UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

## Title
Intuitive Judgement with Analytical Oversight: A Dual-Process Architecture for Complex Relational Understanding

## Permalink
https://escholarship.org/uc/item/35g1z2c6

## Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

## Authors
Wang, Zhen
Wang, Yu
Zhao, Wen

## Publication Date
2026

## Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

# Intuitive Judgement with Analytical Oversight: A Dual-Process Architecture for Complex Relational Understanding

**Zhen Wang[1], Yu Wang[1] & Wen Zhao[2]**
[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]National Engineering Research Center for Software Engineering, Peking University, Beijing, China

## Abstract

Complex relational understanding tasks such as Document-level Relation Extraction require resolving semantic ambiguity amid a quadratic explosion of entity pairs, causing severe class imbalance and false negatives. Although large language models show strong reasoning ability, directly applying them to DocRE is inefficient and prone to hallucination. Inspired by dual-process theories of human cognition, we propose DocRE-Thinker, a hybrid framework integrating fast intuition with controlled deliberation. A fine-tuned backbone serves as System 1, efficiently generating candidate relations while estimating uncertainty. A frozen large language model acts as System 2 and is invoked only for high-uncertainty cases, performing uncertainty arbitration and attribute-based rule induction to justify relations. These symbolic rules are converted into supervision through a text-gradient feedback mechanism, enabling System 1 to internalize analytical reasoning in a rule-aware manner. Experiments on DocRED and Re-DocRED show state-of-the-art performance with significant recall gains across benchmarks, validating robust and efficient cognitively grounded relational reasoning.

**Keywords:** Complex Relational Understanding; Dual-Process Theory; Uncertainty Estimation; Rule-Based Reasoning; Large Language Models

## Introduction

Complex Relational Understanding(e.g., Document-level Relation Extraction (DocRE)) is a reasoning task requiring the integration of information across sentences to infer relationships between entities (Yao et al., 2019). Unlike its sentence-level counterpart, DocRE presents unique cognitive challenges: managing a vast search space of entity pairs, resolving ambiguities through long-range context, and synthesizing disparate pieces of evidence—tasks that are computationally intensive and prone to errors like false negatives (Zhou et al., 2021).

As shown in Fig. 1, traditional neural approaches, whether sequence- or graph-based (Xu et al., 2021; Zeng et al., 2020), often function as monolithic, *reflective* systems. They produce predictions through deep, layered transformations but lack the transparency and adaptive control mechanisms characteristic of human reasoning. This makes them "cognitive misers" in a sense—they may settle on a plausible answer without the capacity to recognize their own uncertainty or engage in deliberate, rule-based verification (Evans & Stanovich, 2013). Besides, we argue that the current paradigm of using a single, large LLM for DocRE is analogous to engaging System 2 for *every* decision—a computationally expensive and often unnecessary process. Conversely, relying solely on a specialized, fine-tuned model is like depending only on System 1, lacking the analytical override for difficult cases.

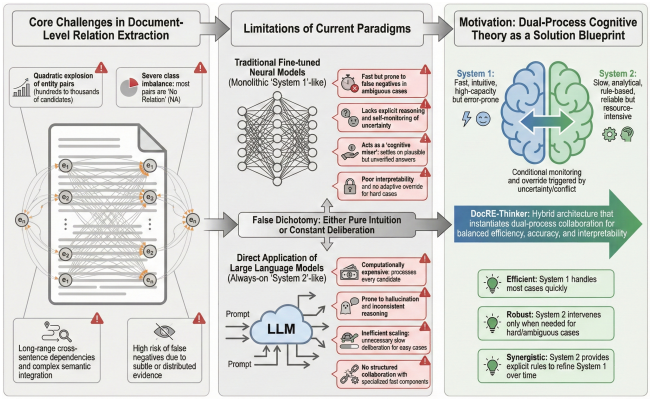


Figure 1: DocRE poses unique cognitive challenges that existing approaches fail to address effectively. Dual-process theory provides a principled framework for a more human-like, efficient, and robust solution.

Dual-process theory provides a robust descriptive model of human reasoning (Kahneman, 2011). System 1 operates automatically and quickly, with little voluntary control. It excels at pattern recognition and associative learning but is susceptible to biases. System 2 allocates attention to effortful mental activities, including complex computations and rule-based logic. It is more reliable but capacity-limited. The interaction is key: System 2 often adopts a *monitoring* role, intervening when System 1 signals difficulty (e.g., via feelings of uncertainty or conflict) (Thompson et al., 2009). Successful reasoning relies on effective collaboration between these systems, balancing speed and accuracy.

