SG-XICL: COORDINATING EVIDENCE AND REASONING SCOPES IN LARGE-LABEL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) enables large language models (LLMs) to perform new tasks from a small set of demonstrations without task-specific parameter updates. With a large label space, however, a separate design question is which labels should receive dedicated demonstration evidence and explicit reasoning for a given query. We propose Subspace-Guided X-ICL (SG-XICL), which constructs a query-conditioned candidate label subspace to coordinate label-aligned demonstration evidence and candidate-wise reasoning while retaining the full task label space for final prediction. Across six language models and eight classification datasets, SG-XICL achieves higher average Accuracy than the four evaluated ICL baselines on their supported comparisons. Compared with full-label X²-ICL, SG-XICL keeps dedicated reasoning local and incurs substantially lower online token burden over common-supported conditions. These results support separating the scope of dedicated explicit processing from the full prediction space in large-label ICL.

1 INTRODUCTION

In-context learning (ICL) enables pretrained language models to perform new tasks from instructions and demonstrations without task-specific parameter updates (Brown et al., 2020). In large-label classification, however, the challenge extends beyond choosing which demonstrations to include. The model must make a decision over a broad legal label space, while the labels that warrant dedicated evidence and comparison can vary with the query. This creates a context-organization problem: which labels should receive explicit demonstration evidence and label-wise reasoning for the current query?

Existing approaches address different parts of this problem. Retrieval-based ICL adapts the demonstrations to each query (Milios et al., 2023; Li et al., 2024), explanation-augmented ICL enriches demonstrations with natural-language explanations (He et al., 2024), and X^2-ICL systematically expands label-conditioned reasoning across the full label set (Honda & Oka, 2025). Long-context work further studies ICL with substantially larger demonstration sets (Bertsch et al., 2025). These directions leave open a common structural question: how should explicit demonstration evidence and label-wise reasoning be coordinated around the labels that are plausible competitors for a particular query?

Our design starts from a distinction between the legal prediction space and the scope of dedicated explicit processing. Keeping every task label available for the final prediction does not require assigning every label dedicated demonstration evidence and label-wise reasoning for every query. We therefore treat a query-conditioned candidate label subspace as a local scope for evidence and comparison rather than as a restricted answer set.

We instantiate this design with **Subspace-Guided X-ICL (SG-XICL)**. For each query, SG-XICL constructs a candidate label subspace, selects label-aligned references, augments them with own-label explanations, and performs candidate-wise reasoning over the same local scope. The final prediction nevertheless remains over the complete task label space. Figure 1 summarizes how this organization differs from representative ICL variants.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

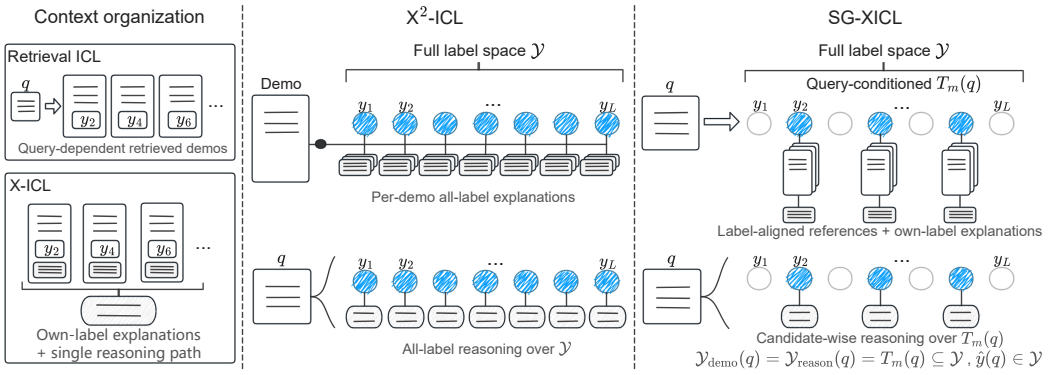


Figure 1: Comparison of context organization and explicit reasoning scope across representative ICL variants. SG-XICL uses a query-conditioned candidate label subspace to coordinate label-aligned evidence and candidate-wise reasoning while retaining the full task label space for the final decision.

Across six language models and eight classification datasets spanning multiple label scales and task families, SG-XICL achieves higher average Accuracy than the four evaluated ICL baselines on their supported comparisons. Its selected references provide higher gold-label demonstration coverage than similarity-based Retrieval ICL on all eight datasets, while its dedicated reasoning scope remains local rather than expanding over the full label set. Relative to full-label X^2-ICL, this organization also incurs substantially lower online token burden on common-supported settings. Targeted sensitivity and structural analyses further show that broader explicit processing is not uniformly beneficial under the evaluated conditions.

Together, these results motivate treating the coordination of evidence and reasoning scopes as a distinct design problem in large-label ICL.

2 Related Work

Demonstration Selection and Large-Label ICL. Demonstration selection ranks examples by semantic relevance or learned usefulness, while later work considers subset composition or coverage (Liu et al., 2022; Rubin et al., 2022; Ye et al., 2023a; Xin et al., 2026). In large-label ICL, Milios et al. study query-specific retrieval (Milios et al., 2023), and Li et al. study how demonstration number and source affect large-label ICL (Li et al., 2024). Label information can also enter selection directly: CLL-RetICL combines sentence-query and sentence-label similarity (Lin et al., 2025). In zero-shot extreme multi-label classification, ICXML constructs candidate-label sets before reranking (Zhu & Zamani, 2024), while BULLS performs bottom-up competition over a hierarchy constructed from learned label representations (Wei et al., 2027). Long-context work also studies substantially larger demonstration sets (Bertsch et al., 2025). SG-XICL assigns a different role to the candidate label subspace: it determines which labels receive dedicated demonstration evidence and candidate-wise reasoning, while the full task label space remains available for final prediction.

Explanation-Augmented and Label-Wise ICL. Another line augments ICL with explicit intermediate information. Chain-of-thought elicits reasoning traces, and natural-language explanations can support ICL, with their form and composition affecting usefulness (Wei et al., 2022; Lampinen et al., 2022; Ye et al., 2023b). X-ICL augments labeled demonstrations with natural-language explanations and uses a single query-side reasoning path (He et al., 2024; Honda & Oka, 2025), whereas X^2-ICL systematically reasons over the full label set (Honda & Oka, 2025). SG-XICL instead coordinates explanation-augmented reference evidence and candidate-wise reasoning within a query-conditioned local scope while retaining full-space prediction. Because generated explanations need not faithfully reflect the basis of a model's prediction, we treat them as explicit contextual information rather than faithful accounts of how the prediction was produced (Ye & Durrett, 2022; Turpin et al., 2023).

