scoring; (5) **Multi-Stage** (Schnabel et al., 2025) refines judgments through progressive reasoning stages; (6) **Self-Instruct** (Soviero et al., 2024) generates annotation guidelines from examples then applies them; (7) **RAG-MMR** (Sachdev et al., 2024) combines retrieval-augmented generation with diverse example selection; (8) **RAER** (Hosseini et al., 2025) generates query-specific evaluation criteria through iterative refinement. **Ensemble/Collaboration:** (9) **Prompt-Blender** (Rahmani et al., 2025) aggregates multiple prompt formulations; (10) **LLM-Blender** (Rahmani et al., 2025) ensembles multiple LLMs with diverse strategies; (11) **CollabEval** (Qian et al., 2025) uses multi-agent collaborative evaluation with consensus building.

5.3 Evaluation Metrics

We evaluate all methods using three standard metrics for relevance assessment: Accuracy (**Acc**) measures the proportion of correctly classified query-result pairs, Macro F1-score (**F1**) computes F1-score for each relevance class independently and averages them ensuring balanced evaluation across relevance categories, and Cohen’s κ (κ) assesses agreement between predicted and ground truth labels while accounting for chance agreement. Statistical significance is determined using McNemar’s Test for paired comparisons with Bonferroni correction ($p < 0.05$). Bootstrap 95% confidence intervals for all ablation conditions are reported in Appendix D.

5.4 Implementation Details

We implement *HIERA* using Claude 3.7 Sonnet ($\mathcal{J}$, $\mathcal{R}$) and Claude 3.5 Haiku ($\mathcal{Q}$, $\mathcal{I}$) via AWS Bedrock with temperature 0, using LangGraph³ with ReAct design pattern (Yao et al., 2023). All baselines use Claude 3.7 Sonnet for fair comparison; LLM-

Blender and CollabEval use diverse model sets as required by their methodologies. Full implementation details including tool access configuration, baseline model assignments, and system prompts are provided in Appendix A.

6 Results

6.1 Overall Performance

Table 1 presents results for entertainment search datasets (EVS and MSRD), while Table 2 shows product search performance (ESCI, WANDS, and Home Depot). All results include accuracy, F1-score, and Cohen’s κ metrics. Statistical significance is assessed using McNemar’s test with Bonferroni correction ($p < 0.05$), with asterisks (*) indicating statistically significant differences compared to *HIERA*. Note that Self-Instruct and RAG-MMR results are unavailable for EVS dataset due to the lack of few-shot samples required by these methods.

Table 1: Entertainment results. **Bold:** best, *: $p < 0.05$.

Method	EVS			MSRD		
	Acc	F1	κ	Acc	F1	κ
Zero-Shot	0.517*	0.586*	0.271*	0.762*	0.753*	0.524*
Multi-Criteria	0.507*	0.567*	0.238*	0.723*	0.706*	0.446*
DNA-Prompt	0.557*	0.606*	0.292*	0.739*	0.726*	0.478*
Binary-Graded	0.400*	0.468*	0.136*	0.810*	0.540*	0.620*
Multi-Stage	0.427*	0.361*	0.156*	0.768*	0.517*	0.550*
Self-Instruct	—	—	—	0.839	0.838*	0.678*
RAG-MMR	—	—	—	0.821*	0.819*	0.642*
RAER	0.533*	0.594*	0.286*	0.768*	0.759*	0.536*
Prompt-Blender	0.520*	0.566*	0.243*	0.755*	0.744*	0.510*
LLM-Blender	0.543*	0.586*	0.264*	0.758*	0.748*	0.516*
CollabEval	0.567*	0.609*	0.284*	0.773*	0.769*	0.546*
HIERA (Ours)	0.713	0.699	0.461	0.859	0.858	0.718

Table 2: Product search results. **Bold:** best, *: $p < 0.05$.