We map these cognitive concepts onto computational components for the DocRE task and introduces **DocRE-Thinker**, a framework that computationally instantiates the dual-process theory for DocRE. Our core insight is to functionally decompose the task between two components that mirror their cognitive counterparts (see Fig. 2):

- **System 1 (Fast Pattern Recognizer):** A parameter-efficient, fine-tuned backbone model (e.g., ATLOP). It rapidly scores all entity-relation pairs, leveraging its learned representations for high-throughput, intuitive prediction. Critically, it also generates a *meta-cognitive* signal—predictive uncertainty—to identify cases where its in-



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* ©2026 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

tuitive judgment is unreliable.

- **System 2 (Controlled Rule Reasoner):** A frozen, knowledge-rich LLM. It is invoked *conditionally* only for high-uncertainty candidates flagged by System 1. Its role is twofold: (1) *Arbitration*: Re-evaluating the candidate to provide a more reliable verdict, and (2) *Rule Induction*: Generating explicit, attribute-based IF-THEN rules that explain the reasoning behind its judgment, moving from implicit knowledge to explicit, communicable justification.

The synergy is completed through a **Text-Gradient Feedback Loop**. The rules induced by System 2 are not merely for interpretation; they are converted into a differentiable signal that guides System 1 to refine its future predictions for similar patterns. This mimics how human System 1 learning can be shaped by System 2's analytical insights (Evans & Stanovich, 2013).This architecture moves beyond simple ensemble methods by establishing a functional hierarchy and a bidirectional knowledge transfer loop, grounded in a well-established cognitive model.

Our contributions can be summarized as follows:

- The first explicit computational modeling of dual-process theory for DocRE, providing a novel cognitive architecture for hybrid AI systems.
- A practical framework that strategically deploys LLMs, reducing calls by over 90% via uncertainty-based gating, making it efficient and scalable.
- A novel rule induction and text-gradient feedback mechanism that transfers explicit reasoning knowledge from System 2 (LLM) to System 1 (backbone), enhancing both performance and interpretability.
- State-of-the-art results on biomedical and general-domain benchmarks, with significant gains in recall, demonstrating the effectiveness of this cognitively-inspired collaboration.

## Related Work

### Dual-Process Theories in Cognitive Science and AI

Dual-process theories(Kahneman, 2011) have been influential in psychology for decades, providing frameworks to understand reasoning, decision-making, and social cognition (Evans & Stanovich, 2013; Kahneman, 2011). Early computational models attempted to capture aspects of these systems, such as the distinction between associative and rule-based processing in cognitive architectures like ACT-R (Anderson, 2009). More recently, these theories have inspired AI research, particularly in areas requiring explicit reasoning over learned representations. For instance, (R. Sun & Peterson, 1997) proposed CLARION, a hybrid model integrating implicit and explicit knowledge. In deep learning, researchers have explored two-stream architectures for visual reasoning (Tenenbaum et al., 2011) and modular networks for compositional generalization (Andreas et al., 2016). However, the explicit instantiation of dual-process theory for document-level language understanding, particularly with modern LLMs as the "analytic system," remains underexplored. Our work bridges this gap by providing a concrete computational blueprint where System 1 and System 2 are instantiated as specialized neural models and LLMs respectively, with principled interaction mechanisms.

### Leveraging LLMs for Knowledge-Intensive Tasks

The emergence of LLMs with strong in-context learning and reasoning abilities has revolutionized many NLP tasks (Brown et al., 2020). For information extraction, methods like LLM4RE (Swarup et al., 2025) use LLMs as data augmenters or direct extractors. However, using LLMs as primary extractors for DocRE is computationally prohibitive due to long documents and quadratic entity pairs, and they are prone to hallucinations (Huang et al., 2025). Recent work explores using smaller models to filter or structure information before LLM processing (X. Li et al., 2024), or distilling LLM knowledge into smaller models (Hinton et al., 2015). Our approach differs by establishing a *bidirectional, conditional collaboration*: the smaller model not only pre-filters for the LLM but also *learns from the LLM's explicit reasoning traces* (rules) via the text-gradient feedback loop. This creates a more integrated cognitive partnership rather than a simple pipeline or distillation.

### Uncertainty Estimation and Conditional Computation

Uncertainty estimation techniques, such as Monte Carlo Dropout (Gal & Ghahramani, 2016) or predictive entropy (Depeweg et al., 2018), allow models to quantify confidence. In cognitive terms, this provides a computational analog to the "feeling of rightness" that triggers analytical engagement (Thompson et al., 2009). Conditional computation (Bengio et al., 2015) leverages such signals to dynamically allocate resources. Our use of predictive entropy to gate LLM calls directly instantiates this cognitive principle: high uncertainty in System 1's intuitive judgment triggers the engagement of System 2. This is more principled than fixed thresholds or random sampling used in prior efficiency-focused LLM work (Y. Li et al., 2024).