3 METHOD

We study in-context learning for large-label-space single-label classification. For each query, SG-XICL constructs a query-conditioned candidate label subspace that coordinates label-aligned demonstration evidence and candidate-wise reasoning while retaining the full task label space for the final decision.

Problem Setup. Let q denote an unlabeled test query and let $\mathcal{Y} = \{y_1, \ldots, y_L\}$ denote the full task label space, where $L = |\mathcal{Y}|$. We consider a single-label classification setting in which the complete task label space $\mathcal{Y}$ is known and available at inference time, and the fixed model returns one prediction $\hat{y}(q) \in \mathcal{Y}$ for each query. Let $\mathcal{D}_{\text{ref}}$ denote a labeled reference pool of pairs (u_i, y_i), where u_i contains the task input and any required task-specific conditioning information, and $y_i \in \mathcal{Y}$ is its observed label. The query q has the same task-input structure as u_i but no observed label. For each $y \in \mathcal{Y}$, define the label-specific reference pool as

$$\mathcal{D}_{\text{ref}}(y) = \{u_i : (u_i, y_i) \in \mathcal{D}_{\text{ref}},\ y_i = y\}\,.$$

Each label-specific pool is assumed to contain enough references for SG-XICL's finite selections.

SG-XICL Overview. SG-XICL organizes explicit evidence and reasoning around a query-conditioned candidate label subspace. Given q, it constructs $T_m(q)$, selects label-aligned references $S(q)$, builds the own-label explanation collection $E(q)$, and forms the candidate-indexed reasoning collection $R(q)$. Figure 2 shows this construction sequence rather than a full dependency graph.

Let $\mathcal{Y}_{\text{demo}}(q)$ denote the labels represented by the selected demonstrations, and let $\mathcal{Y}_{\text{reason}}(q)$ denote the labels that receive dedicated candidate-wise reasoning. SG-XICL uses the same candidate label subspace for both scopes:

$$\mathcal{Y}_{\text{demo}}(q) = \mathcal{Y}_{\text{reason}}(q) = T_m(q) \subseteq \mathcal{Y}, \qquad \hat{y}(q) \in \mathcal{Y}. \tag{1}$$

Equation (1) separates the local explicit scope from the legal prediction space: $T_m(q)$ determines which labels receive dedicated demonstration evidence and candidate-wise reasoning but does not restrict the final output, which remains in $\mathcal{Y}$.

Query-Conditioned Candidate Label Subspace. The first component determines which labels receive dedicated downstream evidence and reasoning. Similarity-based retrieval ranks demonstrations by query similarity and leaves the selected label composition to example-level ranking (Li et al., 2024). SG-XICL instead aggregates query-reference evidence by label to construct $T_m(q)$, which defines the local scope for downstream stages.

For each label $y \in \mathcal{Y}$, SG-XICL identifies the r references in $\mathcal{D}_{\text{ref}}(y)$ that are most similar to q. We write sim for cosine similarity between retrieval representations and use $\mathrm{Top}_k(A, g)$ to denote the k elements of A with the highest values under score g. The label score is the mean similarity over this local neighborhood, and the m highest-scoring labels form the candidate subspace:

$$\begin{gathered}\mathcal{N}_r(q, y) = \mathrm{Top}_r\left(\mathcal{D}_{\text{ref}}(y), \mathrm{sim}(q, \cdot)\right), \\ s(q, y) = \frac{1}{r} \sum_{u \in \mathcal{N}_r(q,y)} \mathrm{sim}(q, u) \qquad T_m(q) = \mathrm{Top}_m\left(\mathcal{Y}, s(q, \cdot)\right).\end{gathered} \tag{2}$$

Here, r controls the reference neighborhood used to score each label and m the candidate breadth. Averaging over the top-r references makes $s(q, y)$ depend on a local set rather than one nearest reference. The score only ranks labels; it is not a class probability or calibrated confidence score. Candidate construction does not use the query's gold label, so $T_m(q)$ need not contain it. Thus, $T_m(q)$ localizes explicit evidence and reasoning while scoring all labels in $\mathcal{Y}$.

Label-Aligned Demonstration Selection. Given $T_m(q)$, SG-XICL selects label-aligned references:

$$\begin{gathered}S_b(q, y) = \mathrm{Top}_b\left(\mathcal{D}_{\text{ref}}(y), \mathrm{sim}(q, \cdot)\right), \qquad y \in T_m(q), \\ S(q) = \bigcup_{y \in T_m(q)} \{(u, y) : u \in S_b(q, y)\}.\end{gathered} \tag{3}$$

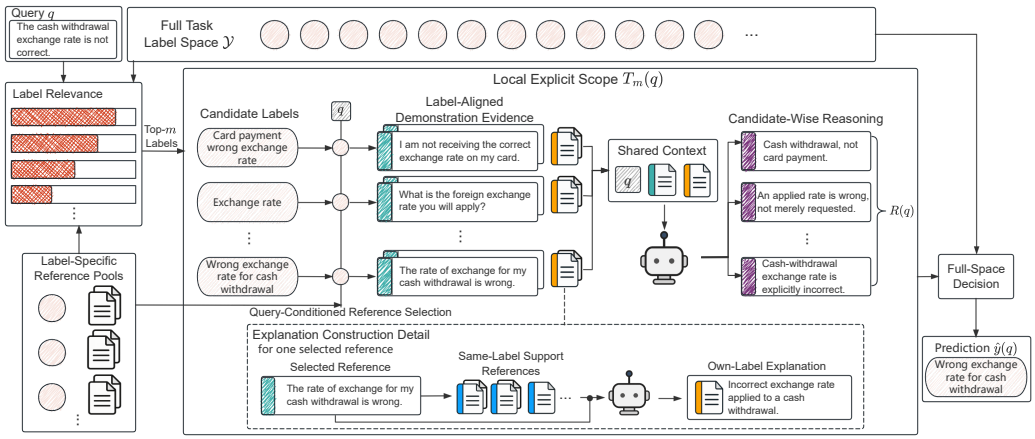


Figure 2: Overview of SG-XICL. The query-conditioned candidate label subspace coordinates label-aligned demonstration evidence and candidate-wise reasoning within a local explicit scope, while final prediction remains over the full task label space.