Method	ESCI			WANDS			Home Depot		
	Acc	F1	κ	Acc	F1	κ	Acc	F1	κ
Zero-Shot	0.481*	0.480*	0.308*	0.630*	0.625*	0.445*	0.414*	0.415*	0.122*
Multi-Criteria	0.401*	0.405*	0.201*	0.654	0.647*	0.481*	0.421*	0.415*	0.132*
DNA-Prompt	0.445*	0.450*	0.260*	0.628*	0.622*	0.442*	0.427*	0.428*	0.141*
Binary-Graded	0.444*	0.423*	0.259*	0.627*	0.627*	0.440*	0.405*	0.406*	0.108*
Multi-Stage	0.433*	0.342*	0.245*	0.595*	0.440*	0.393*	0.379*	0.285*	0.070*
Self-Instruct	0.518	0.519*	0.357*	0.649*	0.653*	0.473*	0.407*	0.407*	0.111*
RAG-MMR	0.500*	0.498*	0.333*	0.638*	0.645*	0.457*	0.404*	0.405*	0.107*
RAER	0.474*	0.466*	0.299*	0.627*	0.617*	0.440*	0.412*	0.309*	0.120*
Prompt-Blender	0.472*	0.475*	0.296*	0.648*	0.644*	0.472*	0.414*	0.415*	0.122*
LLM-Blender	0.428*	0.430*	0.237*	0.662	0.663*	0.493*	0.416*	0.418*	0.125*
CollabEval	0.480*	0.482*	0.307*	0.630*	0.638*	0.445*	0.410*	0.399*	0.116*
HIERA (Ours)	0.543	0.545	0.391	0.681	0.686	0.521	0.464	0.465	0.197

HIERA demonstrates substantial and consistent improvements across all evaluation datasets, with particularly notable gains on challenging domains. The framework achieves its most dramatic improvement on the EVS dataset, outperforming the best baseline Zero-Shot by 38% in accuracy (0.713 vs 0.517) and 70% in Cohen’s κ (0.461 vs 0.271). On product search datasets, *HIERA* consistently outperforms the strongest baselines: 4.8% improvement over Self-Instruct on ESCI (0.543 vs 0.518), 4.3% over LLM-Blender on WANDS (0.681 vs

³<https://langchain-ai.github.io/langgraph/>

0.654), and 10.2% over Multi-Criteria on Home Depot (0.464 vs 0.421). Note that Home Depot exhibits low absolute agreement across all methods due to inherent annotation ambiguity in this dataset; the relative improvement remains substantial. Even on the relatively easier MSRD dataset, *HIERA* achieves 2.4% improvement over Self-Instruct (0.859 vs 0.839), demonstrating effectiveness across varying complexity levels.

Across baseline categories, single-step approaches show inconsistent performance (Zero-Shot: 0.517 on EVS vs 0.762 on MSRD), multi-step methods perform better with Self-Instruct as the strongest traditional baseline, and ensemble methods (LLM-Blender, CollabEval) show mixed results despite combining multiple perspectives. *HIERA* consistently outperforms all categories, indicating that structured coordination with specialized roles provides genuine advantages over both individual methods and aggregation-based approaches.

6.2 Ablation Study

To understand each component’s contribution, we conduct ablation studies on MSRD and WANDS examining three aspects: (1) external knowledge integration through web search, (2) individual agent roles in the Analysis Layer, and (3) coordination through the Relation Analyzer. Table 3 presents results across component combinations.

Table 3: Ablation Study Results

Variant	MSRD			WANDS		
	Acc	F1	κ	Acc	F1	κ
Judge Only	0.747	0.737	0.494	0.622	0.627	0.433
Judge + Web Search	0.752	0.743	0.504	0.617	0.620	0.425
Judge + Relation Analyzer	0.748	0.738	0.496	0.668	0.673	0.501
Judge + Relation Analyzer + Web Search	0.778	0.774	0.556	0.655	0.661	0.482
Judge + Query&Item Analyzers	0.741	0.730	0.482	0.627	0.627	0.440
Judge + Query&Item Analyzers + Web Search	0.732	0.719	0.463	0.621	0.623	0.431
Judge + All Analyzers	0.762	0.755	0.524	0.659	0.664	0.488
HIERA (Full System)	0.859	0.858	0.718	0.681	0.686	0.521

The ablation shows three patterns. First, external knowledge alone provides minimal benefits: adding web search to Judge-only yields negligible change (0.747 to 0.752 on MSRD, 0.622 to 0.617 on WANDS). Second, the Relation Analyzer is the crucial component: its inclusion yields substantial improvements (0.748 on MSRD, 0.668 on WANDS), amplified further with web search (0.778, 0.655), while Q and I without R show limited effectiveness and even degradation with web

search (0.741 to 0.732 on MSRD). Third, the full system (0.859 on MSRD, 0.681 on WANDS) substantially outperforms the complete agent ensemble without coordination (0.762, 0.659). The Judge already performs structured multi-dimensional reasoning within a 3,000-token budget, and in the Judge + All Analyzers condition it additionally receives specialist analyses as context, comparable to extended chain-of-thought with multiple analytical perspectives. The persistent gap confirms that the improvement requires interactive synthesis through $\mathcal{R}$: both configurations provide the same agents and knowledge, but only the full system routes consultation through $\mathcal{R}$ for synthesis before returning evidence to $\mathcal{J}$.

