Task-Grounded and EL2N-Informed RAG Switch for Domain-Specific Information Extraction

Xingyu Zhu[1,2], Robert Bossy [1], Balazs Nagy[2], Laszlo Vidacs[2], Claire Nédellec [1]

[1]INRAE, Université Paris-Saclay, Paris, France
[2]Department of Software Engineering, University of Szeged, Szeged, Hungary

xingyu.zhu@inrae.fr

Abstract

Retrieval-augmented generation (RAG) is increasingly popular in information extraction (IE), especially in biological and chemical domains where external knowledge is abundant. Existing RAG-enhanced IE systems, however, typically retrieve for every input, even though retrieval may be unnecessary, introduce distracting evidence, and increase inference cost. To address this issue, we propose RAG Switch, a retrieval decision framework for IE that combines difficulty-aware signals with task-grounded supervision. Specifically, we approximate model-specific training difficulty, quantified by Error L2-Norm (EL2N), from inference-time uncertainty signals, and use it to guide retrieval decisions. We then define retrieval necessity directly from instance-level extraction outcomes without retrieval, yielding task-grounded supervision for learning retrieval decision. Experiments on multiple biological and chemical IE benchmarks across diverse IE backbones show that RAG Switch consistently outperforms representative retrieval decision baselines and achieves a strong efficiency–performance trade-off in end-to-end RAG-enhanced IE pipelines.

1 Introduction

Retrieval-Augmented Generation (RAG) has been widely shown to improve performance across a variety of NLP tasks by grounding generation in relevant external knowledge during inference (Lewis et al., 2020; Wang et al., 2019). In information extraction (IE), this paradigm is particularly appealing for specialized domains such as biology and chemistry, where large-scale structured databases provide abundant domain knowledge that can potentially enhance extraction accuracy. As a result, integrating RAG into IE systems has become an increasingly active research direction (Li et al., 2025; Zhan et al., 2025).

https://github.com/xizhu-hub/RAG-Switch-IE

However, most existing RAG-enhanced IE approaches invoke retrieval indiscriminately for every input instance. Such brute-force retrieval is not always optimal. On the one hand, unnecessary retrieval may introduce distracting or irrelevant evidence, which can even degrade extraction accuracy. On the other hand, it introduces substantial computational overhead and latency, limiting the practicality of RAG-enhanced IE systems. These limitations motivate the need for adaptive retrieval, namely deciding for each individual instance whether retrieval should be triggered.

To alleviate similar inefficiencies, recent work in question answering and knowledge-intensive generation has explored adaptive retrieval mechanisms that dynamically determine whether retrieving external knowledge is necessary for a given input. One representative direction is to introduce an explicit retrieval decision module that can be trained. For example, Adaptive-RAG (Jeong et al., 2024) formulates this problem as query complexity prediction and trains a classifier to route inputs to no-retrieval, single-step retrieval, or multi-step retrieval pipelines. CRAG (Yan et al., 2024) further introduces a lightweight retrieval evaluator that estimates the quality of retrieved evidence and triggers corrective actions when retrieval appears unreliable. Marina et al. (2025) eliminate the need for LLM-based introspection by training a feature-based classifier, enabling high-efficiency, model-agnostic retrieval routing.

At the same time, a growing body of work has moved away from separately trained retrieval decision modules and instead relies on training-free LLMs operating directly at inference time. For example, RetrievalQA (Zhang et al., 2024) directly prompts LLMs to make a yes/no judgment about whether retrieval is required before answering a query. Self-RAG (Asai et al., 2023) internalizes this decision into the generation process by introducing special reflection tokens whose predicted

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 15319–15336
October 24-29, 2026 ©2026 Association for Computational Linguistics

probabilities govern whether retrieval is triggered during decoding. DRAGIN (Su et al., 2024) detects real-time information needs from token uncertainty and contextual salience, while FLARE (Jiang et al., 2023) triggers retrieval when forward predictions contain low-confidence tokens.

While these methods have shown promise in broader NLP settings, explicit adaptive retrieval for IE remains relatively underexplored compared with question answering and knowledge-intensive generation (Zhan et al., 2025; Li et al., 2025). More importantly, when transferred to IE pipelines, existing adaptive retrieval approaches still suffer from two key limitations.

First, current methods lack a clear and task-grounded definition of when additional information retrieval is beneficial for IE tasks. In many cases, the utility of retrieval decisions is evaluated only through the final performance of an end-to-end RAG-enhanced IE pipeline. However, this overall performance is jointly affected by multiple factors, including retriever quality, evidence relevance, and the behavior of the IE model. As a result, the contribution of the retrieval decision itself is difficult to isolate, making it hard to derive targeted supervision for retrieval decision learning and to evaluate retrieval decisions in a principled manner.

Second, many adaptive retrieval approaches rely on large language models (LLMs) to directly judge retrieval necessity. Such strategies not only incur additional cost from API usage, but also raise concerns about reliability. Prior work (Barkan et al., 2025; Zhang et al., 2025) has shown that LLMs may exhibit systematic overconfidence in self-evaluation, which already casts doubt on the trustworthiness of LLM-based judgments. This concern becomes even more pronounced when an external LLM is used to determine whether retrieval is necessary for a target IE model. In such cases, the resulting decision may not reflect the target model's actual difficulty profile and knowledge requirements.

To address these limitations, we propose RAG Switch, a retrieval decision framework tailored to IE pipelines, particularly named entity recognition (NER) and relation extraction (RE) settings. Our central idea is to replace heuristic or LLM-dependent retrieval decisions with signals that are directly grounded in the target task and the target IE model.

Specifically, we first introduce a task-grounded supervision signal for retrieval utility in IE. To isolate retrieval utility from retrieval quality and strategy, we define it under an idealized assumption used only for label construction, where retrieval does not negatively affect the extraction performance of the underlying IE pipeline, but may yield no improvement in the worst case. Retrieval is then regarded as useful for an instance when the underlying IE pipeline cannot correctly solve it without retrieval, as measured by instance-level F1. This formulation yields supervision that is directly aligned with the target extraction objective, enabling more targeted training and principled evaluation of the switch.

In addition, we introduce EL2N as an auxiliary signal for retrieval decision making. Retrieval is most useful when the target IE model cannot reliably solve an instance using its own internal knowledge. This makes the difficulty in model-specific instance-level training, as quantified by EL2N, a natural signal to decide whether external knowledge is needed. However, EL2N is typically observable only through training dynamics and therefore cannot be directly accessed during inference. To bridge this gap, we first perform a no-retrieval inference pass to obtain uncertainty-related signals from the IE model, including prediction confidence, entropy, and margin statistics, and then use these signals to approximate EL2N at inference time.

Our contributions are summarized as follows:

1. We propose a task-grounded supervision strategy for retrieval decisions in IE by defining retrieval necessity directly from instance-level extraction outcomes achieved without retrieval. This supervision is directly aligned with the target extraction objective, making retrieval decisions easier to train and evaluate in a principled manner.

2. We introduce EL2N as a model-specific training difficulty signal for retrieval decision making, and approximate it at inference time using uncertainty-related signals. To the best of our knowledge, prior work has not explicitly approximated training-time difficulty from inference-time uncertainty signals as a proxy for retrieval decision making.

3. Experiments across multiple biological and chemical IE datasets and backbones show that RAG Switch consistently outperforms representative retrieval decision baselines, maintains a strong trade-off between efficiency and performance, and exhibits strong generalizability across diverse IE pipeline settings.

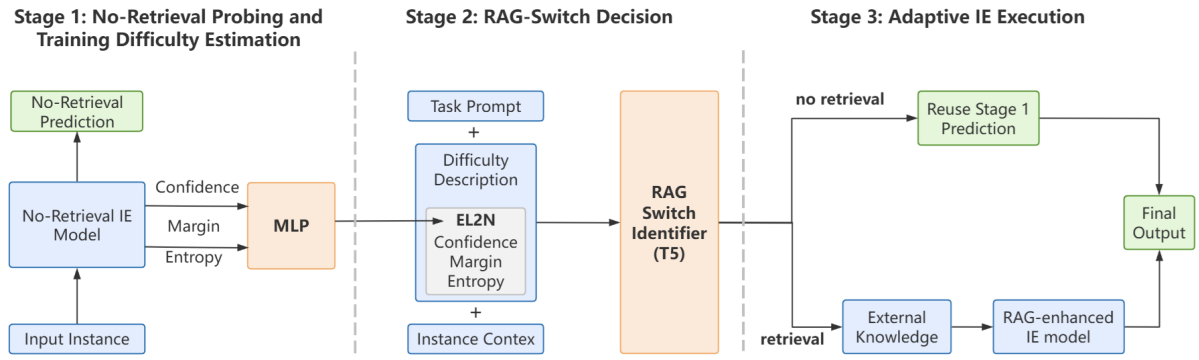


Figure 1: Overview of RAG Switch integrated into an RAG-enhanced IE inference pipeline. The framework consists of three stages. In Stage 1, a no-retrieval IE model produces a no-retrieval prediction and difficulty-aware signals, from which an MLP predicts an EL2N-based difficulty score. In Stage 2, a T5-based identifier takes the task prompt, difficulty description, and instance context to determine whether retrieval is needed. In Stage 3, the system either directly reuses the Stage-1 prediction or routes the instance to the retrieval-augmented IE path.