The parameter b controls per-label evidence depth. Using the same b for all candidates fixes the visible reference count rather than encoding upstream scores through repeated examples; it does not imply equal candidate likelihood. References in $\mathcal{D}_{\text{ref}}$ become demonstration evidence for q only when selected into $S(q)$, which is passed downstream.

Support-Conditioned Own-Label Explanations. Selected references provide input-label pairs but do not state the task-relevant evidence supporting their labels. SG-XICL therefore attaches an own-label explanation to each selected reference. X-ICL associates natural-language explanations with labeled demonstrations (He et al., 2024); SG-XICL likewise uses reference-side own-label explanations but conditions each one on additional same-label references.

For each selected reference $(u_i, y_i) \in S(q)$, SG-XICL constructs a same-label explanation-support set C_i centered on u_i rather than q, with u_i itself excluded. For support depth h,

$$C_i = \text{Top}_h\left(\mathcal{D}_{\text{ref}}(y_i) \setminus \{u_i\}, \text{sim}(u_i, \cdot)\right) \qquad e_i = G_{\text{exp}}\left(u_i, y_i, C_i\right). \tag{4}$$

The explanation e_i describes task-relevant evidence in u_i that supports its observed label y_i. It is an own-label reference explanation, not an explanation of the current query or alternative candidate labels. References in C_i condition explanation construction but do not become selected demonstrations. G_{exp} denotes an abstract explanation generator instantiated in the experimental setup.

Since q is not an input to G_{exp}, e_i can be reused whenever u_i is selected. Define $E(q) = \{e_i : (u_i, y_i) \in S(q)\}$; each e_i remains paired with u_i, and $E(q)$ depends on q only through $S(q)$.

Candidate-Wise Reasoning with Full-Space Decision. Reference explanations in $E(q)$ describe evidence for the observed labels of selected references, not the current query under alternative candidates. SG-XICL therefore introduces one query-side reasoning coordinate for each $y \in T_m(q)$. X^2-ICL explores label-conditioned reasoning over the full label set (Honda & Oka, 2025), whereas SG-XICL performs label-wise reasoning over $T_m(q)$, matching the demonstration-evidence scope.

Let G_{query} denote the query-side reasoning operator. For each $y \in T_m(q)$, it evaluates q with respect to y while conditioning on the shared localized context $S(q)$ and $E(q)$:

$$\rho_y(q) = G_{\text{query}}\left(q, y, S(q), E(q)\right), \qquad y \in T_m(q) \qquad R(q) = \left(\rho_y(q)\right)_{y \in T_m(q)}. \tag{5}$$

Each reasoning coordinate is candidate-indexed and conditioned on shared $S(q)$ and $E(q)$, so evidence from other candidates can inform the evaluation. A coordinate may support y, find insufficient evidence, or identify conflict with the query; it need not justify y. Systematic reasoning here denotes candidate-indexed coverage rather than unconstrained repeated sampling.

Equation (5) defines the explicit reasoning collection, not the final prediction rule. The model makes one decision $\hat{y}(q) \in \mathcal{Y}$ using q, $T_m(q)$, $S(q)$, $E(q)$, $R(q)$, and the full task label space. Labels outside $T_m(q)$ receive no dedicated reasoning coordinate but remain legal outputs; SG-XICL therefore uses a single full-space decision rather than a local prediction followed by a fallback stage.

4 EXPERIMENTS

We evaluate SG-XICL across multiple datasets and models, then examine how its query-conditioned local scope changes evidence access, explicit processing, within-scope information organization, and the structural coordination of local evidence and reasoning.

Experimental Setup. We evaluate eight classification datasets spanning multiple label-space sizes and task families, with 500 frozen evaluation queries per dataset and 4,000 Main queries in total. The evaluation covers six language models: DeepSeek-v4-pro, DeepSeek-v4-flash, Qwen3-14B, Qwen3-8B, Mistral-Nemo-Instruct-2407, and Mistral-7B-Instruct-v0.3. Table 1 summarizes the dataset statistics.

Table 1: Datasets and reference-pool statistics for the Main evaluation.

Dataset	Task	$\lvert\mathcal{Y}\rvert$	Reference Pool
SemEval-2010 Task 8 (Hendrickx et al., 2010)	Relation	18	7,999
DialogRE-SR (Yu et al., 2020)	Dialogue relation	30	5,819
KBP37 (Zhang & Wang, 2015)	Relation	37	17,510
Banking77 (Casanueva et al., 2020)	Intent	77	9,992
Wiki80 (Han et al., 2019)	Relation	80	50,367
FewRel (Han et al., 2018)	Relation	124	67,078
CLINC150 (Larson et al., 2019)	Intent	150	17,993
MAVEN-ETC (Wang et al., 2020)	Event type	165	77,976

We compare SG-XICL with four baselines representing fixed demonstrations, query-dependent example retrieval, explanation-augmented single-path reasoning, and systematic full-label reasoning: ICL, Retrieval ICL, X-ICL, and X^2-ICL, respectively (Li et al., 2024; He et al., 2024; Honda & Oka, 2025). Unless otherwise stated, SG-XICL uses the canonical configuration $r = 10$, $m = 8$, $b = 1$, and $h = 8$. Accuracy is the primary metric and Macro-F1 provides corroborating evidence. The randomized baselines use seeds 1000 and 2000, whereas SG-XICL uses one canonical deterministic run. All methods use the same frozen queries, legal task label vocabulary, eight-demonstration budget, and frozen model-runtime condition. Method-faithful requests exceeding the available context are recorded as unsupported rather than truncated or semantically modified. Detailed dataset preparation, baseline implementations, prompts, generation settings, and runtime configurations are provided in Appendix A and Appendix B.

Overall Performance. Table 2 compares all five methods across six models and eight datasets.

Over common-supported model–dataset cells, SG-XICL shows aggregate Accuracy differences of $+12.00$ percentage points (pp) over ICL, $+2.42$ pp over Retrieval ICL, and $+7.46$ pp over X-ICL across all 48 cells. Relative to X^2-ICL, the difference is $+16.44$ pp over the 37 cells where both methods are context-supported. Macro-F1 follows the same aggregate direction. These results establish broad complete-method effectiveness rather than universal improvement in every condition. The smaller margin over Retrieval ICL in particular hides substantial model-dependent variation: SG-XICL exceeds Retrieval ICL on all eight DeepSeek-v4-flash datasets, whereas Retrieval ICL exceeds SG-XICL on all eight Mistral-Nemo-Instruct-2407 datasets. X^2-ICL also has 11 context-unsupported model–dataset cells, which are excluded from common-supported comparisons rather than treated as zero-performance outcomes. The complete-method results therefore establish the empirical phenomenon but do not identify whether the differences arise from evidence access, evidence organization, explanation structure, candidate construction, or reasoning organization. We next examine how the methods allocate label-relevant evidence and explicit processing.