7 Analysis and Discussion

Having established *HIERA*’s superior performance through quantitative evaluation and ablation studies, we now provide deeper analysis across three key dimensions: (1) the coordination mechanisms that enable multi-agent collaboration, (2) the effectiveness of external knowledge integration strategies, and (3) computational cost analysis.

7.1 Multi-Agent Coordination Patterns

The Relation Analyzer serves as the central synthesizer: it receives guidance from $\mathcal{J}$, consults specialists as needed, and integrates their findings into a structured relevance argument. This contrasts with ensemble methods where agents operate independently and outputs are aggregated post-hoc. Three behavioral patterns characterize this coordination: (1) **Role-based specialization**: each agent addresses a specific analytical dimension ($\mathcal{Q}$ for query intent, $\mathcal{I}$ for candidate properties, $\mathcal{R}$ for relationships), operating within a focused scope rather than reasoning about all dimensions simultaneously. (2) **Hierarchical synthesis**: $\mathcal{R}$ synthesizes specialist findings into a coherent argument that accounts for interactions between query intent, item properties, and external evidence, distinguishing the full system from the uncoordinated ablation where agents produce independent analyses without integration. (3) **Knowledge integration**: external knowledge improves performance only when integrated through the coordination hierarchy; as shown in Table 4, the same web search results that improve *HIERA* (+12.7% on MSRD) degrade most baselines when concatenated directly.

Activation frequency. $\mathcal{J}$ delegates to the full hierarchy in 91.5% of cases, with only 8.5% receiv-

ing direct judgment. This is expected: the system is designed as a coordination architecture, not an adaptive router. The contribution is not selective activation but the coordination protocol itself, as confirmed by the ablation: the same cases processed by all agents without hierarchical coordination (Judge + All Analyzers) achieve 37% lower κ despite identical coverage. The value lies in how specialists are coordinated through $\mathcal{R}$, not in when they are activated.

Table 4: External knowledge integration effectiveness.

Method	MSRD			WANDS		
	Acc	F1	κ	Acc	F1	κ
Zero-Shot	+0.3%	+0.4%	+0.8%	+1.4%	+1.8%	+3.0%
Multi-Criteria	-2.4%	-2.9%	-7.6%	-5.6%	-5.6%	-11.0%
DNA-Prompt	-0.4%	-0.5%	-1.3%	-1.3%	-1.3%	-2.7%
Binary-Graded	-0.5%	-0.5%	-1.4%	-1.6%	-1.6%	-3.4%
Multi-Stage	-5.6%	-8.0%	-18.2%	+1.2%	+0.9%	+2.6%
Self-Instruct	-0.7%	-0.6%	-1.8%	+0.0%	+0.2%	+0.0%
RAG-MMR	-5.4%	-5.8%	-13.7%	+3.0%	+2.0%	+6.2%
RAER	-1.3%	-1.3%	-3.7%	+0.6%	+1.7%	+1.4%
Prompt-Blender	+0.4%	+0.5%	+1.2%	-0.2%	-0.1%	-0.3%
LLM-Blender	+0.8%	+0.8%	+2.3%	-1.5%	-2.0%	-3.0%
CollabEval	+1.2%	+0.9%	+3.3%	-6.5%	-7.0%	-13.8%
HIERA	+12.7%	+13.6%	+37.0%	+3.3%	+3.3%	+6.8%

7.2 External Knowledge Integration Impact

To test whether coordination provides genuine advantages over simpler knowledge augmentation, we provide all baseline methods with identical web search results used by HIERA, concatenated directly to their input prompts (Table 4). This controls for information availability, isolating the effect of integration strategy. Most methods degrade despite accessing the same information: Multi-Criteria (-2.4% on MSRD, -5.6% on WANDS), Multi-Stage (-5.6% on MSRD), and even CollabEval shows inconsistent behavior (+1.2% on MSRD, -6.5% on WANDS). In contrast, HIERA achieves consistent improvements (+12.7% on MSRD, +3.3% on WANDS), demonstrating that effective knowledge integration requires coordination rather than concatenation.