2 Related Work

Recent work on adaptive retrieval has mainly focused on question answering and knowledge-intensive generation tasks, aiming to balance task performance and retrieval cost by dynamically deciding when retrieval should be triggered. One line of work trains an explicit decision component for retrieval control. CRAG (Yan et al., 2024) introduces a lightweight evaluator to estimate whether retrieved evidence is reliable enough to support downstream prediction and triggers corrective actions when the evidence appears unreliable. RouteRAG (Guo et al., 2025) further extends this idea by learning a retrieval policy under efficiency constraints, so that retrieval control can be optimized jointly with system cost. Adaptive-RAG (Jeong et al., 2024) formulates retrieval control as a supervised routing problem, learning a query-complexity classifier that maps inputs to no-retrieval, single-step retrieval, or multi-step retrieval modes. Marina et al. (2025) decouple retrieval routing from LLM states, using a lightweight classifier based on extrinsic query features to achieve efficient, model-agnostic decision-making. These methods show that retrieval decisions can be learned explicitly to balance retrieval cost and task performance.

Another line of work does not train a separate decision component, but instead relies on the same LLM that performs the target NLP task to estimate retrieval necessity. RetrievalQA (Zhang et al., 2024) prompts an LLM to output a yes/no retrieval judgment before answering, while Self-RAG (Asai et al., 2023) internalizes retrieval control into decoding through reflection tokens. Related methods such as FLARE (Jiang et al., 2023) and DRAGIN (Su et al., 2024) further use generation-time confidence or uncertainty signals to trigger retrieval during generation. These methods avoid training an additional controller and suggest that model-side uncertainty signals can aid retrieval decisions.

Despite these advances, existing adaptive retrieval methods are not designed directly for IE, especially for NER and RE. In particular, when transferred to IE pipelines, they still leave open two issues central to our setting: how to define retrieval supervision in a way that is directly grounded in extraction outcomes, and how to avoid relying on retrieval judgments that are not directly tied to the behavior of the target IE model. Our method addresses these gaps by constructing supervision from no-retrieval extraction outcomes and introducing model-specific difficulty-aware signals from the target IE model.

3 Methodology

RAG Switch is built around two components for retrieval decision making in IE: model-specific training difficulty as the decision signal, and task-grounded supervision from no-retrieval extraction outcomes as the decision objective. Figure 1 shows the overall framework and its integration into an IE pipeline with RAG.

3.1 Training Difficulty as an Auxiliary Signal

We introduce a model-specific training difficulty as an auxiliary signal for retrieval decision making.

To operationalize this notion, we seek a difficulty measure that is specific to the target IE model itself. A natural reference for this purpose is *Error L2-Norm* (EL2N) (Paul et al., 2021; Sabbineni et al., 2023), which measures the discrepancy between the model prediction and the training target at the instance level. In our setting, EL2N is particularly suitable because it provides an instance-level estimate of how difficult a sample is for the target IE model to fit during training.

Given an input instance x, let the IE model produce a sequence of predictive distributions $p_1, p_2, \ldots, p_T$, where $p_t \in \mathbb{R}^{K_t}$ denotes the predictive distribution at prediction step t, K_t is the size of the corresponding prediction space, and T is the number of prediction steps associated with the instance. Let q_t denote the target distribution at step t during training. We use the following instance-level EL2N score:

$$\mathrm{EL2N}(x) = \frac{1}{T}\sum_{t=1}^{T} \|p_t - q_t\|_2^2 .$$

A higher EL2N indicates a greater discrepancy between the predictive distributions and the target output, and therefore a higher training difficulty for the instance with respect to the IE model.

However, EL2N depends on the training target q_t and is therefore not available at inference time, when retrieval decisions must be made. To address this, we consider three inference-time uncertainty signals computed directly from the IE model's predictive distribution: *confidence*, *margin*, and *entropy*. These signals characterize complementary aspects of the predictive distribution and are closely related to the error structure captured by EL2N.

For each prediction step t, let

$$p_t = (p_{t,1}, \ldots, p_{t,K_t}).$$

The confidence is the maximum predicted probability,

$$c_t = \max_j p_{t,j},$$

the largest margin is the gap between the two highest probabilities,

$$m_t = p_{t,(1)} - p_{t,(2)},$$

and the predictive entropy is

$$H_t = -\sum_{j=1}^{K_t} p_{t,j} \log p_{t,j},$$

where $p_{t,(1)}$ and $p_{t,(2)}$ are the largest and second-largest values in p_t, respectively.

To explain why these signals can serve as informative proxies for EL2N, we first consider the common one-hot target case, where the relationship between prediction uncertainty and instance-level error becomes especially transparent. We extend the analysis to the more general target distribution setting in Appendix A.

Under the one-hot target case, let y_t denote the correct target index at step t, so that q_t places all probability mass on y_t. The step-level contribution to EL2N can be decomposed as

$$\|p_t - q_t\|_2^2 = (1 - p_{t,y_t})^2 + \sum_{j \neq y_t} p_{t,j}^2. \quad (1)$$

This decomposition shows that the prediction error is jointly determined by the probability mass missing from the correct target and by how the remaining mass is distributed over incorrect alternatives.

When the correct target is also the highest prediction, i.e., $p_{t,y_t} = c_t$, one can derive the bound

$$(1 - c_t)^2 \frac{K_t}{K_t - 1} \;\le\; \|p_t - q_t\|_2^2 \;\le\; 2(1 - c_t)^2. \quad (2)$$

This shows that in this regime EL2N is tightly controlled by $1 - c_t$.

When the highest prediction is incorrect, confidence alone is no longer sufficient. In this case, one can show that

$$\|p_t - q_t\|_2^2 \ge (1 - c_t + m_t)^2 + c_t^2, \quad (3)$$

which indicates that the error also depends on the local competition structure around the correct target, captured by the margin. Even when confidence and margin are fixed, the remaining probability mass can still be distributed differently across alternatives, so entropy provides complementary information about the global spread of the predictive distribution. Detailed derivations of the above bounds, together with the extension to general target distributions, are given in the Appendix A.

We aggregate the step-level uncertainty statistics over the full instance by averaging across prediction steps:

$$c(x),\ m(x),\ H(x) = \frac{1}{T}\sum_{t=1}^{T} (c_t,\ m_t,\ H_t).$$

Based on these aggregated uncertainty signals, we form the three-dimensional feature vector

$$z(x) = [\, c(x),\, m(x),\, H(x) \,],$$

and use a two-layer MLP to estimate the corresponding EL2N score:

$$\hat{d}(x) = \mathrm{MLP}(z(x)),$$

where $\hat{d}(x)$ denotes the predicted EL2N difficulty of instance x. The predictor is trained with mean squared error loss:

$$\mathcal{L}_{\text{diff}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{d}(x_i) - \mathrm{EL2N}(x_i) \Big)^2 .$$

Based on the predicted EL2N $\hat{d}(x)$, we obtain an inference-time proxy of training difficulty from a single forward pass of the IE model. Together with the three uncertainty signals themselves, this predicted difficulty forms the difficulty-aware signals used by the RAG Switch.

3.2 Task-grounded Supervision for Retrieval Switching

The identifier is the final decision component in RAG Switch that predicts whether retrieval should be triggered for a given input instance. To train such a module, we require supervision that directly reflects whether triggering retrieval is likely to be beneficial for the target instance. Rather than relying on indirect heuristics, we construct a task-grounded supervision signal from the behavior of the upstream no-retrieval pipeline itself.