Gold-Label Evidence and Conditional Predictive Behavior. We first ask whether query-dependent example retrieval and SG-XICL expose the same label-level evidence. As shown in Figure 3, SG-XICL achieves higher gold-label demonstration coverage than Retrieval ICL on all eight evaluated datasets. The gap is strongly dataset-dependent, however: Retrieval ICL already reaches 98.9% coverage on both Banking77 and CLINC150, compared with 99.8% for SG-XICL. Thus, the result is not that Retrieval generally fails to retrieve gold-label evidence, but that example-level relevance and

Table 2: Main predictive performance in Accuracy/Macro-F1 (%). Baselines are means over two random seeds, and SG-XICL is deterministic. Values are rounded to one decimal for display, and bolding is determined from unrounded results. Bold marks the highest value for each metric within each model and dataset combination. N/A denotes a complete prompt exceeding the model-specific context limit. Average is the unweighted mean over all eight datasets and is reported only for rows with complete eight-dataset results. Seed variability is reported in Appendix C.

Model	Method	SemEval	DialogRE	KBP37	Bank77	Wiki80	FewRel	CLINC150	MAVEN	Average
DeepSeek v4-pro	ICL	82.3/83.4	79.6/73.4	50.0/52.0	82.2/81.4	75.9/75.4	70.8/69.9	95.5/95.0	55.9/53.7	74.0/73.0
	Retrieval ICL	87.4/88.3	83.0/80.2	**52.4**/54.4	91.1/90.9	80.4/79.8	79.7/79.6	**99.1/99.0**	63.0/61.5	79.5/79.2
	X-ICL	82.7/83.1	80.9/75.4	48.7/50.8	82.6/81.9	74.4/74.1	68.4/67.1	96.5/96.0	55.0/53.2	73.7/72.7
	X^2-ICL	83.0/82.9	79.7/74.4	50.8/51.3	82.7/81.9	75.3/74.8	68.2/66.7	95.6/95.1	57.1/55.2	74.1/72.8
	SG-XICL	**89.2/89.7**	**86.4/80.3**	50.4/**54.9**	**93.6/93.2**	**84.0/83.8**	**81.8/81.8**	99.0/99.0	**66.8/66.1**	81.4/81.1
DeepSeek v4-flash	ICL	63.8/72.1	81.6/76.5	33.9/38.5	78.2/79.4	74.6/75.7	67.3/68.1	91.2/92.8	52.7/50.9	67.9/69.3
	Retrieval ICL	66.0/74.5	82.4/79.9	40.4/46.7	81.0/84.9	78.8/79.9	75.2/77.2	95.9/97.1	58.7/57.7	72.3/74.7
	X-ICL	76.7/78.4	81.7/76.7	42.0/44.4	82.6/81.8	75.5/75.0	70.4/69.1	96.0/95.6	52.6/50.3	72.2/71.4
	X^2-ICL	82.0/83.1	82.0/76.1	44.1/46.4	83.1/82.1	78.5/78.0	71.7/70.6	96.8/96.4	54.9/52.6	74.1/73.2
	SG-XICL	**85.4/86.2**	**86.2/82.2**	**48.4/54.3**	**95.2/95.1**	**84.8/84.7**	**82.6/82.4**	**98.8/98.8**	**64.2/63.0**	80.7/80.8
Qwen3-14B	ICL	41.9/38.8	38.7/35.5	31.9/30.5	75.3/73.5	42.1/40.5	38.1/35.9	86.9/86.4	39.7/38.7	49.3/47.5
	Retrieval ICL	**59.5/59.4**	42.6/**42.3**	**43.1/42.5**	88.3/88.0	**61.9/62.2**	**59.1/58.6**	96.8/96.6	57.7/57.0	63.6/63.3
	X-ICL	48.8/48.0	42.1/40.4	33.8/32.7	75.1/74.0	44.7/43.8	40.1/37.7	87.6/86.4	36.7/35.3	51.1/49.8
	X^2-ICL	44.3/41.5	38.7/36.2	20.7/20.5	72.6/71.0	37.7/37.1	*N/A*	*N/A*	*N/A*	*N/A*
	SG-XICL	53.2/53.8	**44.4**/41.0	30.4/33.3	**89.0/88.5**	59.0/58.9	56.6/57.0	**97.2/97.2**	**61.4/60.0**	61.4/61.2
Mistral-Nemo Instruct-2407	ICL	33.8/27.3	41.3/39.4	28.2/24.1	69.2/68.3	41.2/40.6	31.2/29.2	83.3/81.5	34.2/33.2	45.3/43.0
	Retrieval ICL	**44.0/41.9**	**42.5/42.8**	**34.6/32.6**	**88.3/88.2**	**59.0/59.7**	**53.5/54.3**	**96.7/96.7**	**56.7/56.6**	59.4/59.1
	X-ICL	36.8/32.2	36.6/34.4	33.9/32.0	68.8/67.8	41.0/40.9	33.6/32.0	82.8/81.1	29.4/27.1	45.4/43.4
	X^2-ICL	24.8/21.3	18.5/20.7	5.6/6.7	1.1/1.8	3.2/4.5	0.0/0.0	3.4/4.8	0.0/0.0	7.1/7.5
	SG-XICL	43.4/40.4	30.8/29.6	28.6/27.4	82.6/82.0	46.6/46.9	43.0/41.8	94.0/93.8	42.6/41.7	51.5/50.5
Qwen3-8B	ICL	28.1/24.6	32.5/32.8	23.7/22.3	68.9/67.8	40.2/40.3	30.1/28.3	83.6/82.9	32.4/29.5	42.4/41.1
	Retrieval ICL	40.6/38.4	34.3/**36.6**	**31.3/32.7**	**87.7/87.4**	**58.9/58.6**	51.1/52.2	**96.8/96.6**	51.7/51.2	56.6/56.7
	X-ICL	35.6/31.0	33.9/31.4	27.8/27.2	68.7/66.6	42.0/41.9	32.0/31.0	85.0/84.0	28.9/25.8	44.2/42.4
	X^2-ICL	33.6/30.2	31.6/32.4	18.7/19.4	68.8/68.0	31.5/32.9	*N/A*	*N/A*	*N/A*	*N/A*
	SG-XICL	**45.2/45.4**	**37.2**/35.2	30.4/31.8	83.6/82.7	55.0/56.0	**55.4/54.7**	93.6/93.3	**56.6/54.9**	57.1/56.7
Mistral-7B Instruct-v0.3	ICL	21.2/19.4	1.2/2.5	9.0/9.5	23.5/28.7	24.1/23.7	15.7/16.0	38.6/42.8	17.1/18.1	18.8/20.1
	Retrieval ICL	29.2/31.2	1.1/1.8	13.9/16.0	16.2/20.4	**39.4/42.7**	**41.5/44.6**	9.5/12.8	**40.3/43.0**	23.9/26.6
	X-ICL	**35.9**/30.1	**31.5/31.7**	**18.5/16.7**	**63.7/64.0**	31.9/32.5	22.1/22.2	77.0/75.8	27.5/26.7	38.5/37.4
	X^2-ICL	21.0/20.8	22.8/21.8	12.1/11.6	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*	*N/A*
	SG-XICL	33.6/**31.3**	21.8/24.4	13.4/13.5	62.0/63.4	34.0/37.0	31.4/31.4	**79.8/81.0**	26.0/25.7	37.8/38.5