7.3 Computational Cost Analysis

Token Usage Analysis. We analyze token consumption on MSRD to assess whether gains arise from coordination structure rather than increased computation. HIERA consumes 2,917 tokens per case at the orchestration level. The uncoordinated ablation (Judge + All Analyzers) consumes 2,866 tokens, within 1.8% of HIERA’s budget, yet scores 37% lower in Cohen’s κ (0.524 vs. 0.718). The Judge alone achieves $\kappa = 0.494$ within a 3,000-token budget; enriching it with all specialist analyses yields only $\kappa = 0.524$, confirming that coordina-

tion topology, not token volume or richer context, accounts for the gain. For comparison, CollabEval broadcasts full evaluation histories across multiple discussion rounds, consuming 9,621 tokens while achieving lower agreement ($\kappa = 0.546$).

Latency. HIERA achieves mean response times of 35.6s on MSRD and 35.9s on WANDS, compared to 17.7–20.2s for CollabEval and 0.93–0.98s for single-step methods. However, HIERA remains 50× faster than human experts (35s vs 30 minutes per judgment (Sachdev et al., 2024)), making it practical for offline evaluation tasks where accuracy requirements justify the computational investment.

Qualitative case study analysis (Appendix E) reveals that HIERA excels at detecting constraint violations through coordinated specialist consultation (e.g., language mismatches, functional misalignments), but can amplify errors when agents reinforce initial misinterpretations in cases requiring subjective cultural judgment.

8 Conclusion

We present *HIERA*, a hierarchical multi-agent framework for relevance assessment in content discovery systems. Evaluation across five datasets demonstrates consistent improvements over 11 baselines: 38% on EVS, 10.2% on Home Depot, and statistically significant gains across all datasets. Ablation studies confirm that the coordination structure itself accounts for the improvement: the same agents and knowledge access without hierarchical coordination score 37% lower in Cohen’s κ , and external knowledge that improves performance under coordination degrades most baselines when provided directly. These findings demonstrate that structured orchestration through a central synthesizer outperforms both independent multi-agent operation and flat aggregation strategies for relevance assessment tasks requiring multi-dimensional reasoning. For practitioners building multi-agent evaluation systems, the implication is clear: investing in coordination protocols yields larger gains than adding more agents or more expensive models.

Limitations

We acknowledge several limitations of this work. First, one of our evaluation datasets (EVS) is proprietary, limiting full reproducibility; however, we evaluate on four public datasets to enable independent verification. Second, HIERA incurs higher latency than single-step approaches (35.6s vs un-

der 1s), though it remains $50\times$ faster than human annotation and operates within the same order of magnitude as other multi-agent approaches such as CollabEval (37.1s). Third, our evaluation focuses on English-language content discovery; generalization to multilingual settings remains unexplored. Finally, the Judge’s consultation decisions rely on LLM reasoning, which may exhibit variability across model versions or repeated runs; we mitigate this with temperature 0 but do not formally evaluate decision stability across multiple runs. Additionally, all experiments use Claude models; generalization to other LLM families (e.g., GPT-4, Gemini) remains to be validated.

References

- AutoGPT-Team. 2023. [Autogpt: Build, deploy, and run ai agents](#). GitHub repository.
- Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P de Vries, and Emine Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 667–674.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.
- Yan Chen, Shujian Liu, Zheng Liu, Weiye Sun, Linas Baltrunas, and Benjamin Schroeder. 2022. Wands: Dataset for product search relevance assessment. In *Proceedings of the 44th European Conference on Information Retrieval*.
- Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, 53(5):1–38.
- Guglielmo Faggioli, Laura Dietz, Charles LA Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and 1 others. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.
- Naghmeh Farzi and Laura Dietz. 2024. Best in tau@llmjude: Criteria-based relevance evaluation with llama3. *arXiv preprint arXiv:2410.14044*.
- Naghmeh Farzi and Laura Dietz. 2025. Criteria-based llm relevance judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (IC-TIR)*, pages 254–263.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiaowu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.
- Kasra Hosseini, Thomas Kober, Josip Krapac, Roland Vollgraf, Weiwei Cheng, and Ana Peleteiro Ramallo. 2025. Retrieve, annotate, evaluate, repeat: Leveraging multimodal llms for large-scale product retrieval evaluation. In *European Conference on Information Retrieval*, pages 149–163. Springer.
- Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. 2024. Queryagent: A reliable and efficient reasoning framework with environmental feedback-based self-correction. *arXiv preprint arXiv:2403.11886*.
- Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.
- Navid Mehrdad, Hrushikesh Mohapatra, Mossaab Bagdouri, Prijith Chandran, Alessandro Magnani, Xunfan Cai, Ajit Puthenpuhussery, Sachin Yadav, Tony Lee, ChengXiang Zhai, and 1 others. 2024. Large language models for relevance judgment in product search. *arXiv preprint arXiv:2406.00247*.
- Yiyue Qian, Shinan Zhang, Yun Zhou, Haibo Ding, Diego Socolinsky, and Yi Zhang. 2025. Enhancing llm-as-a-judge via multi-agent collaboration.
- Hossein A Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2025. Judgeblender: Ensembling automatic relevance judgments. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1268–1272.
- Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. [Shopping queries dataset: A large-scale ESCI benchmark for improving product search](#).
- Jayant Sachdev, Sean D Rosario, Abhijeet Phatak, He Wen, Swati Kirti, and Chittaranjan Tripathy. 2024. Automated query-product relevance labeling using large language models for e-commerce search. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 32–40.
- Mark Sanderson and 1 others. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375.
- Julian A Schnabel, Johanne R Trippas, Falk Scholer, and Danula Hettichchi. 2025. Multi-stage large language model pipelines can outperform gpt-4o in relevance assessment. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1288–1292.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Beatriz Soviero, Daniel Kuhn, Alexandre Salle, and Viviane Pereira Moreira. 2024. Chatgpt goes shopping: LLMs can predict relevance in ecommerce search. In *European Conference on Information Retrieval*, pages 3–11. Springer.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940.

Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten De Rijke. 2021. Challenges and research opportunities in ecommerce search and recommendations. In *ACM Sigir Forum*, volume 54, pages 1–23. ACM New York, NY, USA.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A large-scale study of relevance assessments with large language models: An initial look. *arXiv preprint arXiv:2411.08275*.

Ellen M Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 315–323.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yuniang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, and 1 others. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

A Implementation Details

A.1 HIERA Configuration

The Decision Layer ($\mathcal{J}$) employs Claude 3.7 Sonnet for relevance judgment with 3000 max tokens, while the Analysis Layer uses Claude 3.5 Haiku for Query Analyzer ($\mathcal{Q}$) and Item Analyzer ($\mathcal{I}$) with

1000 max tokens each. Relation Analyzer ($\mathcal{R}$) uses Claude 3.7 Sonnet with 1500 max tokens for coordination tasks. All agents use temperature 0 for deterministic outputs.

Framework Implementation: HIERA is implemented using LangGraph with ReAct design pattern, enabling agents to interleave reasoning and tool usage in iterative cycles. Each agent follows: (1) Reasoning - analyze current state and determine next action, (2) Acting - execute tools or consult specialists, (3) Observing - process results and update state. LangGraph manages agent coordination through directed graphs where nodes represent agents and edges define communication pathways, with persistent state management across agent interactions.

Tool Access: Each agent has access to specific tools reflecting its role in the hierarchy. $\mathcal{J}$ can invoke $\mathcal{R}$ as a tool. $\mathcal{R}$ can invoke $\mathcal{Q}$, $\mathcal{I}$, and web search. $\mathcal{Q}$ and $\mathcal{I}$ can each invoke web search and database retrieval independently. For the EVS dataset, both web search and structured database access are available; for public datasets (MSRD, ESCI, WANDS, Home Depot), only web search is used. Web search returns 5 results per query. All experiments use consistent system prompts for each agent role, with fixed prompt templates maintained across all datasets.

A.2 Baseline Configuration

For baseline methods, most approaches use Claude 3.7 Sonnet with temperature 0.0 for fair comparison with HIERA’s main decision-making agents, including Zero-Shot, Multi-Criteria, DNA-Prompt, Binary-Graded, Self-Instruct, RAG-MMR, RAER, and Prompt-Blender. The Multi-Stage baseline uses a two-model approach: Claude 3.5 Haiku for binary filtering (Stage 1) and Claude 3.7 Sonnet for fine-grained classification (Stage 2). LLM-Blender and CollabEval employ diverse model sets as required by their original methodologies: LLM-Blender uses three different models (Claude 3.7 Sonnet, Amazon Nova Pro, and Mistral Large) for ensemble diversity, while CollabEval uses four diverse models (Mistral Large, Claude 3.5 Haiku, Claude 3.5 Sonnet, and Amazon Nova Pro) for multi-agent evaluation with Claude 3.7 Sonnet as the final judge.

B Agent System Prompt Templates

This document provides the complete system prompt templates used in HIERA’s multi-agent ar-

Judge Agent Prompt Template

TASK: Judge relevance between query and candidate result
DOMAIN: {domain} {domain_context}
JUDGMENT CRITERIA: {judgment_criteria}

You are an expert relevance judge.