Specifically, according to the operational formulation introduced in Section 1, retrieval is considered beneficial for an instance if the IE pipeline cannot solve it correctly *without* retrieval. Let $\mathrm{F1}(x)$ denote the instance-level F1-score achieved by the no-retrieval IE pipeline on instance x. We then define the supervision label for the identifier as

$$y^{\text{id}}(x) = \begin{cases} 0, & \text{if } \mathrm{F1}(x) = 1, \\ 1, & \text{otherwise.} \end{cases}$$

Here, $y^{\text{id}}(x) = 1$ indicates that triggering retrieval is beneficial for the instance under this criterion, whereas $y^{\text{id}}(x) = 0$ indicates that the no-retrieval pipeline already achieves the target extraction outcome. This definition grounds retrieval supervision directly in the target IE objective, and has the practical advantage that it does not depend on any particular retrieval database or strategy.

With these task-grounded labels, we instantiate the identifier with a T5 model and train it in a text-to-text manner, given T5's strong balance between reasoning capability and inference efficiency. Each input consists of three components: a task prompt, a difficulty description, and the instance context. The task prompt defines the retrieval decision objective and explains the role of the difficulty-aware signals introduced in Section 3.1, including confidence, margin, entropy, and the predicted EL2N. The difficulty description presents these signals in a structured *metric: value* format. The instance context provides the semantic information needed to interpret these signals. Let $u(x)$ denote the textualized difficulty description of instance x, and let $\mathrm{ctx}(x)$ denote its instance context. The identifier is trained to generate one of two fixed responses, *Yes, it is difficult* or *No, it is not difficult*, which correspond to labels 1 and 0, respectively. The training objective is the standard sequence-to-sequence negative log-likelihood:

$$\mathcal{L}_{\text{id}} = -\sum_{i=1}^{N} \log p_\theta(r_i \mid \mathrm{Prompt}(u(x_i), \mathrm{ctx}(x_i))) ,$$

where r_i is the target textual response for instance x_i, and θ denotes the identifier parameters.

Based on the task-grounded supervision and the structured input above, the T5 identifier can use its attention mechanism to jointly model the interaction between the difficulty cues and the instance context, and thereby learn to predict whether retrieval should be triggered for the current instance.

4 Experiment

4.1 Experimental Setting

We evaluate RAG Switch on IE pipelines built on three representative backbone architectures: LLaMA2-7B (Touvron et al., 2023), T5-large (Raffel et al., 2020), and BERT-large (Devlin et al., 2019), representing decoder-only, encoder–decoder, and encoder-based models, respectively. Experiments are conducted on three biological and chemical IE benchmarks: Florilege (Falentin et al., 2017), CHEMDNER (Krallinger et al., 2015), and ChemProt (Taboureau et al., 2010). These domains are a natural testbed for retrieval-augmented IE because rich external knowledge is available, while the need for retrieval still varies substantially

across instances. Florilege focuses on microbial ecology in food science and contains structurally complex entity mentions and relations. CHEMDNER is a chemical named entity recognition benchmark derived from PubMed abstracts, requiring fine-grained categorization of chemical mentions. ChemProt is a biological relation extraction dataset centered on chemical–protein interactions.

We compare the performance of RAG Switch with three representative retrieval decision modules adapted from prior methods: RetrievalQA (ReQA) (Zhang et al., 2024), Adaptive-RAG (AdptRAG) (Jeong et al., 2024), and Self-RAG (Asai et al., 2023). Although originally proposed for different settings, these methods all include mechanisms for determining whether to retrieve and can be adapted to IE tasks. In our comparison, we focus specifically on their retrieval decision mechanisms rather than the full original methods. RetrievalQA directly prompts an LLM to make a yes/no retrieval judgment, Adaptive-RAG learns a routing model for retrieval decisions, and Self-RAG integrates retrieval control into generation through special tokens. For fair comparison, we adapt these retrieval decision modules to the IE setting and evaluate them under the same supervision data and evaluation protocol.

In addition, we implement an MLP-based identifier using the same difficulty-aware signals introduced in Section 3.1, including confidence, margin, entropy, and the predicted EL2N-based training difficulty. This variant directly predicts retrieval decisions from these features and serves as a lightweight baseline to examine whether simple decision boundaries over these signals are sufficient.

To evaluate retrieval decision quality, we report Balanced Accuracy (BalAcc), Yes-F1, and No-F1. BalAcc is used as the primary evaluation metric, as it averages recall over positive and negative classes and is robust to class imbalance. Besides, Yes-F1 measures performance on the retrieval-beneficial class, while No-F1 measures performance on the retrieval-unnecessary class. Additional implementation details are provided in Appendix B.

4.2 Main Result

Table 1 reports the main results of different retrieval decision methods across three datasets and three IE backbones. Overall, RAG Switch achieves the highest Balanced Accuracy in all settings, showing that the proposed retrieval decision mechanism generalizes well across heterogeneous IE pipelines. Averaged over datasets, the gain in BalAcc is 10.36 points for LLaMA, 6.68 points for BERT, and 6.80 points for T5.

More concretely, compared with the fine-tuned retrieval decision baseline Adaptive-RAG, our approach yields consistent improvements across all dataset–backbone combinations, with the largest single improvement reaching 23.88 BalAcc points on Florilege in the T5-based setting. We attribute this to the additional difficulty-aware signals used in RAG Switch. Unlike routing decisions derived only from the input query, they directly reflect how difficult an instance is for the target IE model, substantially improving decision quality.

The gap is even larger for the two training-free LLM-based baselines, RetrievalQA and Self-RAG. Across datasets, these methods typically obtain relatively low Yes-F1 while maintaining much higher No-F1, especially in the T5- and LLaMA-based settings. This imbalance indicates that these approaches tend to over-predict the no-retrieval case, failing to reliably identify instances where retrieval is actually required. A likely reason is that the proportion of retrieval-beneficial instances varies substantially across IE backbones: in ChemProt and CHEMDNER, it is around 60% for the BERT-based pipeline but typically only 10–20% for the T5- and LLaMA-based pipelines. As a result, no-retrieval cases dominate for stronger backbones, which also helps explain why LLM-based decision approaches such as RetrievalQA and Self-RAG tend to favor the no-retrieval prediction. More broadly, this also suggests that generic LLM-based judgments are less well-aligned with the retrieval needs of the target IE model.

We also compare two variants of the identifier: a T5-based version and an MLP-based version. Both outperform external baselines in most settings, while the T5-based version performs consistently better, indicating the benefit of modeling interactions between structured difficulty cues and the instance context.

Overall, the results indicate that adaptive retrieval for IE benefits from combining task-grounded supervision with model-specific difficulty signals, enabling the decision identifier to better capture when the target IE model is likely to benefit from external knowledge.

4.3 Analysis of Difficulty Signals

To examine the contribution of different difficulty-aware signals, we remove each component from

Dataset	Method	LLaMA			BERT			T5		
		Yes-F1	No-F1	BalAcc	Yes-F1	No-F1	BalAcc	Yes-F1	No-F1	BalAcc
CHEMDNER	ReQA (LLaMA2-7B)	10.55	89.39	49.52	2.36	50.92	50.10	3.37	86.94	50.35
	ReQA (LLaMA2-13B)	14.68	60.20	51.17	6.53	50.30	49.72	4.39	87.02	50.60
	ReQA (GPT-4.1)	16.40	83.55	50.89	15.28	52.25	53.03	10.88	87.48	52.53
	SelfRAG (LLaMA2-7B)	12.39	68.94	47.30	–	–	–	–	–	–
	AdptRAG (T5-large)	51.75	52.40	65.56	82.28	29.39	58.61	36.41	86.37	60.16
	RAG Switch (MLP)	62.51	62.24	68.54	69.29	56.94	65.38	59.93	86.48	75.21
	RAG Switch (T5-large)	68.54	90.11	76.51	82.29	61.40	71.09	70.17	90.46	81.86
ChemProt	ReQA (LLaMA2-7B)	8.73	90.91	49.91	3.40	54.64	49.33	1.62	90.11	48.57
	ReQA (LLaMA2-13B)	17.27	51.68	49.36	9.25	52.58	47.62	7.65	91.38	51.29
	ReQA (GPT-4.1)	13.16	86.98	50.52	12.71	55.17	50.98	11.92	84.78	48.35
	SelfRAG (LLaMA2-7B)	14.23	64.37	45.79	–	–	–	–	–	–
	AdptRAG (T5-large)	25.35	92.71	58.19	77.42	39.86	59.84	28.56	90.32	58.31
	RAG Switch (MLP)	47.46	80.56	60.23	30.68	61.58	60.47	50.40	89.84	72.58
	RAG Switch (T5-large)	43.80	89.69	71.52	80.86	53.29	66.73	60.61	91.62	80.08
Florilege	ReQA (LLaMA2-7B)	9.15	60.62	51.73	1.65	53.81	45.10	2.79	60.87	50.06
	ReQA (LLaMA2-13B)	22.92	59.43	52.31	12.33	56.63	49.66	8.12	61.33	51.18
	ReQA (GPT-4.1)	1.37	60.19	50.35	5.54	56.94	49.17	16.73	61.29	52.25
	SelfRAG (LLaMA2-7B)	0.59	60.03	50.10	–	–	–	–	–	–
	AdptRAG (T5-large)	43.32	62.52	62.39	74.62	34.90	57.41	31.83	65.97	59.46
	RAG Switch (MLP)	40.08	67.37	61.39	72.30	34.39	55.43	84.13	70.66	76.92
	RAG Switch (T5-large)	80.65	66.75	73.21	78.26	48.28	64.05	86.87	80.79	83.34

Table 1: Retrieval decision module performance comparison across datasets and backbones.