label-level evidence allocation are different objects. A plausible explanation is that similarity retrieval can devote several demonstration slots to nearby instances without ensuring representation of the label identities needed for local discrimination, whereas SG-XICL explicitly allocates references across query-conditioned candidate labels. Consistent with this distinction, SG-XICL often retains gold-label evidence on query–seed coordinates where Retrieval ICL misses it, although the conditional denominator becomes small on datasets where Retrieval coverage is already near saturation.

Higher coverage does not imply guaranteed gold-label inclusion. SG-XICL itself has imperfect coverage on several datasets, exposing a focus–coverage tradeoff: concentrating dedicated processing on a compact candidate label subspace can omit a useful label. This is why the candidate label subspace $T_m(q)$ allocates dedicated evidence and reasoning rather than defining the legal prediction space; the final decision remains over the complete task label space.

The coverage difference also creates a simple alternative explanation for the Main result: SG-XICL might differ only because it more often includes gold-label evidence. We therefore restrict the next comparison to queries for which both the seed-specific fixed ICL-family demonstrations and the SG-XICL selected references contain at least one gold-label example. This restriction matches only binary evidence availability. It does not match evidence quality, query alignment, demonstration ordering, candidate identities, explanation content, or reasoning structure, and therefore is not a fully matched causal comparison. The eligible subsets also shrink substantially on some datasets, consistent with sparse label coverage from a fixed eight-shot demonstration set, although dataset differences prevent a causal label-count interpretation.

Predictive behavior remains heterogeneous under this restriction. SG-XICL is favorable in many DeepSeek conditions, whereas Mistral-Nemo shows a different pattern and X-ICL remains stronger in several cases. Binary gold-label evidence availability is therefore insufficient to account for the observed method differences. Residual variation can plausibly reflect the quality and alignment of available evidence, the organization of local competition, and model-dependent responses to

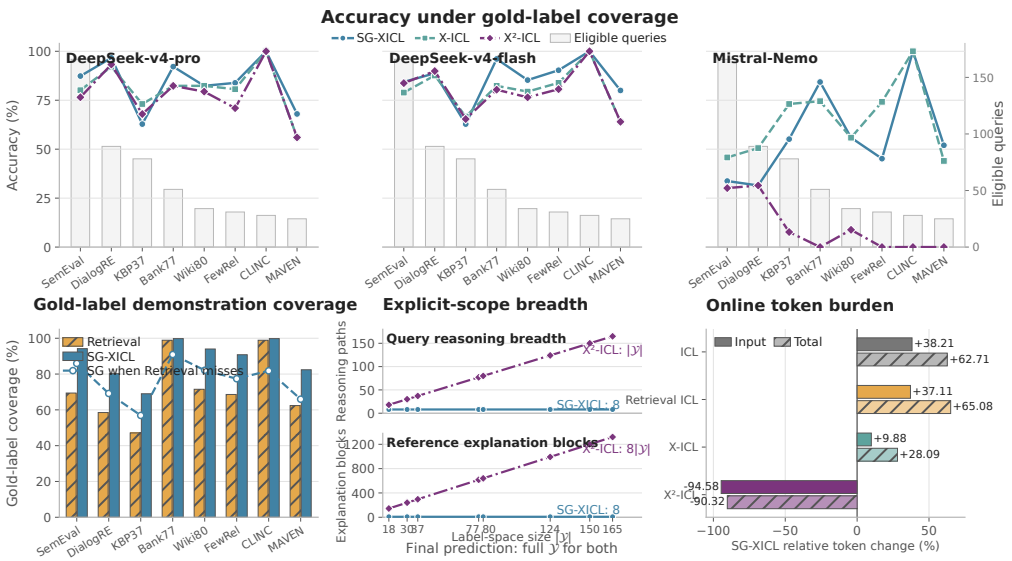


Figure 3: Evidence access and explicit-processing scope. Top: Accuracy on seed-1000 queries where both the fixed ICL-family demonstrations and SG-XICL references contain a gold-label example, for the three models with X^2-ICL supported on all eight datasets; bars show eligible-query counts. Bottom left: gold-label coverage (Retrieval: 500 queries $\times$ 2 seeds; SG-XICL: 500 canonical queries; dashed SG curve conditions on Retrieval misses). Bottom center: SG-XICL uses eight candidate-wise reasoning coordinates and eight visible own-label reference explanations, whereas X^2-ICL reasons over all $|\mathcal{Y}|$ labels and exposes $8|\mathcal{Y}|$ label-conditioned reference explanations; both predict over the full task label space $\mathcal{Y}$. Bottom right: median relative input/total token change of SG-XICL, where positive means greater use. Zero-valued Mistral-Nemo X^2-ICL points are measured supported outcomes.

reasoning scaffolds. Prior X-ICL results suggest that models can differ in how effectively they exploit explanation augmentation, while X^2-ICL additionally requires systematic label-wise multi-path reasoning (He et al., 2024; Honda & Oka, 2025). The present cross-model behavior is consistent with such model–scaffold interactions, but our experiment does not isolate model capability as a causal factor. The remaining differences motivate examining not only which evidence is available, but how much of the label space receives dedicated explicit processing.