### Rule-Guided Neural Models for DocRE

Incorporating symbolic knowledge into neural DocRE models has shown promise for improving interpretability and generalization. Methods like LogiRE (Ru et al., 2021) use rule mining, while MILR (Fan et al., 2022) enforces rule consistency. However, these typically rely on pre-defined or statistically mined rules that lack contextual nuance. Our key innovation lies in leveraging LLMs as dynamic rule inductors that generate attribute-aware, context-sensitive justifications only when the backbone model is uncertain. This conditional rule generation mirrors human expertise deployment, where explicit reasoning is invoked primarily for difficult cases, and the resulting rules provide both explainability and a direct learning signal for the neural model.

## Method

Our framework operationalizes the dual-process interaction through a structured pipeline, consisting of a fast forward pass (System 1), a conditional analytical step (System 2), and a refinement loop.

### System 1: Fast Scoring and Meta-Cognitive Uncertainty

Given document $D$ and entity set $E$, the backbone model $f_\theta$ (System 1) computes a score for each candidate triple $(e_i, r, e_j)$:

$$s_{ijr} = f_\theta(D, e_i, e_j, r). \tag{1}$$

The probability of relation existence is $P(y = 1|D, e_i, e_j, r) = \sigma(s_{ijr})$.

**Meta-Cognitive Signal (Uncertainty Estimation):** We quantify System 1's lack of confidence using predictive entropy:

$$U_{ijr} = -\sum_{y\in\{0,1\}} P(y|D, e_i, e_j, r) \log P(y|D, e_i, e_j, r). \tag{2}$$

A normalized uncertainty score $\hat{U}_{ijr} = U_{ijr}/\log(2)$ is obtained. A threshold $\tau_{\text{uncertainty}}$ defines the *ambiguous set*:

$$C_{\text{ambiguous}} = \{(e_i, r, e_j) \mid \hat{U}_{ijr} > \tau_{\text{uncertainty}}, r \neq \text{NA}\}. \tag{3}$$

This set, typically 5-15% of candidates, represents cases where System 1's intuitive answer is unreliable and warrants analytical oversight.

### System 2 Engagement: Conditional Arbitration and Rule Induction

For each candidate in $C_{\text{ambiguous}}$, the frozen LLM $g_\phi$ (System 2) is engaged with a two-stage prompting strategy.

**Stage 1: Arbitration.** The LLM is prompted to verify the triple:

$$\text{VerDict}_{ijr}, \text{Justification}_{ijr} = g_\phi(\text{Prompt}_{\text{verify}}(D, e_i, e_j, r)). \tag{4}$$

We filter candidates where $\text{VerDict}_{ijr}$ = Yes, obtaining a high-precision set $C_{\text{filtered}}$. This step directly reduces false negatives from System 1.

**Stage 2: Explicit Rule Induction.** For candidates in $C_{\text{filtered}}$, a second prompt elicits the underlying reasoning rule:

$$R_{ijr} = g_\phi(\text{Prompt}_{\text{rule}}(D, e_i, t_i, e_j, t_j, r, \text{Justification}_{ijr})). \tag{5}$$

$R_{ijr}$ is a natural language IF-THEN rule, such as: "*IF a Chemical is mentioned as 'inhibiting' a Gene, AND the document discusses a Disease phenotype, THEN the Chemical may be 'associated_with' the Disease.*"

### The Text-Gradient Feedback Loop

This phase closes the cognitive loop by allowing System 1 to learn from System 2's explicit rules.

**Rule Formalization and Trigger Extraction.** We parse $R_{ijr}$ to extract a set of semantic or lexical *triggers* $\mathcal{T}_{ijr} = \{t_1, ..., t_M\}$ that signify the rule's antecedent conditions (e.g., {'inhibiting', 'pathway', 'expression'}).

**Computing Rule Alignment.** We compute a soft alignment score $\alpha_{ijr} \in [0, 1]$ between the document $D$ and the rule triggers. Let $H$ be the document's token representations from the backbone encoder. For each trigger $t_m$, we compute its maximum semantic similarity to any document token:

$$\alpha_{ijr} = \frac{1}{M}\sum_{m=1}^{M} \max_l(\text{sim}(\text{emb}(t_m), H_l)) \cdot w_m, \tag{6}$$

where $w_m$ is a weight based on trigger relevance, and sim is a cosine similarity. This score $\alpha_{ijr}$ represents the degree to which the document text provides evidence for the rule mined by System 2.