Dataset	Method	BERT	T5	LLaMA
CHEMDNER	All	**71.09**	**81.86**	**76.51**
	-Margin	70.45	80.21	75.26
	-EL2N	67.22	78.53	73.49
	-Confidence	69.90	80.62	73.88
	-Entropy	70.64	81.50	75.53
ChemProt	All	**66.73**	**80.08**	**71.52**
	-Margin	66.52	78.69	71.41
	-EL2N	63.40	76.18	67.33
	-Confidence	64.91	77.04	69.08
	-Entropy	66.06	79.50	71.10
Florilege	All	**64.05**	**83.34**	73.21
	-Margin	63.59	82.44	**73.66**
	-EL2N	61.14	79.86	69.90
	-Confidence	62.10	81.20	71.75
	-Entropy	63.38	82.70	72.42

Table 2: Ablation study on difficulty-aware signals. Results are reported in Balanced Accuracy (BalAcc).

Backbone	Signal	p_r	s_ρ	AUC	mw_p
BERT	EL2N	0.404	**0.490**	**0.792**	**0**
	Margin	0.335	0.389	0.697	5e-247
	Confidence	**0.421**	0.448	0.723	3e-257
	Entropy	0.339	0.362	0.682	4e-231
T5	EL2N	0.427	**0.475**	**0.833**	**0**
	Margin	**0.494**	0.471	0.821	**0**
	Confidence	0.478	0.464	0.820	**0**
	Entropy	0.411	0.465	0.821	**0**
LLaMA	EL2N	**0.564**	**0.407**	**0.888**	**8e-162**
	Margin	0.480	0.380	0.861	3e-140
	Confidence	0.444	0.402	0.886	1e-161
	Entropy	0.440	0.375	0.858	2e-137

Table 3: Correlation between difficulty-aware signals and retrieval decision labels using last-epoch statistics on CHEMDNER as an example. p_r and s_ρ are reported in **absolute** value.

the full feature set and report Balanced Accuracy.

As shown in Table 2, removing any one of the four signals leads to performance degradation in almost all settings, indicating that each provides useful information for retrieval decision making. Among them, removing EL2N leads to the largest degradation in most settings, showing that explicitly approximating model-specific training difficulty provides clear benefit beyond using the original uncertainty signals alone. At the same time, confidence, margin, and entropy remain complementary across datasets and backbones. Overall, these results show that effective retrieval decisions benefit from both inference-time uncertainty signals and an explicit difficulty proxy for the target IE model.

We further analyze the relationship between each signal and the retrieval decision label using Pearson correlation (p_r), Spearman rank correlation (s_ρ), AUC, and the Mann–Whitney U test p-value (mw_p), where all statistics are computed from the last-epoch signals used at inference time. As shown in Table 3, all four signals exhibit high discriminative power and statistically significant associations with the labels across backbones, as indicated by their AUC values and extremely small mw_p. As expected, EL2N shows the strongest association in most settings, while other signals

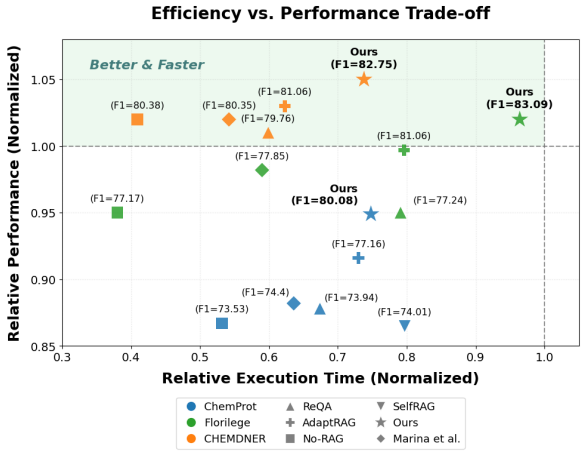


Figure 2: Efficiency–performance trade-off analysis across various tasks. The dashed horizontal and vertical lines represent the performance and execution time of the Full-RAG baseline ($1.0\times$), respectively. Numbers in parentheses indicate the absolute F1 scores.

also maintain robust relationships with the labels. This pattern is consistent with the ablation results and further supports the use of EL2N, showing that mapping inference-time uncertainty signals to model-specific training difficulty yields an especially informative proxy for retrieval decision making. More correlation details are shown in Appendix C.

4.4 End-to-End Trade-off Evaluation

To evaluate the practical impact of RAG Switch, we further conduct end-to-end experiments on full RAG-enhanced IE pipelines across multiple datasets. For relation extraction, we follow BiomedRAG (Li et al., 2025) and build the pipeline on the ChemProt dataset, using LLaMA2-7B model. For entity recognition on CHEMDNER, we follow BIOREADER (Frisoni et al., 2022) and build the pipeline on T5-Large model. These settings are chosen to ensure comparability with their original configurations and experiments. We also consider entity recognition on Florilege, where we construct our own end-to-end entity recognition pipeline based on T5-Large, using SilverCorpus (Zhu et al., 2026) as the retrieval source. We additionally include Marina et al. (2025) as a baseline for comparison, as their approach is recognized for its high computational efficiency. More details are provided in the Appendix D.

We report relative inference time over the entire test set and relative performance (F1-score), both normalized to the Full-RAG setting. As shown in Figure 2, the dashed vertical and horizontal lines divide the performance–efficiency space according to the Full-RAG baseline. Points to the left of the vertical line require less computation than Full-RAG, while points above the horizontal line achieve higher performance. Under this view, our method consistently occupies a favorable region of the trade-off space across IE tasks, and on the two entity recognition datasets it lies in the upper-left quadrant, indicating that it outperforms Full-RAG while using less inference time.

On ChemProt relation extraction, our method uses only about half of the additional time introduced by Full-RAG, while still retaining nearly 95% of its performance. On both entity recognition tasks, our method outperforms Full-RAG. On Florilege, it achieves this improvement with an inference cost close to the Full-RAG setting. On CHEMDNER, the benefit is even more substantial, improving over Full-RAG by 3.84 F1 points while requiring only about 74% of the Full-RAG inference cost. We attribute this advantage to the higher retrieval decision quality of RAG Switch on both retrieval-beneficial and retrieval-unnecessary instances. Although our method introduces an additional forward pass to compute difficulty-aware signals at inference time, this overhead is effectively offset by the reduction in unnecessary retrieval calls. At the same time, more accurate identification of retrieval-beneficial cases helps preserve the gains brought by external knowledge, while filtering out retrieval-unnecessary cases reduces noise introduced by irrelevant retrieved contexts.

5 Conclusion

We presented RAG Switch, a retrieval decision framework for adaptive retrieval in information extraction. By defining retrieval necessity from instance-level extraction outcomes and approximating model-specific training difficulty with a predicted EL2N score derived from inference-time uncertainty, RAG Switch provides a task-grounded way to decide whether retrieval should be triggered for each instance. Experiments on multiple biological and chemical IE benchmarks across diverse backbones showed that our method consistently improves retrieval decision quality over representative baselines, while end-to-end results further demonstrated a strong efficiency–performance trade-off in RAG-enhanced IE pipelines. Beyond the IE setting studied here, these results also suggest that approxi-

mating training-time difficulty from inference-time signals may provide a useful direction for other adaptive inference problems that require model-specific difficulty awareness.

Limitations

RAG Switch requires additional offline computation during training to construct task-grounded retrieval labels. Specifically, these labels are obtained from out-of-fold predictions of the no-retrieval IE pipeline, which introduces an extra preparation step before the identifier can be trained. This design helps prevent data leakage in label construction, at the cost of additional training-side computation.