Explicit Scope and Online Token Burden. Under the canonical configuration, SG-XICL uses eight candidate-wise reasoning coordinates and eight visible own-label reference explanations for each query. X^2-ICL instead assigns dedicated reasoning to all $|\mathcal{Y}|$ labels and exposes $8|\mathcal{Y}|$ label-conditioned reference explanations. Both methods nevertheless retain the full task label space $\mathcal{Y}$ for final prediction. The candidate label subspace $T_m(q)$ therefore defines a query-conditioned explicit-processing budget rather than a restricted answer set.

This difference in scope produces a corresponding hierarchy in online token burden. SG-XICL uses more online tokens than ICL, Retrieval ICL, and X-ICL because it introduces label-aligned explanations and candidate-wise reasoning. Relative to full-label X^2-ICL, however, it reduces median input-token usage by 94.58% and total-token usage by 90.32% over common-supported conditions. SG-XICL therefore occupies an intermediate regime: it spends additional context on local evidence organization and explicit comparison while avoiding full-label expansion. These measurements characterize online token burden rather than latency, monetary cost, throughput, or end-to-end computational efficiency. Taken together, the Main analyses suggest that effective large-label ICL depends not only on obtaining relevant evidence, but on coordinating which labels receive that evidence and how much of the label space is expanded into explicit reasoning. Complete numerical, support, and token details for Figure 3 are provided in Appendix D.

Sensitivity to Evidence Depth and Explanation Scope. Having established a compact query-conditioned explicit scope, we next ask whether simply placing more information inside that scope improves SG-XICL. We conduct a targeted sensitivity analysis on FewRel and MAVEN-ETC using DeepSeek-v4-pro, Qwen3-14B, and Mistral-Nemo-Instruct-2407 with 200 frozen queries per dataset.

We probe two orthogonal expansions while preserving the candidate label subspace: increasing the depth of same-label reference evidence and broadening the labels discussed by each reference explanation.

For evidence depth, we increase the number of same-label references per candidate from the canonical $b = 1$ setting to $b = 2, 3, 4$, while keeping the candidate label subspace and query-side reasoning coordinates fixed. Predictive performance improves initially, the incremental gain becomes smaller, and the aggregate response slightly declines at $b = 4$. Macro-F1 follows the same qualitative pattern, with substantial model–dataset heterogeneity. The results therefore support finite early gains from deeper label-aligned evidence rather than a universal monotonic depth effect. Early additional references may provide complementary realizations of a label, whereas later additions provide diminishing marginal information. This is a bounded interpretation rather than a direct measurement of redundancy. The canonical $b = 1$ setting remains the frozen Main configuration and was not selected as an optimum from this sweep.

The amount of model-visible context behaves differently from prediction quality. Input and total token burden increase monotonically with evidence depth even after the predictive improvement diminishes and becomes non-monotonic. Thus, predictive return saturates before context expansion does. This result does not define an optimal performance–token tradeoff, but it shows that adding more evidence within the local scope continues to incur online context cost after the aggregate predictive return becomes limited.

We then broaden the role of each reference explanation while keeping the candidate label subspace, selected references, query-side reasoning targets, and full task label space fixed. Replacing canonical own-label explanations with explanations covering all labels in the candidate label subspace changes aggregate Accuracy by -0.80 pp and Macro-F1 by -0.71 pp on the paired common-supported panel, while increasing median input tokens by 168.00% and total tokens by 150.38%. The cell-level effects are heterogeneous, so the result does not imply that broader explanations are uniformly harmful. It instead shows that substantially expanding reference-side explanation breadth provides no aggregate predictive improvement on this targeted panel despite a large increase in model-visible context.

Together, the depth and breadth analyses show that more explicit information within a fixed local scope is not uniformly more useful. Its value depends on quantity, placement, and the role assigned to that information. The pattern is consistent with a role-specialized organization in which references primarily provide label-aligned grounding, while cross-label comparison is concentrated in query-side candidate-wise reasoning rather than repeated within every reference explanation. We treat this as a design interpretation rather than a causal mechanism claim. This observation motivates the final question: if simply adding more information is insufficient, does the structural organization of the local scope itself matter?

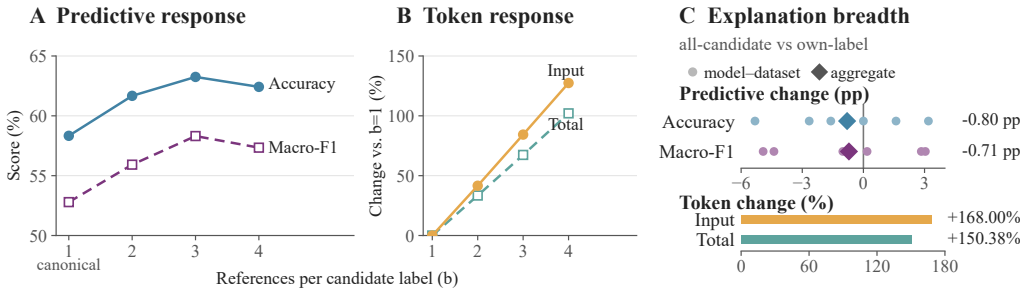


Figure 4: Within-scope sensitivity of SG-XICL. The targeted sensitivity panel uses DeepSeek-v4-pro, Qwen3-14B, and Mistral-Nemo-Instruct-2407 on FewRel and MAVEN-ETC, with a deterministic 200-query projection from each frozen Main-500 set. **(A)** Aggregate Accuracy and Macro-F1 across the six model–dataset cells as same-label evidence depth increases from canonical $b = 1$ to $b = 4$; $b = 1$ reuses the corresponding Main SG-XICL outputs, while $b > 1$ are sensitivity conditions. **(B)** Input- and total-token changes on the same 200-query panel, measured relative to $b = 1$. **(C)** All-candidate versus own-label explanation effects on the exact 188/200 paired common-support queries in each cell; predictive differences are all-candidate minus own-label, and token changes use the same paired universe. Full cell-level results, support accounting, and aggregation details are provided in Appendix E.

Ablation of Local-Scope Coordination. We evaluate four predefined structural contrasts on DeepSeek-v4-flash across MAVEN-ETC, SemEval-2010 Task 8, and DialogRE-SR, using the same 500 frozen Main queries per dataset. These contrasts answer different structural questions and do not form an additive component decomposition.