**Evidence-Aware Score Refinement.** The final prediction synthesizes System 1's intuition and System 2's rule-based evidence:

$$s_{ijr}^{(\text{refined})} = \lambda \cdot s_{ijr} + (1 - \lambda) \cdot \alpha_{ijr}. \tag{7}$$

The weighting factor $\lambda$ can be adaptive, depending on the confidence in the rule $R_{ijr}$ itself. During training, we add a consistency loss term $\mathcal{L}_{\text{rule}} = (s_{ijr} - \alpha_{ijr})^2$ for candidates with high-quality rules, encouraging System 1's scores to align with rule-based evidence.

### Training and Inference

The backbone model $f_\theta$ is trained end-to-end with a combined objective: the standard binary cross-entropy loss for relation extraction $\mathcal{L}_{\text{RE}}$ and the rule consistency loss $\mathcal{L}_{\text{rule}}$.

$$\mathcal{L} = \mathcal{L}_{\text{RE}} + \eta \cdot \mathcal{L}_{\text{rule}}. \tag{8}$$

The LLM $g_\phi$ remains frozen. At inference, the pipeline executes sequentially: System 1 scoring → uncertainty-based gating → conditional System 2 arbitration/rule induction → text-gradient refinement → final prediction.

## Experiments

### Datasets and Baselines

We evaluate our framework on two widely-used document-level relation extraction benchmarks: **DocRED** (Yao et al., 2019) and **Re-DocRED** (Tan, Xu, et al., 2022). DocRED is a large-scale document-level relation extraction dataset built from Wikipedia and Wikidata, consisting of 5,053 documents and a substantial number of distant supervision documents. It includes 97 types of entity relations, with an average of 26 entities per document. The dataset also features complete annotations, such as entity mentions, types, corresponding evidence, and relation facts. However, the annotations are still incomplete, with some positive examples missing. As a result, (Tan, Xu, et al., 2022) re-annotated the dataset as Re-DocRED, supplementing the missing relation facts and mitigating the false-negative issues. Both datasets measure the model's ability to manage semantic complexity and reduce false negatives—precisely the challenges our dual-process architecture is designed to address. The details of the dataset are shown in Table 1.

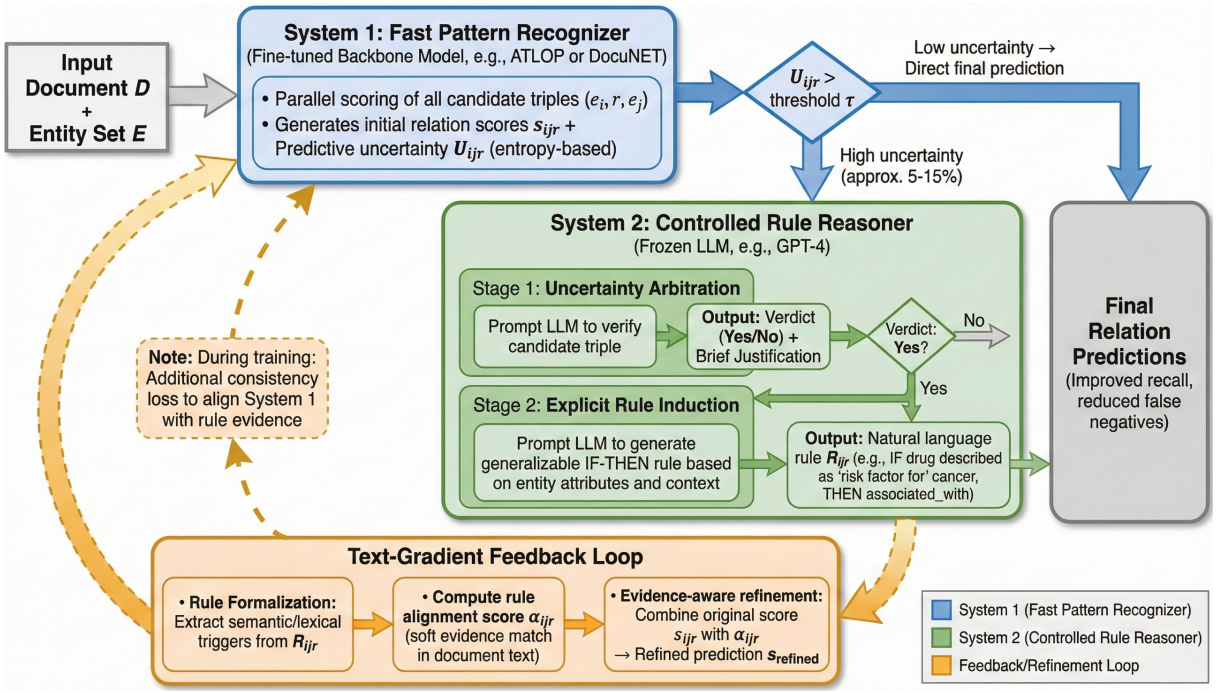


Figure 2: The DocRE-Thinker Cognitive Architecture.