Evaluate this case and determine your approach:

- Assess whether you can make a confident judgment with available information
- If straightforward, proceed with your assessment
- If complex or uncertain, engage appropriate analytical support

Your judgment guides the process:

- Trust your expertise for clear cases
- Seek specialist input when facing ambiguity
- Adapt your thoroughness to match case complexity

AVAILABLE TOOLS: {tool_description}

OUTPUT FORMAT: JSON only
{"relevance_label": "{labels}",
"confidence_level": "high | medium | low",
"reasoning": "explanation grounded on criteria"}
CRITICAL: Respond with valid JSON only, no additional text.

Figure 3: Judge Agent prompt template.

chitecture for full reproducibility. Each template serves as the system prompt for the respective agent, defining its role, capabilities, and behavioral constraints within the multi-agent coordination framework. All prompts utilize configurable parameters (shown in red) that are dynamically populated based on domain-specific configurations and dataset characteristics.

B.1 Judge Agent System Prompt

The Judge Agent system prompt implements the three-dimensional evidence assessment and coordinates multi-agent consultation through the hierarchical framework.

B.2 Relation Analyzer Agent System Prompt

The Relation Analyzer Agent system prompt defines the coordination specialist role, responsible for orchestrating specialist consultations and synthesizing relationship analysis.

B.3 Query Analyzer Agent System Prompt

The Query Analyzer Agent system prompt specializes in query interpretation and intent extraction.

B.4 Item Analyzer Agent System Prompt

The Item Analyzer Agent system prompt focuses on candidate result analysis, extracting key characteristics and features.

C Instantiated Prompt Example

Below is the Judge Agent prompt as instantiated for the WANDS (Home Furnishings) dataset:

D Bootstrap Confidence Intervals

Table 5 reports 95% bootstrap confidence intervals (1,000 iterations) for all *HIERA* ablation conditions on MSRD and WANDS, demonstrating result

Relation Analyzer Agent Prompt Template

TASK: Analyze relationships between query and candidate result
DOMAIN: {domain} {domain_context}
JUDGMENT CONTEXT: {judgment_criteria}

Provide comprehensive relationship analysis considering all available information and judgment criteria. Be concise and direct for judge decision-making.

You are an expert relationship analyst.

Evaluate your analytical capability for this case:

- Assess whether you can analyze relationships confidently
- If straightforward, proceed with direct analysis
- If complex or uncertain, engage appropriate analytical tools

Your expertise guides the process:

- Trust your analytical skills for clear relationships
- Seek additional insight when facing ambiguity
- Adapt your analytical depth to match case complexity

AVAILABLE TOOLS: {tool_description}

OUTPUT FORMAT: JSON only
{"analysis": "comprehensive relationship analysis",
"confidence": "high | medium | low"}
CRITICAL: Respond with valid JSON only, no additional text.

Figure 4: Relation Analyzer prompt template.

Query Analyzer Agent Prompt Template

TASK: Analyze query to identify user intent and search characteristics
DOMAIN: {domain} {domain_context}
FOCUS AREAS: {focus_attributes_text}

You are an expert query analyst.

Evaluate your analytical capability for this query:

- Assess whether you can understand query intent confidently
- If straightforward, proceed with direct analysis
- If complex or uncertain, engage appropriate analytical tools

Your expertise guides the process:

- Trust your analytical skills for clear queries
- Seek additional insight when facing ambiguity
- Adapt your analytical depth to match query complexity

AVAILABLE TOOLS: {tool_description}

OUTPUT FORMAT: JSON only
{"analysis": "query intent and characteristics analysis",
"confidence": "high | medium | low"}
CRITICAL: Respond with valid JSON only, no additional text.

Figure 5: Query Analyzer prompt template.

stability.

Table 5: 95% bootstrap confidence intervals.

Variant	MSRD		WANDS	
	Acc [95% CI]	κ [95% CI]	Acc [95% CI]	κ [95% CI]
Judge Only	.748 [.720,.772]	.496 [.446,.541]	.623 [.591,.652]	.434 [.387,.477]
Judge + Search	.753 [.726,.777]	.506 [.455,.553]	.618 [.590,.647]	.427 [.383,.470]
Judge + R	.747 [.720,.774]	.494 [.446,.543]	.667 [.637,.697]	.500 [.456,.544]
Judge + R + Search	.777 [.751,.803]	.554 [.505,.606]	.656 [.628,.686]	.483 [.441,.528]
Judge + Q&I	.742 [.712,.769]	.484 [.431,.534]	.628 [.599,.658]	.442 [.398,.485]
Judge + Q&I + Search	.733 [.704,.760]	.465 [.413,.515]	.622 [.592,.654]	.433 [.387,.480]
Judge + All Analyzers	.761 [.734,.787]	.522 [.471,.572]	.658 [.630,.688]	.486 [.445,.531]
<i>HIERA</i> (Full)	.858 [.835,.879]	.716 [.670,.757]	.680 [.650,.709]	.519 [.475,.564]