Table 3: Structural contrasts on DeepSeek-v4-flash across MAVEN, SemEval, and DialogRE. Each cell reports Accuracy/Macro-F1 Δ (pp) in the indicated direction. Rows are predefined non-additive contrasts; all conditions retain full-$\mathcal{Y}$ prediction.

Contrast	MAVEN	SemEval	DialogRE	Mean
Label salience (C1–C0)	-1.20/-0.81	-0.60/-1.24	+1.20/+1.09	-0.20/-0.32
Reasoning structure (C4–C1)	+2.80/+2.54	+5.60/+6.90	+22.00/+23.11	+10.13/+10.85
Own-label evidence (C4–C2)	+2.60/+2.58	+25.00/+26.45	+22.40/+23.45	+16.67/+17.49
Candidate construction (C4–C3)	+1.40/+0.92	+18.40/+19.15	+22.40/+22.38	+14.07/+14.15

We first test whether a candidate label subspace helps merely because it makes a few plausible labels more salient. The C1–C0 contrast exposes the same candidate labels more explicitly while retaining a generic single reasoning path. Its mixed and near-neutral pattern does not support a consistent benefit from candidate visibility alone. A useful local scope therefore requires more than making plausible labels salient.

We next examine how the local scope is used once it has been constructed. The C4–C1 reasoning-structure contrast favors canonical SG-XICL on all three datasets, indicating that explicit candidate-wise comparison differs materially from a generic reasoning path. The C4–C2 own-label-evidence contrast is also positive on all three datasets. Importantly, C2 retains the same selected references and candidate-wise reasoning but removes their visible own-label explanations, so this contrast concerns explicit label-aligned grounding rather than raw reference presence. Both contrasts show substantial variation in magnitude across datasets, and Macro-F1 follows the same qualitative direction. Taken together, they are consistent with complementary roles: reference-side information grounds candidate labels, while query-side candidate-wise reasoning compares the grounded alternatives. This interpretation also aligns with the explanation-breadth analysis above, where repeating cross-label comparison inside every reference added substantial context without improving aggregate prediction.

Finally, the C4–C3 contrast asks whether any plausible query-dependent candidate construction is sufficient. C3 scores each label using its single nearest reference, whereas canonical SG-XICL aggregates support from multiple nearby references. The canonical construction is favored on all three datasets. This indicates that candidate construction is not merely a preprocessing detail because it determines which labels subsequently receive aligned evidence and dedicated comparison. Neighborhood aggregation may be less sensitive to a single extreme match than a single-nearest rule, but the experiment does not directly isolate that mechanism. Changes in candidate membership also propagate to the downstream evidence and reasoning context, so this contrast should be interpreted as an upstream structural alternative rather than the isolated contribution of a scoring function.

Overall, the ablations support a coordinated local structure rather than a ranking of independent components. Candidate construction determines **what** should be compared, label-aligned evidence determines how those candidates are **grounded**, and candidate-wise reasoning determines how they are **compared**. Illustrative query-level success and failure cases are provided in Appendix G.

5 Conclusion

In this paper, we introduce SG-XICL, a framework for coordinating label-aligned demonstration evidence and candidate-wise reasoning in large-label ICL. SG-XICL constructs a query-conditioned candidate label subspace to coordinate dedicated demonstration evidence and candidate-wise reasoning while retaining the full task label space for final prediction. Across six models and eight datasets, SG-XICL achieves higher average Accuracy than the four evaluated ICL baselines on their supported comparisons. These findings support a broader principle for large-label ICL: dedicated explicit reasoning need not expand with the full decision space, but can instead focus on plausible competitors for the current query, enabling selective explicit processing without restricting the final prediction space.

AI USE STATEMENT

Generative AI tools were used to assist with and polish the writing of this paper. In project code development, generative AI was used only in an auxiliary role; the core design, implementation structure, and final construction of the codebase were completed and reviewed by the authors. No other part of the research process relied on generative AI, and the authors take responsibility for the final paper and project artifacts.

REPRODUCIBILITY STATEMENT

The formal SG-XICL procedure and notation are specified in Section 3, and the shared evaluation protocol is summarized in Section 4. Appendix A documents dataset construction, deterministic evaluation-scope sampling, baseline implementations, model and runtime conditions, and parsing and validity rules, while Appendix B provides the prompt and explanation-generation templates used in the experiments. Appendices C–F provide the complete quantitative results and supporting evidence for the Main, evidence-and-scope, sensitivity, and ablation analyses. An accompanying supplementary code package provides implementation and evaluation resources for reproducing and auditing the reported experimental pipeline. Because fixed version identifiers were not available or recorded for every API-hosted model and upstream model checkpoint, future reruns may not reproduce every model-generated response exactly. The reported experimental procedure, configurations, and evaluation pipeline are documented to support scientific reproduction, but exact output-level replay across all environments is not claimed.

REFERENCES

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12119–12149, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd workshop on natural language processing for conversational AI*, pp. 38–45, 2020.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4803–4809, 2018.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. Opennre: An open and extensible toolkit for neural relation extraction. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP): system demonstrations*, pp. 169–174, 2019.

Xuanli He, Yuxiang Wu, Oana-Maria Camburu, Pasquale Minervini, and Pontus Stenetorp. Using natural language explanations to improve robustness of in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13477–13499, 2024.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Katrin Erk and Carlo Strapparava (eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006/.

Ukyo Honda and Tatsushi Oka. Exploring explanations improves the robustness of in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23693–23714, 2025.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, 2022.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1311–1316, 2019.

Zhan Li, Fanghui Liu, Volkan Cevher, and Grigorios Chrysos. Demonstrations in in-context learning for llms with large label space. In *First Workshop on Long-Context Foundation Models@ ICML 2024*, 2024.

Chaohao Lin, Kaida Wu, Peihao Xiang, Yanzhao Wu, and Ou Bai. Cll-reticl: Contrastive linguistic label retrieval-based in-context learning for text classification via large language models. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1575–1590, 2025.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 173–184, 2023.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 2655–2671, 2022.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. Maven: A massive general domain event detection dataset. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1652–1671, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Mingzhu Wei, Rui Song, and Rui Mao. Bottom-up competition based on label hierarchy trees for in-context learning with large label space. *Information Fusion*, 138:104727, 2027. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2026.104727. URL https://www.sciencedirect.com/science/article/pii/S1566253526006032.