In addition, while RAG Switch consistently improves end-to-end efficiency over Full-RAG in our experiments, the magnitude of this advantage depends on several factors in the target IE pipeline, including the proportion of retrieval-beneficial instances, the cost of retrieval itself, and the additional inference overhead introduced after retrieval is incorporated. We discuss this boundary case in more detail and provide corresponding practical recommendations in the Appendix E.

Acknowledgments

This work was supported by the FAIROmics project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101120449.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-Reflective Retrieval Augmented Generation. In *NeurIPS 2023 workshop on instruction tuning and instruction following*.

Casey O Barkan, Sidney Black, and Oliver Sourbut. 2025. Do Large Language Models Know What They Are Capable Of? In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Hélène Falentin, Estelle Chaix, Bedis Dridi, Philippe Bessieres, Solange Buchin, Stéphanie-Marie Deutsch, Magalie Weber, Robert Bossy, Sandra Derozier, Bruno Perret, and 1 others. 2017. Florilege: A Database Gathering Microbial Phenotypes of Food Interest. In *4th International Conference on Microbial Diversity 2017*.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. BIOREADER: A Retrieval-enhanced Text-to-text Transformer for Biomedical Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793.

Yucan Guo, Miao Su, Saiping Guan, Zihao Sun, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025. RouteRAG: Efficient Retrieval-Augmented Generation from Text and Graph via Reinforcement Learning. *arXiv preprint arXiv:2512.09487*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The Drugs and Chemical Names Extraction Challenge. *Journal of cheminformatics*, 7(Suppl 1):S1.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in neural information processing systems*, 33:9459–9474.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. 2025. BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine. *Journal of Biomedical Informatics*, 162:104769.

Maria Marina, Nikolay Ivanov, Sergey Pletenev, Mikhail Salnikov, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. LLM-Independent Adaptive RAG: Let the Question Speak for Itself. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8697–8709.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep Learning on a Data Diet:

Finding Important Examples Early in Training. *Advances in Neural Information Processing Systems*, 34:20596–20607.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *Journal of machine learning research*, 21(140):1–67.

Anusha Sabbineni, Nikhil Anand, and Maria Minakova. 2023. Comprehensive benchmarking of entropy and margin based scoring metrics for data selection. *arXiv preprint arXiv:2311.16302*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic Retrieval Augmented Generation Based on the Real-time Information Needs of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013.

Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, and 1 others. 2010. ChemProt: A Disease Chemical Biology Database. *Nucleic acids research*, 39(suppl_1):D367–D372.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and 1 others. 2019. Improving Natural Language Inference using External Knowledge in the Science Questions Domain. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7208–7215.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. *arXiv preprint arXiv:2401.15884*.

Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. RAMIE: Retrieval-Augmented Multi-Task Information Extraction with Large Language Models on Dietary Supplements. *Journal of the American Medical Informatics Association*, 32(3):545–554.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025. Generative Verifiers: Reward Modeling as Next-Token Prediction. In *International Conference on Learning Representations*, volume 2025, pages 12476–12505.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-Form Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975.

Xingyu Zhu, Claire Nédellec, and Robert Bossy. 2026. Silver Corpus for Plant-based Food Fermentation Information Extraction. Recherche Data Gouv. Draft version. DOI: https://doi.org/10.57745/MI7Q9W.

A Analysis of EL2N and Inference-Time Uncertainty Signals

In this appendix, we analyze the relationship between EL2N and the inference-time uncertainty signals used in our method. We first consider the common one-hot target case, where the connection between prediction uncertainty and instance-level error is especially transparent, and then extend the analysis to the more general setting where the target at each prediction step is an arbitrary probability distribution.

A.1 One-Hot Target Case

Assume that the target distribution at prediction step t is one-hot, i.e., q_t places all probability mass on the correct label y_t. Then the step-level EL2N term can be decomposed as

$$\|p_t - q_t\|_2^2 = (1 - p_{t,y_t})^2 + \sum_{j \neq y_t} p_{t,j}^2. \quad (4)$$

Equation 4 shows that the prediction error is determined by two components: the probability mass missing from the correct target, and the way the remaining mass is distributed over incorrect alternatives.

Case 1: the correct target is also the top prediction. We first consider the regime where the correct target is also the top prediction, i.e., $p_{t,y_t} = c_t$. From Equation 4, the step-level EL2N term becomes

$$\|p_t - q_t\|_2^2 = (1 - c_t)^2 + \sum_{j \neq y_t} p_{t,j}^2. \quad (5)$$

Since the probabilities over all incorrect alternatives sum to the residual mass,

$$\sum_{j \neq y_t} p_{t,j} = 1 - c_t, \quad (6)$$

the squared residual term is bounded as follows.

For the upper bound, since all terms are nonnegative,

$$\sum_{j \neq y_t} p_{t,j}^2 \leq \left(\sum_{j \neq y_t} p_{t,j} \right)^2 = (1 - c_t)^2. \quad (7)$$

For the lower bound, under a fixed total mass $\sum_{j \neq y_t} p_{t,j} = 1 - c_t$, the sum of squares is minimized when the residual mass is uniformly distributed over the $K_t - 1$ incorrect alternatives. Hence,

$$\sum_{j \neq y_t} p_{t,j}^2 \geq (K_t - 1) \left(\frac{1 - c_t}{K_t - 1} \right)^2 = \frac{(1 - c_t)^2}{K_t - 1}. \quad (8)$$

Substituting Equations 7 and 8 into Equation 5, we obtain

$$\frac{K_t}{K_t - 1}(1 - c_t)^2 \;\leq\; \|p_t - q_t\|_2^2 \;\leq\; 2(1 - c_t)^2, \quad (9)$$

which shows that when the correct target is also the top prediction, EL2N is tightly controlled by $1 - c_t$.

Case 2: the top prediction is incorrect. We now consider the regime where the top prediction is incorrect, so that $p_{t,y_t} \neq c_t$. In this case, the correct target cannot exceed the second-largest probability:

$$p_{t,y_t} \leq p_{t,(2)} = c_t - m_t. \quad (10)$$

Therefore,

$$1 - p_{t,y_t} \geq 1 - (c_t - m_t) = 1 - c_t + m_t, \quad (11)$$

and squaring both sides yields

$$(1 - p_{t,y_t})^2 \geq (1 - c_t + m_t)^2. \quad (12)$$

On the other hand, since the top prediction is incorrect, the top probability c_t is itself part of the residual error term in Equation 4. Hence,

$$\sum_{j \neq y_t} p_{t,j}^2 \geq c_t^2. \quad (13)$$

Substituting Equations 12 and 13 into Equation 4, we obtain

$$\begin{aligned} \|p_t - q_t\|_2^2 &= (1 - p_{t,y_t})^2 + \sum_{j \neq y_t} p_{t,j}^2 \\ &\geq (1 - c_t + m_t)^2 + c_t^2, \end{aligned} \quad (14)$$

which shows that when the top prediction is wrong, confidence alone is insufficient and the local competition structure captured by the margin becomes crucial.

Why entropy is still necessary. The one-hot analysis above explains why confidence acts as a primary anchor of the prediction error and why margin becomes important when the top prediction is incorrect. However, confidence and margin do not fully determine the second term in Equation 4. Even when the top score and the top-two gap are

fixed, the remaining probability mass can still be distributed in very different ways across the incorrect alternatives, leading to different EL2N values. Entropy captures this global spread of probability mass and therefore provides complementary information that is not recoverable from confidence and margin alone.

In summary, the one-hot case already reveals the main intuition behind our proxy design: confidence captures the dominant alignment with the target, margin refines the local competition structure when the top prediction is unreliable, and entropy reflects the global dispersion of the predictive distribution.

A.2 Extension to General Target Distributions

We now extend the analysis to the more general setting where the target at step t is an arbitrary probability distribution

$$q_t = (q_{t,1}, \dots, q_{t,K_t}) \in \Delta^{K_t - 1}.$$

Let

$$p_t = (p_{t,1}, \dots, p_{t,K_t})$$

denote the predictive distribution. Define

$$j_t^* = \arg\max_j q_{t,j},$$
$$q_t^{(1)} = q_{t,j_t^*}, \qquad q_t^{(2)} = \max_{j \neq j_t^*} q_{t,j}.$$

where $q_t^{(1)}$ and $q_t^{(2)}$ are the largest and second-largest target probabilities, respectively.