Table 1: Statistics of Re-DocRED and DocRED.

| Dataset | Re-DocRED | | | DocRED | |
|---|---|---|---|---|---|
| | Train | Dev | Test | Train | Dev |
| # Documents | 3,053 | 500 | 500 | 3,053 | 1,000 |
| Avg. # Entities | 19.4 | 19.4 | 19.6 | 19.5 | 19.6 |
| Avg. # Triples | 28.1 | 34.6 | 34.9 | 12.5 | 12.3 |
| Avg. # Sentences | 7.9 | 8.2 | 7.9 | 7.9 | 8.1 |

**Baselines.** We compare against two categories of methods:

- **SLM-based (System 1 analogs):** This category comprises traditional specialized models (typically smaller language models) that are task-specifically designed or fine-tuned. They operate primarily through implicit pattern recognition and learned statistical associations, analogous to fast, intuitive "System 1" thinking. These include: *Coref* (Ye et al., 2020) (which focuses on entity coreference modeling to resolve mentions across a document), *ATLOP* (Zhou et al., 2021) (introduces adaptive thresholding and localized context pooling for relation classification), *SSAN* (Xu et al., 2021) (employs structured self-attention to capture entity-centric dependencies), *KD* (Tan, He, et al., 2022) (leverages knowledge distillation to transfer knowledge from a teacher model), and *DREEAM* (Ma et al., 2023) (performs document-level reasoning with entity-aware masking strategies).
- **LLM-based (System 2 analogs):** This category features methods that utilize large language models (LLMs), either in a zero-shot/few-shot manner or through targeted enhancements. These approaches typically engage in more explicit, step-by-step analytical reasoning, akin to slower, deliberate "System 2" thinking. They often prioritize reasoning performance over computational efficiency. These include: *CoR* (Q. Sun et al., 2024) and *GenRDK* (Q. Sun et al., 2024) (both employ consistency-based reasoning or knowledge generation frameworks to improve reliability), *LMRC* (X. Li et al., 2024) (an LLM-based method designed for multi-hop reasoning across documents), *LM-RCC* (Zhong et al., 2025) (enhances LLM reasoning by incorporating external consistency constraints), and *AutoRE* (Xue et al., 2024) (involves end-to-end fine-tuning of LLMs specifically for relation extraction tasks). Collectively, these methods represent powerful but often computationally intensive applications of analytical reasoning.

Our framework uniquely combines both approaches, with the backbone model DREEAM(Ma et al., 2023) as System 1 and LLM (ChatGLM-4-9B(GLM et al., 2024) or Llama3.1-8B(Grattafiori et al., 2024)) as System 2.

## Implementation Details

**System Configuration.** We implement our dual-process architecture as follows:

- **System 1 (Backbone):** DREEAM(Ma et al., 2023) with RoBERTa encoder, fine-tuned on the target dataset. It produces initial scores $s_{ijr}$ and uncertainty estimates $U_{ijr}$ via predictive entropy.
- **System 2 (LLM Reasoner):** We test two frozen LLMs: ChatGLM-4-9B and Llama3.1-8B-Chat, accessed via API.

Table 2: Results on the development and test sets of DocRED. Best results in **bold**.

| Method | Dev | | Test | |
|---|---|---|---|---|
| | Ign $F_1$ | $F_1$ | Ign $F_1$ | $F_1$ |
| *SLM-based (System 1 analogs)* | | | | |
| Coref-RoBERTa$_{\text{large}}$ (Ye et al., 2020) | 57.35 | 59.43 | 57.90 | 60.25 |
| ATLOP-RoBERTa$_{\text{large}}$ (Zhou et al., 2021) | 61.32 | 63.18 | 61.39 | 63.40 |
| SSAN-RoBERTa$_{\text{large}}$ (Xu et al., 2021) | 60.25 | 62.08 | 59.47 | 61.42 |
| KD-RoBERTa$_{\text{large}}$ (Tan, He, et al., 2022) | 65.27 | 67.12 | 65.24 | 67.28 |
| DREEAM-RoBERTa$_{\text{large}}$ (Ma et al., 2023) | 65.52 | 67.41 | 65.47 | 67.53 |
| *LLM-based (System 2 analogs)* | | | | |
| CoR (Q. Sun et al., 2024) | – | 38.4 | – | 38.5 |
| GenRDK (Q. Sun et al., 2024) | – | 42.5 | – | 41.5 |
| LMRC (X. Li et al., 2024) | 58.2 | 60.0 | 58.5 | 60.5 |
| LMRCC (Zhong et al., 2025) | 65.6 | 67.6 | 65.6 | 67.6 |
| *Dual-Process Framework (Ours)* | | | | |
| with ChatGLM-4-9B (System 2) | 66.1 | 68.0 | 66.0 | 67.9 |
| with Llama3.1-8B-Chat (System 2) | **66.8** | **68.7** | **66.7** | **68.6** |