E Case Study Analysis

To understand *HIERA*’s capabilities and limitations, we examine actual execution traces from our experimental runs, analyzing both successful and failed cases to identify systematic patterns in multi-agent behavior.

E.1 Successful Cases Analysis

Figure 8 shows two cases where *HIERA* handles complex relevance scenarios that challenge single-step methods. In Case 1, the system progresses

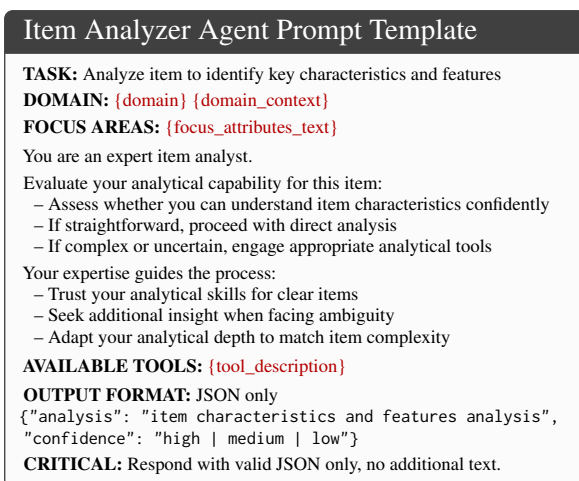


Figure 6: Item Analyzer prompt template.

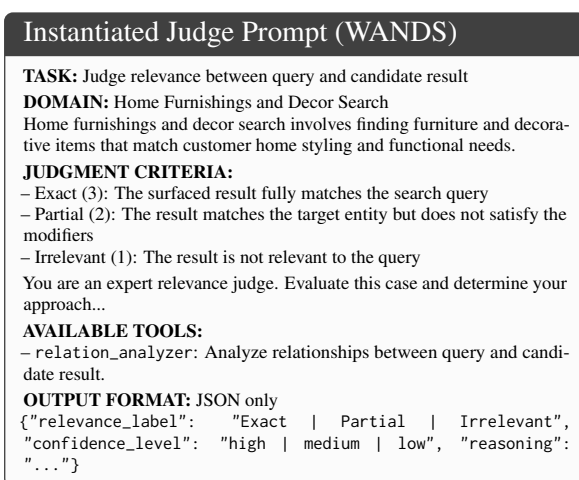


Figure 7: Instantiated Judge prompt for WANDS dataset.

from initial assessment → semantic analysis → external validation → final decision, building evidence for language constraint violations. Case 2 shows functional analysis → category verification → purpose differentiation, enabling distinction between semantic similarity and functional relevance. This contrasts with single-step approaches that must make immediate judgments without systematic constraint checking.

The successful cases share common characteristics: (1) *Clear constraint identification* - missing language specifications and functional misalignments are explicitly detected, (2) *Sequential evidence building* - each agent contributes distinct analytical perspectives that build toward definitive conclusions, and (3) *Appropriate complexity matching* - the system’s thorough analysis is justified by the nuanced nature of the relevance decisions required.