For a general target distribution, the step-level EL2N term becomes

$$\begin{aligned} \|p_t - q_t\|_2^2 &= \sum_j (p_{t,j} - q_{t,j})^2 \\ &= \|p_t\|_2^2 + \|q_t\|_2^2 - 2\langle p_t, q_t \rangle. \end{aligned} \tag{15}$$

Equation 15 shows that EL2N depends both on the overall shape of the predictive distribution and on its alignment with the target distribution. This already suggests why inference-time uncertainty signals can serve as informative proxies: confidence and margin constrain the local shape of p_t, while entropy captures its global dispersion.

Case 1: the dominant target index is also the top prediction. Assume that the model's top prediction agrees with the dominant target index, i.e.,

$$p_{t,j_t^*} = c_t.$$

Then

$$\|p_t - q_t\|_2^2 = (c_t - q_t^{(1)})^2 + \sum_{j \neq j_t^*} (p_{t,j} - q_{t,j})^2. \tag{16}$$

The residual terms satisfy

$$\sum_{j \neq j_t^*} p_{t,j} = 1 - c_t, \qquad \sum_{j \neq j_t^*} q_{t,j} = 1 - q_t^{(1)}.$$

Hence,

$$\begin{aligned} \sum_{j \neq j_t^*} (p_{t,j} - q_{t,j}) &= (1 - c_t) - (1 - q_t^{(1)}) \\ &= q_t^{(1)} - c_t. \end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\sum_{j \neq j_t^*} (p_{t,j} - q_{t,j})^2 \geq \frac{(q_t^{(1)} - c_t)^2}{K_t - 1}.$$

Substituting this into Equation 16 gives

$$\|p_t - q_t\|_2^2 \geq (c_t - q_t^{(1)})^2 \left(1 + \frac{1}{K_t - 1}\right). \tag{17}$$

For an upper bound, we use

$$(a - b)^2 \leq a^2 + b^2 \quad \text{for } a, b \geq 0,$$

which yields

$$\sum_{j \neq j_t^*} (p_{t,j} - q_{t,j})^2 \leq \sum_{j \neq j_t^*} p_{t,j}^2 + \sum_{j \neq j_t^*} q_{t,j}^2.$$

Since

$$\sum_{j \neq j_t^*} p_{t,j}^2 \leq (1 - c_t)^2, \qquad \sum_{j \neq j_t^*} q_{t,j}^2 \leq (1 - q_t^{(1)})^2,$$

we obtain

$$\|p_t - q_t\|_2^2 \leq (c_t - q_t^{(1)})^2 + (1 - c_t)^2 + (1 - q_t^{(1)})^2. \tag{18}$$

Equations 17 and 18 generalize the one-hot correct-top regime. They show that when the model's top prediction aligns with the dominant target index, EL2N is still controlled by how close the confidence c_t is to the dominant target mass $q_t^{(1)}$.

Case 2: the top prediction disagrees with the dominant target index. Now assume that the model's top prediction is different from the dominant target index, so that

$$p_{t,j_t^*} \neq c_t.$$

Since j_t^* is not the top-predicted index, we have

$$p_{t,j_t^*} \leq p_{t,(2)} = c_t - m_t.$$

Therefore,

$$q_t^{(1)} - p_{t,j_t^*} \geq q_t^{(1)} - c_t + m_t.$$

Let k denote the top-predicted index, so that $p_{t,k} = c_t$ and $k \neq j_t^*$. Since j_t^* is the dominant target index, the target probability at k cannot exceed the second-largest target mass:

$$q_{t,k} \leq q_t^{(2)}.$$

Hence,

$$p_{t,k} - q_{t,k} \geq c_t - q_t^{(2)}.$$

Using these two coordinates in the EL2N sum yields

$$\|p_t - q_t\|_2^2 \geq (q_t^{(1)} - c_t + m_t)^2 + (c_t - q_t^{(2)})^2. \quad (19)$$

Equation 19 generalizes the one-hot wrong-top regime. It shows that when the dominant target index is not the top prediction, confidence alone is insufficient: the error depends jointly on the overconfident wrong top prediction and on the competition structure around the dominant target, which is precisely what the margin captures.

Why entropy remains necessary. The bounds above explain why confidence and margin are informative, but they still do not fully determine EL2N. Even when c_t and m_t are fixed, the remaining probability mass can be distributed in different ways across the non-top alternatives, which changes both the shape term $\|p_t\|_2^2$ in Equation 15 and the residual discrepancy terms in Equations 16–19. Entropy captures this global spread of probability mass and therefore provides complementary information that is not recoverable from confidence and margin alone.

In summary, the general target distribution case preserves the same intuition as the one-hot analysis: confidence acts as the primary anchor of prediction error, margin refines the local competition structure when the highest prediction is unreliable, and entropy captures the global dispersion of the predictive distribution. This justifies using these inference-time uncertainty signals to predict EL2N and construct the difficulty proxy used by the switch.

B Additional Experimental Details

B.1 Data Split and Training Details

All datasets used in this work are publicly available benchmark datasets and are used only for research and evaluation purposes. We follow the licenses or terms of use specified by the original dataset providers. We do not redistribute the raw datasets; users should obtain them from the official sources.

For all datasets, we split the data into training, validation, and test sets with a ratio of 8.1:0.9:1.

For the EL2N proxy model introduced in Section 3.1, we use a two-layer MLP with hidden size 256 for each layer. The model is trained with learning rate 5×10^{-4} for 50 epochs.

For the identifier introduced in Section 3.2, we instantiate the model with T5-large. The identifier is trained with learning rate 3×10^{-5} for 15 epochs.

For the MLP-based identifier baseline in Section 4.1, we use the same four input features, namely confidence, margin, entropy, and the predicted EL2N-based training difficulty. The classifier consists of two hidden layers with hidden size 64, and is trained with learning rate 3×10^{-4} for at most 100 epochs, with early stopping patience set to 20.

All experiments are conducted on two NVIDIA A6000 GPUs, each with 48GB memory.

B.2 Construction of Retrieval Labels

A key step in training the identifier is obtaining retrieval labels for the training set. To obtain such labels, we use *out-of-fold predictions* from the no-retrieval IE pipeline.

Specifically, we partition the original training data into 10 folds. For each round, we train the no-retrieval IE pipeline on 9 folds and compute instance-level F1 on the held-out fold. Repeating this process over all 10 folds yields out-of-fold F1 scores for every training instance. We then derive the binary supervision labels according to the rule in Section 3.2: instances with perfect no-retrieval performance are assigned label 0, and the remaining instances are assigned label 1.

This procedure ensures that each training instance is labeled by a model that was not trained on

that instance, thereby reducing leakage and producing a more faithful estimate of whether the instance can be solved without retrieval.

More concretely, instance-level F1 is computed for each individual instance by comparing the predicted output set against the gold output set under the task-specific evaluation criterion. An instance receives perfect F1 only when all target items are predicted exactly correctly, with no missing or spurious outputs. These instance-level F1 scores are then used to derive the binary retrieval labels.

B.3 Baseline Adaptation Details

All external baselines considered in this work were originally proposed for question answering settings, but the retrieval decision modules involved in these methods can be run independently and therefore adapted to generative IE tasks.

For evaluation, we first obtain the ground-truth retrieval labels for each backbone model on each IE task using the procedure described in Section B.2. These labels are used as the ground truth for evaluating retrieval decision quality.

For RetrievalQA and Self-RAG, the retrieval decision relies on large language models to perform self-assessment or to assess the target IE model. Therefore, we adapt these methods to the IE setting by modifying their prompts while preserving their original retrieval decision mechanisms. Specifically, we use the following prompt template:

> Give a question, determine whether you/[backbone model name] model need to retrieve external resources, such as real-time search engines, Wikipedia, or databases, to answer the question correctly. Only answer Yes or No.

For Self-RAG, deeper modification of the original model would require reconstructing the training data and retraining the model. Therefore, we only use the publicly available T5-based checkpoint released on HuggingFace.

For Adaptive-RAG, the original retrieval decision routing is simplified into a binary Yes/No retrieval decision. Moreover, instead of constructing supervision from retrieval outcomes as in the original method, we train its retrieval decision module using exactly the same supervision data as our identifier. To ensure fair comparison, the training setting is kept the same as that of our identifier.

Overall, although these methods were originally proposed for different settings, we focus specifically on their retrieval decision modules rather than the full original methods, and adapt them to the IE setting under the same supervision data and the same evaluation protocol.

B.4 Additional Evaluation Details

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, with respect to the binary retrieval labels. Here, the positive class corresponds to the retrieval-beneficial class, and the negative class corresponds to the retrieval-unnecessary class.