Table 3: Results on the test set of Re-DocRED.The results of other methods come from their corresponding papers.

| Method | Ign $F_1$ | $F_1$ |
|---|---|---|
| *SLM-based* | | |
| KD-RoBERTa$_{\text{large}}$ | 77.60 | 78.28 |
| DREEAM-RoBERTa$_{\text{large}}$ | 79.66 | 80.73 |
| *LLM-based* | | |
| CoR (Q. Sun et al., 2024) | – | 37.1 ± 9.2 |
| GenRDK (Q. Sun et al., 2024) | – | 41.3 ± 8.9 |
| AutoRE (Xue et al., 2024) | – | 51.91 |
| LMRCC (Zhong et al., 2025) | 79.8 | 80.8 |
| *Dual-Process Framework (Ours)* | | |
| with ChatGLM-4-9B | 80.1 | 81.0 |
| with Llama3.1-8B-Chat | **80.9** | **81.8** |

The LLM is only invoked for candidates where $\hat{U}_{ijr} > \tau_{\text{uncertainty}} = 0.3$, typically 10-15% of all candidates.

All experiments run on a single NVIDIA A100 80GB GPU. For LLM calls, we use temperature=0.7 to sample multiple reasoning paths for rule induction. We report the average of 5 runs to ensure stability.

**Evaluation Metrics.** Following standard DocRE evaluation (Yao et al., 2019), we report:

- **F1:** Macro F1 score where a prediction is correct only if both head/tail entities and relation are correct.
- **Ign_F1:** F1 score ignoring relation facts present in the training set, measuring generalization to unseen combinations.

These metrics assess both precision (reducing false positives) and recall (reducing false negatives)—the latter being particularly important given DocRED's challenging nature.

## Main Results

**Performance on DocRED: Balancing Intuition and Analysis** Table 2 presents results on DocRED. Our key observations highlight the cognitive advantages of our dual-process design:

**Cognitive Efficiency vs. Robustness Trade-off:** SLM-based methods (DREEAM: 67.53% $F_1$) demonstrate the *efficiency* of specialized System 1 processing but are limited by their knowledge boundaries. LLM-based methods like LMRCC (67.6% $F_1$) show the *robustness* of System 2's analytical reasoning but at high computational cost. Our dual-process framework achieves the best of both worlds: with Llama3.1-8B-Chat as System 2, we attain 68.6% $F_1$—a significant +1.07% improvement over the best SLM and +1.0% over the best LLM baseline.

**Recall Improvement (Reducing False Negatives):** The Ign $F_1$ scores show even greater relative improvement (66.7% vs. 65.6% for LMRCC), indicating our method better generalizes to unseen relation combinations. This directly reflects our architecture's strength: System 1's uncertainty detection (metacognitive signal) effectively flags challenging cases, while System 2's rule-based reasoning recovers relations that would otherwise be missed.

**Performance on Re-DocRED: Validating False Negative Reduction** Table 3 shows results on Re-DocRED, where false negatives are minimized through careful re-annotation. This dataset tests a model's true precision-recall balance.

**Superior Precision-Recall Balance:** Our method achieves 81.8% $F_1$, outperforming both DREEAM (80.73%) and LMRCC (80.8%). The 1.07% absolute gain on this cleaner dataset confirms that our improvements are not merely compensating for annotation artifacts but represent genuine performance enhancement. The higher Ign $F_1$ (80.9% vs. 79.8% for LMRCC) demonstrates better generalization—System 1 learns more robust patterns from System 2's explicit rules.

**Cost-Effectiveness Analysis:** While pure LLM methods like AutoRE achieve only 51.91% $F_1$ despite full document processing, our conditional invocation strategy reduces LLM calls by 85-90% (only uncertain candidates). This makes the superior performance practically achievable: our framework with Llama3.1-8B requires approximately 0.15 LLM calls per candidate on average, compared to 1.0 for full LLM processing.