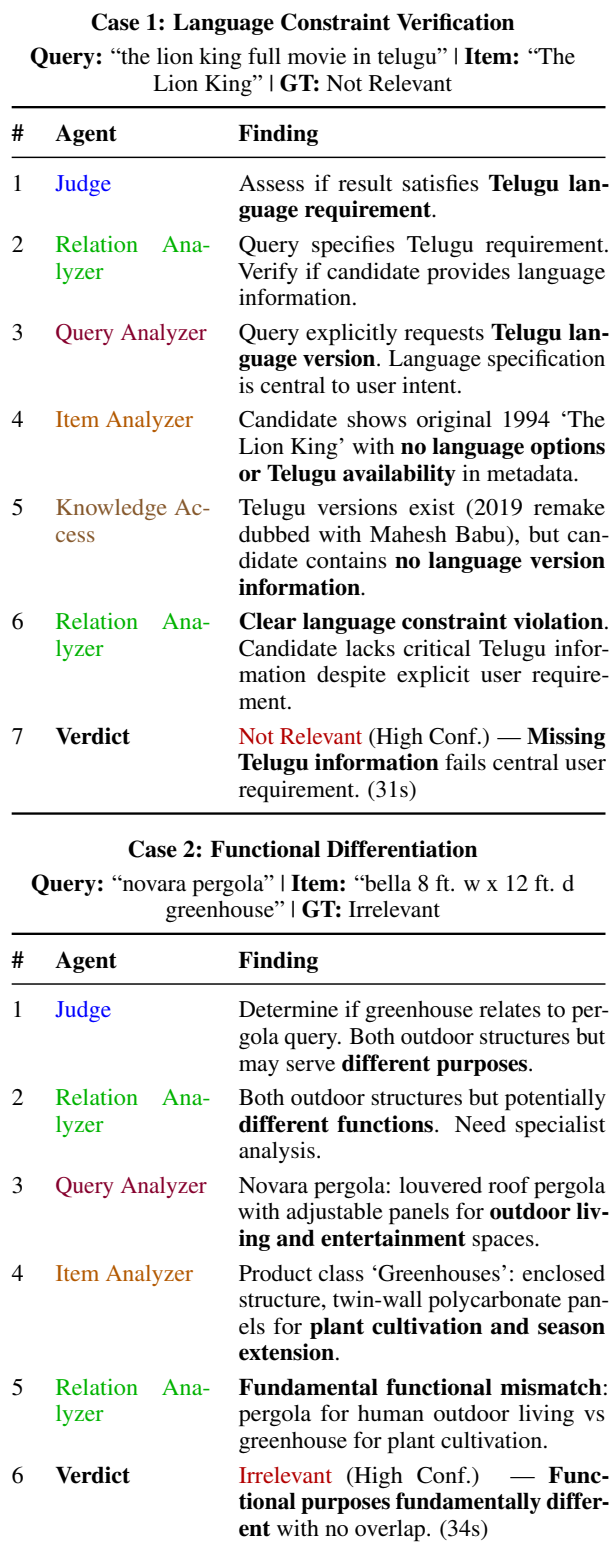


Figure 8: HIERA Success Cases: Hierarchical reasoning steps for relevance judgment

E.2 Failure Cases Analysis

Figure 9 shows failure cases where *HIERA* produced incorrect judgments despite sophisticated reasoning. These cases reveal systematic limitations: cultural over-interpretation (the Relation Analyzer’s orchestration leads specialists to reinforce inappropriate cultural associations) and specificity over-enforcement (systematic consultation becomes overly rigid, with agents collectively rejecting partially relevant results that human annotators accept).

Case 1: False Positive (Over-interpreting Cultural Connections)		
Query: “zen” Item: “japanese floating koi figurine” GT: Irrelevant		
#	Agent	Finding
1	Judge	Query ‘zen’ is broad, referring to Japanese philosophical concept emphasizing simplicity, mindfulness, and harmony with nature.
2	Relation Analyzer	Need to assess cultural connections . Consult specialists for query interpretation and item analysis.
3	Query Analyzer	Zen represents home decor style search with minimalist, natural design principles. Japanese cultural connection suggests potential match.
4	Item Analyzer	Japanese decorative figurine with cultural symbolism. Zen-inspired aesthetic with minimalist design, suitable for water features.
5	Relation Analyzer	Koi fish symbolize perseverance and harmony, connecting to Zen concepts of mindfulness and nature.
6	Verdict	Partial (Med. Conf.) — Cultural connections justify partial relevance. (39.8s)
Case 2: False Negative (Over-emphasizing Content Specificity)		
Query: “eiffel tower wall art” Item: “calter 6 piece framed art set” GT: Partial		
#	Agent	Finding
1	Judge	Query requests wall art featuring Eiffel Tower. Need to verify content specificity .
2	Relation Analyzer	Need to analyze query specificity and item content match. Consulting specialists.
3	Query Analyzer	Specific request for Eiffel Tower wall art. Potential variations include photographic prints, paintings, canvas with Paris-themed imagery.
4	Item Analyzer	6-piece contemporary geometric wall art set in black and white. Modern geometric motif design with no Eiffel Tower content .
5	Relation Analyzer	Product matches ‘Wall Art’ category but lacks specific Eiffel Tower imagery . Clear content mismatch.
6	Verdict	Irrelevant (High Conf.) — Lacks Eiffel Tower imagery requested in query. (35.9s)

Figure 9: HIERA Error Cases: False positive and false negative examples