The recall on the positive class and the recall on the negative class are defined as

$$\mathrm{Recall}_+ = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$
$$\mathrm{Recall}_- = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

Balanced Accuracy (BalAcc) is then computed as the average of these two recalls:

$$\mathrm{BalAcc} = \frac{1}{2}\left(\mathrm{Recall}_+ + \mathrm{Recall}_-\right).$$

Yes-F1 measures performance on the retrieval-beneficial class. Let

$$\mathrm{Precision}_+ = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$
$$\mathrm{Recall}_+ = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}.$$

Then Yes-F1 is defined as

$$\text{Yes-F1} = \frac{2\cdot \mathrm{Precision}_+ \cdot \mathrm{Recall}_+}{\mathrm{Precision}_+ + \mathrm{Recall}_+}.$$

No-F1 measures performance on the retrieval-unnecessary class. Let

$$\mathrm{Precision}_- = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}},$$
$$\mathrm{Recall}_- = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}.$$

Then No-F1 is defined as

$$\text{No-F1} = \frac{2\cdot \mathrm{Precision}_- \cdot \mathrm{Recall}_-}{\mathrm{Precision}_- + \mathrm{Recall}_-}.$$

Statistical significance test. The main retrieval decision results are averaged over 10-fold cross-validation on test sets. We conduct statistical significance testing against the strongest baseline using a paired t-test, with significance level set to $p < 0.05$.

Backbone	Signal	p_r	s_ρ	AUC	mw_p
BERT	EL2N	**0.364**	**0.400**	**0.730**	**5e-18**
	Margin	0.232	0.228	0.631	4e-7
	Confidence	0.269	0.327	0.687	3e-13
	Entropy	0.316	0.230	0.635	1e-12
T5	EL2N	**0.587**	**0.593**	**0.845**	**1e-40**
	Margin	0.547	0.590	0.842	5e-40
	Confidence	0.523	0.531	0.808	9e-33
	Entropy	0.509	0.514	0.799	7e-31
LLaMA	EL2N	**0.466**	**0.485**	**0.783**	**1e-27**
	Margin	0.381	0.394	0.730	9e-19
	Confidence	0.407	0.472	0.776	2e-26
	Entropy	0.277	0.389	0.727	2e-18

Table 4: Correlation between difficulty-aware signals and retrieval decision labels using last-epoch statistics on Florilege. p_r and s_ρ are reported in absolute value.

Backbone	Signal	p_r	s_ρ	AUC	mw_p
BERT	EL2N	0.429	**0.559**	**0.832**	**0**
	Margin	0.392	0.474	0.781	7e-258
	Confidence	**0.475**	0.486	0.788	7e-271
	Entropy	0.360	0.459	0.772	5e-242
T5	EL2N	**0.552**	**0.496**	**0.861**	**0**
	Margin	0.491	0.467	0.855	**0**
	Confidence	0.524	**0.496**	**0.861**	**0**
	Entropy	0.504	0.474	0.858	**0**
LLaMA	EL2N	**0.559**	**0.485**	**0.843**	**5e-196**
	Margin	0.408	0.477	0.838	4e-183
	Confidence	0.464	0.480	0.840	2e-192
	Entropy	0.494	0.422	0.802	4e-150

Table 5: Correlation between difficulty-aware signals and retrieval decision labels using last-epoch statistics on ChemProt. p_r and s_ρ are reported in absolute value.

C Additional Correlation Results

Table 4 and Table 5 provide the corresponding correlation results on Florilege and ChemProt, complementing the CHEMDNER example shown in Table 3 of the main text.

Overall, the results on these two datasets follow a pattern similar to that of CHEMDNER. Across backbones, all four signals remain statistically associated with the retrieval decision labels, as reflected by their AUC values and small mw_p values. EL2N continues to rank among the strongest signals in most settings, especially in terms of AUC, while confidence, margin, and entropy also remain informative. This is consistent with our motivation for introducing EL2N: uncertainty-related signals are useful, but an explicit model-specific training difficulty proxy provides a stronger signal for retrieval decision making. These supplementary results therefore provide additional evidence that the signal–label relationships observed in the main text are not specific to CHEMDNER, but are consistently observed across the other datasets as well.

D End-to-End Trade-off Evaluation Details

D.1 End-to-End Pipeline Implementation

In this section, we additionally include Marina et al. (2025) as a baseline for comparison, as their approach is recognized for its high computational efficiency. Marina et al. decide whether retrieval is needed using lightweight external features rather than LLM-based uncertainty estimation. It trains a classifier on signals such as entity popularity, question type, question complexity, and context relevance. Due to differences in task settings, we retained only three types of features in our experiments: frequency, question complexity, and context relevance, and simplified the routing decision to a binary yes/no choice.

For BiomedRAG and BIOREADER, we reproduce the original settings as closely as possible, including the backbone model, chunking strategy, and retrieval strategy. The retrieval source is also kept identical to the original papers, namely PubMed.

For Florilege, we build an end-to-end entity recognition pipeline based on T5 in a text generation format. The prompt is:

> Extract taxon, use, phenotype, molecule and habitat entities from the given context. If RAG information is provided, use it when helpful. If RAG is None, rely only on the context. Context: [instance context]. RAG: [retrieval info].

For retrieval, our goal is to simulate a retrieval-enhanced setting in which the retrieved content is relevant but still contains noise. We therefore use SilverCorpus as the retrieval source. SilverCorpus is highly related in topic to Florilege. It is constructed from PubMed abstracts filtered for plant fermentation themes, and then annotated with relevant entities and relations using GPT-4.1. Since some entity types in SilverCorpus fully overlap with those in Florilege, we directly sample matched instances from SilverCorpus and return all annotated entity content in the matched instance as retrieved examples for enhancement. This design also helps reduce the RAG process time cost.

In terms of runtime workflow, for all end-to-end pipelines and all baseline retrieval decision modules, we adopt the deployment structure that is

Dataset	Improved (%)	Unchanged (%)	Degraded (%)
Florilege	37.29	59.32	3.30
CHEMDNER	26.43	70.72	2.85
ChemProt	43.80	54.48	1.72

Table 6: Effect of retrieval on instances predicted by RAG Switch as retrieval-beneficial and routed to retrieval. Changes are measured by instance-level F1 relative to the corresponding no-retrieval extraction.

most favorable to their efficiency and performance. Specifically, the corresponding baseline retrieval decision module makes its decision from the instance context before the task model is invoked in the inference pipeline. For RAG Switch, the deployment follows Figure 1 exactly.

D.2 Retrieval Outcomes on Routed Instances

Our task-grounded supervision is constructed from no-retrieval extraction outcomes and is therefore independent of any particular retrieval database or retrieval strategy. To further examine the realized effect of retrieval in the end-to-end RAG-enhanced IE pipelines, we analyze the instances predicted by RAG Switch as retrieval-beneficial and consequently routed to retrieval.

Table 6 reports the proportions of routed instances whose instance-level F1 scores are improved, unchanged, or degraded after retrieval. Retrieval improves 37.29%, 26.43%, and 43.80% of the routed instances on Florilege, CHEMDNER, and ChemProt, respectively, whereas degradation occurs in only 3.30%, 2.85%, and 1.72% of the cases.

These results characterize the realized behavior of the instantiated RAG pipelines rather than a guaranteed effect of RAG Switch. RAG Switch identifies instances for which retrieval is potentially useful based on the failure of no-retrieval IE, while the eventual benefit of retrieval also depends on factors such as the retriever, retrieved evidence, context construction, prompting strategy, and IE model. Consequently, some routed instances may remain unchanged. Nevertheless, degradation is consistently rare across all three datasets, while a substantial proportion of routed instances benefit from retrieval.

D.3 Retrieval-Rate Sensitivity Analysis

To investigate the relationship between retrieval coverage and end-to-end performance, we conduct a retrieval-rate sensitivity analysis under the same

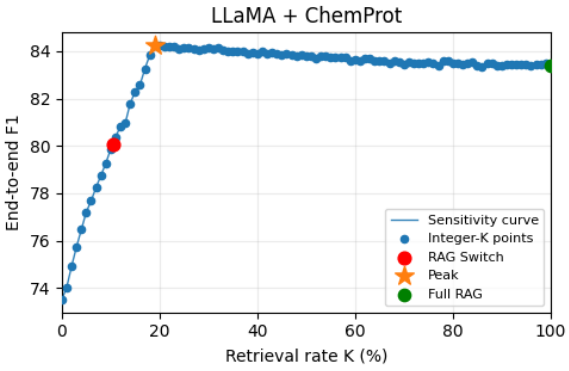


Figure 3: Retrieval-rate sensitivity curve for LLaMA on ChemProt. The red marker, star, and green marker indicate the RAG Switch operating point, peak performance, and Full-RAG setting, respectively.