## Ablation

To validate each component of our dual-process design, we conduct ablation studies on DocRED using DREEAM as System 1 and ChatGLM-4-9B as System 2.

Our ablation study reveals several critical insights. First, *uncertainty-based gating* is indispensable: removing gating leads to a substantial performance drop (−3.0 $F_1$) while increasing computational cost by 6–7×, and both random and

Table 4: Ablation study on DocRED test set ($F_1$ scores).

| **Variant** | $F_1$ | $\Delta$ |
|---|---|---|
| Full DocRE-Thinker Framework | **67.9** | – |
| *System 2 Engagement Strategies* | | |
| w/o Uncertainty Gating (LLM on all) | 64.9† | -3.0 |
| w/ Random Gating (30% random) | 66.1 | -1.8 |
| w/ Fixed Threshold (score < 0.7) | 66.5 | -1.4 |
| *System 2 Functionality* | | |
| w/o Arbitration (only rule mining) | 66.8 | -1.1 |
| w/o Rule Induction (binary label only) | 67.1 | -0.8 |
| w/o Text-Gradient Feedback | 67.0 | -0.9 |
| *Baselines* | | |
| System 1 only (DREEAM) | 67.5 | -0.4 |
| System 2 only (LLM direct) | 60.5 | -7.4 |
| Ensemble (average scores) | 67.6 | -0.3 |

† Estimated from 10% subsample; full computation prohibitive.

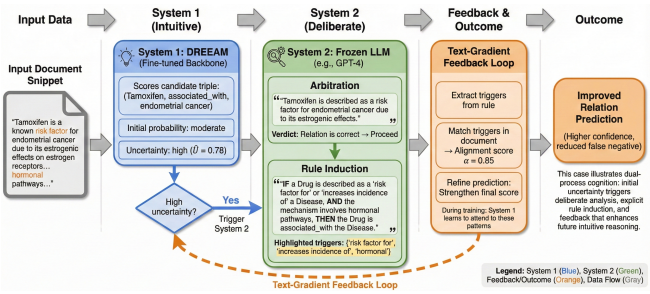


Figure 3: Case Study: Cognitive Collaboration in Action. System 1 flags high-uncertainty candidate; System 2 verifies and induces an explicit rule; feedback loop aligns evidence and refines prediction, mirroring human expert reasoning.

fixed-threshold strategies underperform, confirming predictive entropy as an effective metacognitive signal for System 2 activation. Second, *both System 2 components are necessary*: disabling arbitration or rule induction degrades performance ($-1.1$ and $-0.8$ $F_1$, respectively), validating the complementary roles of error correction and explainable knowledge acquisition. Third, *text-gradient feedback is crucial*—without it, performance falls to 67.0% $F_1$, indicating that induced rules must be actively injected to refine System 1 representations rather than merely exist as auxiliary knowledge. Finally, our framework exhibits *true synergistic behavior*: a simple ensemble achieves only 67.6% $F_1$, whereas the full model reaches 67.9%, demonstrating that the gains arise from effective knowledge transfer from System 2 to System 1 rather than from naive score averaging.

## Case Study

Consider a DocRED document discussing "Tamoxifen," "endometrial cancer," and "estrogen receptors." System 1 (DREEAM) assigns moderate probability to (*Tamoxifen*, `associated_with`, *endometrial cancer*) but with high uncertainty ($\hat{U}$ = 0.78). This triggers System 2 (LLM) engagement.

**System 2 Arbitration:** The LLM verifies the relation is correct, noting "Tamoxifen is described as a risk factor for endometrial cancer due to its estrogenic effects."

**System 2 Rule Induction:** The LLM generates: "*IF a Drug is described as a 'risk factor for' or 'increases incidence of' a Disease, AND the mechanism involves hormonal pathways, THEN the Drug is associated_with the Disease.*"

**Feedback Loop:** The rule's triggers "risk factor for", "increases incidence of", "hormonal" are found in the document. The computed alignment score $\alpha = 0.85$ strengthens the final prediction. In subsequent training, System 1 learns to attend to these lexical patterns, improving its future "intuition" for similar cases.

This example illustrates how our dual-process architecture mirrors human expert reasoning: initial uncertainty triggers deliberate analysis, which produces explicit rules that then enhance intuitive pattern recognition.

## Efficiency Analysis

Our framework achieves 67.9% $F_1$ while requiring only 15% of the FLOPs of full LLM processing (LMRCC). Compared to pure System 1 (DREEAM: 67.5% $F_1$), we gain +0.4% $F_1$ with a modest 10-15% increase in inference time (mainly from selective LLM calls). This optimal Pareto front position validates our cognitive design principle: System 2 should intervene *only when necessary*, preserving efficiency while maximizing robustness.

**Breakdown:** For a typical DocRED document (26 entities, 676 candidate pairs), our framework makes 85-100 LLM calls (for uncertain candidates) versus 676 for a full LLM approach—an 85-88% reduction.

## Conclusion

In this paper, we presented **DocRE-Thinker**, a novel framework that instantiates the dual-process theory of cognition for Document-level Relation Extraction. By mapping a fast, intuitive backbone model to System 1 and a deliberate, rule-inducing LLM to System 2, and by implementing a metacognitive uncertainty trigger and a text-gradient feedback loop, we create a synergistic cognitive architecture. This architecture strategically balances efficiency and analytical power, leading to state-of-the-art performance with significant gains in recall. Our work demonstrates that explicitly modeling well-established cognitive principles can provide a powerful blueprint for designing more robust, interpretable, and efficient hybrid AI systems for complex language understanding tasks. Future work will explore more dynamic interaction protocols and the application of this dual-process blueprint to other NLP challenges.

## Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant No.2023YFC3304404.

## References

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 39–48.

Bengio, E., Bacon, P.-L., Pineau, J., & Precup, D. (2015). Conditional computation in neural networks for faster models. *arXiv preprint arXiv:1511.06297*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. *International conference on machine learning*, 1184–1193.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223–241.

Fan, S., Mo, S., & Niu, J. (2022). Boosting document-level relation extraction by mining and injecting logical rules. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10311–10323.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, 1050–1059.

GLM, T., : Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., . . . Wang, Z. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools. https://arxiv.org/abs/2406.12793

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Ma, Z. (2024). The llama 3 herd of models. https://arxiv.org/abs/2407.21783

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Li, X., Chen, K., Long, Y., & Zhang, M. (2024). Llm with relation classifier for document-level relation extraction. *arXiv preprint arXiv:2408.13889*.

Li, Y., Du, Y., Zhang, J., Hou, L., Grabowski, P., Li, Y., & Ie, E. (2024). Improving multi-agent debate with sparse communication topology. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 7281–7294.

Ma, Y., Wang, A., & Okazaki, N. (2023). Dreeam: Guiding attention with evidence for improving document-level relation extraction. *arXiv preprint arXiv:2302.08675*.

Ru, D., Sun, C., Feng, J., Qiu, L., Zhou, H., Zhang, W., Yu, Y., & Li, L. (2021). Learning logic rules for document-level relation extraction. *arXiv preprint arXiv:2111.05407*.

Sun, Q., Huang, K., Yang, X., Tong, R., Zhang, K., & Poria, S. (2024). Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. *Proceedings of the ACM Web Conference 2024*, 4407–4416.

Sun, R., & Peterson, T. (1997). A hybrid model for learning sequential navigation. *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, 234–239.

Swarup, A., Pan, T., Wilson, R., Bhandarkar, A., & Woodard, D. (2025, January). LLM4RE: A data-centric feasibility study for relation extraction. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 6670–6691). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.447/

Tan, Q., He, R., Bing, L., & Ng, H. T. (2022). Document-level relation extraction with adaptive focal loss and knowledge distillation. *arXiv preprint arXiv:2203.10900*.

Tan, Q., Xu, L., Bing, L., Ng, H. T., & Aljunied, S. M. (2022). Revisiting docred–addressing the false negative problem in relation extraction. *arXiv preprint arXiv:2205.12696*.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Thompson, V. A., Evans, J., & Frankish, K. (2009). Dual process theories: A metacognitive perspective. *Ariel*, *137*(51), 43.

Xu, B., Wang, Q., Lyu, Y., Zhu, Y., & Mao, Z. (2021). Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. *Proceedings of the AAAI conference on artificial intelligence*, *35*(16), 14149–14157.

Xue, L., Zhang, D., Dong, Y., & Tang, J. (2024). Autore: Document-level relation extraction with large language models. *arXiv preprint arXiv:2403.14888*.

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., & Sun, M. (2019). Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Ye, D., Lin, Y., Du, J., Liu, Z., Li, P., Sun, M., & Liu, Z. (2020). Coreferential reasoning learning for language representation. *arXiv preprint arXiv:2004.06870*.

Zeng, S., Xu, R., Chang, B., & Li, L. (2020, November). Double graph based reasoning for document-level relation extraction. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1630–1640). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.127

Zhong, H., Wei, X., & Zhang, H. (2025). Leveraging llm for enhancing document-level relation extraction with correction and completion. *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, 2364–2369.

Zhou, W., Huang, K., Ma, T., & Huang, J. (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. *Proceedings of the AAAI conference on artificial intelligence*, *35*(16), 14612–14620.