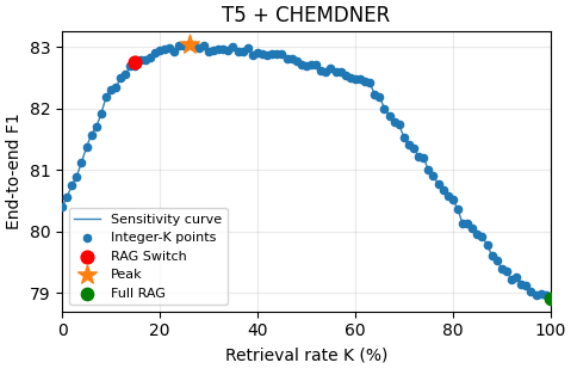


Figure 4: Retrieval-rate sensitivity curve for T5 on CHEMDNER. The red marker, star, and green marker indicate the RAG Switch operating point, peak performance, and Full-RAG setting, respectively.

RAG-enhanced IE settings. Specifically, test instances are ranked according to the retrieval-need score predicted by RAG Switch, and retrieval is progressively enabled for the top-$K\%$ instances according to this ranking. The corresponding end-to-end F1 curves are shown in Figure 3 to 5.

Across all three IE settings, the end-to-end F1 initially increases as retrieval is activated for more beneficial instances, demonstrating that the retrieval-need score provides an effective ranking of instances according to their expected retrieval benefits. The improvement gradually saturates before reaching the peak region, where small plateaus and fluctuations are observed due to the discrete instance-level routing decisions. After the peak point, further increasing the retrieval coverage leads to a gradual performance degradation, suggesting that indiscriminate retrieval over additional instances may not always benefit the end-to-end IE performance.

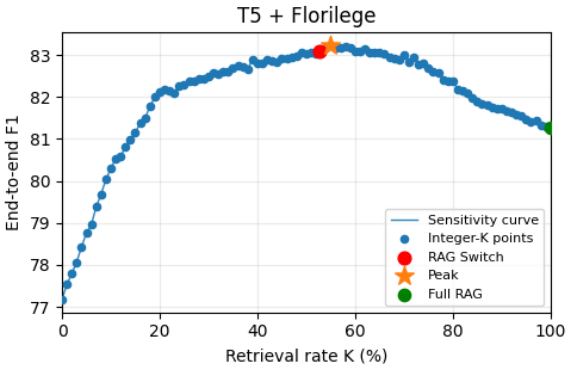


Figure 5: Retrieval-rate sensitivity curve for T5 on Florilege. The red marker, star, and green marker indicate the RAG Switch operating point, peak performance, and Full-RAG setting, respectively.

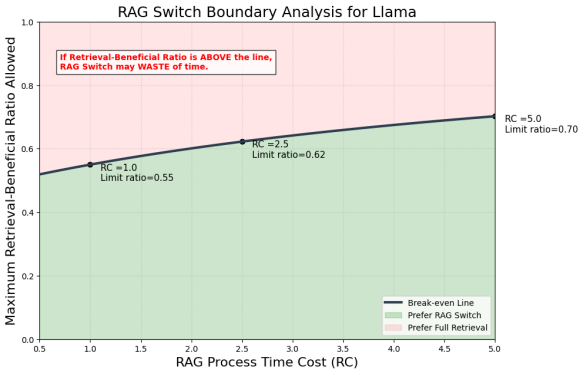


Figure 6: RAG Switch boundary analysis of end-to-end time efficiency for LLaMA model.

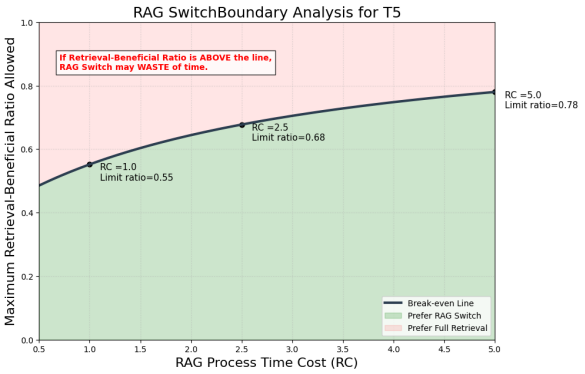


Figure 7: RAG Switch boundary analysis of end-to-end time efficiency for T5 model.

The optimal retrieval rates are substantially lower than the 100% Full-RAG setting, with peak performance achieved at: LLaMA+ChemProt: F1=84.25 at K=19%, T5+CHEMDNER: F1=83.04 at K=26%, and T5+Florilege: F1=83.23 at K=55%. These results demonstrate that selective retrieval over a properly ranked subset of instances can achieve superior end-to-end performance compared with uniformly applying retrieval to all instances.

The original RAG Switch decisions activate retrieval for 10.5%, 14.9%, and 52.8% of instances on ChemProt, CHEMDNER, and Florilege, respectively, achieving end-to-end F1 scores of 80.08, 82.75, and 83.09. These operating points fall within the low-to-moderate retrieval-rate regions where substantial performance gains have already been obtained, while avoiding retrieval activation for a large proportion of instances. Therefore, the sensitivity analysis further validates that the retrieval-need score learned by RAG Switch enables effective instance-level retrieval selection and achieves a favorable performance–efficiency trade-off compared with Full-RAG.

E Boundary Analysis of End-to-End Time Efficiency

To better understand the boundary conditions of the end-to-end time efficiency of RAG Switch, we further analyze the break-even point between our switching strategy and the Full-RAG setting under different IE backbones. Figure 6 to 8 show the corresponding boundaries for LLaMA, T5, and BERT, respectively. These plots are intended as practical references for deployment.

The retrieval decision module of RAG Switch introduces an average inference overhead of 0.22s per instance in our experiments. In this analysis, the RAG process time cost is estimated from the average end-to-end statistics reported in Section 4.4. It denotes the time required by the RAG process from receiving the query to returning the final retrieved results, including possible embedding, matching, and re-ranking stages. For each backbone, the boundary curve indicates the maximum proportion of retrieval-beneficial instances under which the switching strategy remains more time-efficient than Full-RAG.

Overall, the boundary analysis suggests that, under the large majority of realistic operating conditions, RAG Switch remains more time-efficient than Full-RAG. This advantage becomes even more pronounced as the time cost of the RAG process increases. Empirically, we observe that when the base model achieves stronger task performance, the proportion of retrieval-beneficial instances is typically well below the theoretical boundary, often falling within the range of 10%–30%. This suggests that, in regularly trained settings, only a rela-

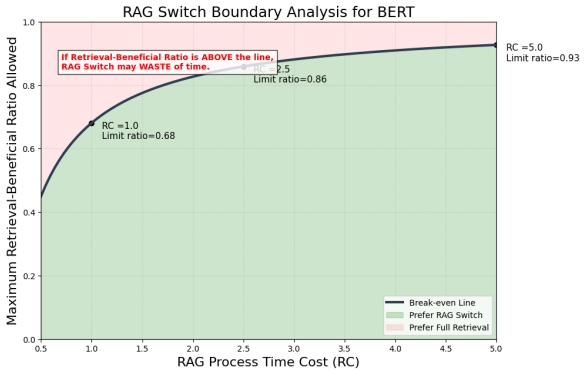


Figure 8: RAG Switch boundary analysis of end-to-end time efficiency for BERT model.

tively small subset of instances truly benefits from retrieval, while applying Full-RAG to all instances introduces unnecessary overhead. For weaker models such as BERT, the retrieval-beneficial proportion can be higher and may lie closer to the boundary, often around 40%–60%, which makes the time-efficiency concern more relevant. However, BERT-style encoders are substantially faster than generative models or larger LLMs, so the additional cost of the no-retrieval probe accounts for a smaller share of the total runtime. At the same time, the higher retrieval-beneficial proportion means that selective retrieval also has greater potential to improve task performance, since more instances can genuinely benefit from retrieval. Therefore, even in this regime, the switching strategy can still remain practically attractive in terms of the overall efficiency–performance trade-off.

For few-shot or otherwise less stable settings, the boundary depends more sensitively on the underlying pipeline behavior. As a practical heuristic, one may refer to the performance of the pipeline without RAG enhancement: when its performance remains above roughly 40 F1-scores, switching is still likely to be favorable in many cases. In addition, reducing the length of the instance context, for example by splitting it into shorter fragments, may simultaneously reduce inference time and lower the proportion of retrieval-beneficial instances, thereby making the switching strategy more attractive.