Jiayi Xin, Xiang Li, Evan Qiang, Weiqing He, Tianqi Shang, Weijie J Su, and Qi Long. Ucs: Estimating unseen coverage for improved in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10965–10981, 2026.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pp. 39818–39833. PMLR, 2023a.

Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392, 2022.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4469–4484, 2023b.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4927–4940, 2020.

Dongxu Zhang and Dong Wang. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*, 2015.

Yaxin Zhu and Hamed Zamani. ICXML: An in-context learning framework for zero-shot extreme multi-label classification. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2086–2098, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.134. URL https://aclanthology.org/2024.findings-naacl.134/.

A Experimental and Reproducibility Details

This appendix specifies data construction, the fixed evaluation scope, shared scientific assets, baseline implementations, model-runtime conditions, and output validation for the Main experiment. It complements the compact setup in the main text without redefining methods or results.

A.1 Dataset Construction and Source Mapping

Table 4 records the prepared reference/test construction and task adaptation for the eight task views. Relation-classification views preserve entity order and relation direction. Cross-role task-identical duplicates are removed from the reference side while retaining the test record. MAVEN documents and DialogRE dialogues are not split across reference and test roles.

Dataset	Reference/test construction and task adaptation
FewRel	Reference: deterministic holdout remainder. Test: deterministic label-stratified holdout. Sentence plus ordered entity pair to one relation label; holdout split seed 42.
Wiki80	Reference: train. Test: validation. Sentence plus ordered entity pair to one relation label.
CLINC150	Reference: train + validation. Test: test split. 150 in-scope intents; out-of-scope examples and the OOS label are excluded.
Banking77	Reference: train. Test: test split. All 77 intents are retained without merging semantically similar intents.
KBP37	Reference: train + validation. Test: test split. Sentence plus ordered entity pair to one relation label; `no_relation` remains a valid canonical label.
SemEval-2010 Task 8	Reference: train. Test: test split. Sentence plus ordered entity pair to one relation label; `Other` remains a valid canonical label.
MAVEN-ETC	Reference: official train adapted view. Test: official validation adapted view. Event-type classification for a provided marked positive trigger, rather than full event detection.
DialogRE-SR	Reference: official train + dev adapted view. Test: official test adapted view. One ordered argument pair is predicted within its dialogue; the full dialogue may be used as task input.

Table 4: Reference/test construction and task-specific adaptation used by the Main experiment.

A.2 Main Evaluation Scope and Deterministic Sampling

The Main scope contains exactly 500 fixed queries per dataset, for 4,000 queries in total. The subset is constructed once from dataset ID, query ID, and the prepared gold label used only for evaluation-set stratification. Each canonical label receives one query first. The remaining quota is allocated approximately in proportion to residual label capacity using deterministic largest-remainder allocation with canonical-label-ID tie breaking. Within each label, queries are ordered by a fixed SHA256 key derived from the dataset and query identifiers. No model prediction, embedding, similarity

score, prompt length, context-support outcome, latency, cost, or downstream metric enters scope construction.

Gold labels are used by this one-time scope builder only. After the query IDs are fixed, downstream selection, prompt construction, context-support decisions, and inference operate on a gold-free active query view. The same query IDs are shared across all models, methods, and applicable seeds, and no performance-based resampling or model-specific substitution is permitted.

A.3 Models, Shared Assets, and Canonical Configuration

The four randomized baselines use seeds 1000 and 2000. SG-XICL uses one canonical deterministic run per model×dataset setting and is not duplicated as pseudo-replicates. In paper-facing notation, the canonical SG-XICL configuration is

$$r = 10, \qquad m = 8, \qquad b = 1, \qquad h = 8,$$

where h is the number of same-label support examples used to construct each own-label reference explanation. All methods retain the complete canonical task label vocabulary as the legal prediction space.

Similarity-based components use Qwen3-Embedding-8B. Embedding inference uses bfloat16 with SDPA, a maximum input length of 1,024 tokens, L2-normalized 4,096-dimensional outputs stored in float32, and exact cosine similarity. Explanation-based methods use DeepSeek-v4-pro as the shared offline explanation generator where required, with temperature 0, thinking enabled, and a maximum output budget of 384,000 tokens. These explanations are generated independently of the formal test-query content and may be reused when the same reference is selected by different queries.

A.4 Baseline Implementation and Fidelity

The four baselines are instantiated as method-faithful implementations under the shared evaluation protocol. They preserve their defining demonstration, explanation, and reasoning structures while sharing the fixed Main query scope, complete legal label vocabulary, eight-demonstration budget, and the corresponding model-runtime condition.

Baseline	Implementation and fidelity control
ICL	Eight fixed demonstrations per active seed, constructed by stable seed-indexed SHA256 ranking over reference IDs. Seeds 1000 and 2000 use disjoint demonstration sets within each dataset.
Retrieval ICL	For each dataset and active seed, a query-independent finite pool is formed with seven references per canonical label by stable seed-controlled SHA256 ranking. The pool is flattened and the query-specific top eight examples are selected by exact cosine similarity. Ties are resolved by similarity descending and then canonical record ID ascending. No unique-label post-filter is applied, so repeated labels are allowed.
X-ICL	Uses the same eight demonstration IDs and ordering as ICL under the same active seed. The model-visible change is the addition of own-label reference explanations and one query-side reasoning path. Exact prompt rendering is given in Appendix B.
X^2-ICL	Uses the same eight demonstration IDs and ordering as ICL and X-ICL under the same active seed. Each demonstration is paired with label-conditioned explanations over the complete canonical label space, with full-label query-side reasoning. A method-faithful request that is unsupported under the fixed runtime is recorded as `N/A_ctx`, not truncated or simplified.

Table 5: Baseline implementation and fidelity controls used in the Main experiment.

A.5 Runtime and Generation Configuration

Each online model is evaluated under one fixed Main runtime condition shared across methods. Optional context-extension mechanisms are not enabled method-wise or request-wise merely to make a longer prompt runnable. A complete method-faithful request that exceeds the fixed model condition is recorded as context-unsupported rather than truncated or simplified.

The two API conditions use `deepseek-v4-pro` and `deepseek-v4-flash`. Both use temperature 0, thinking enabled, and a maximum output budget of 384,000 tokens, with no project-side context extension. Immutable provider revisions were not pinned, so exact replay of provider-generated responses is not claimed.

Local model	Context	Output	Core decoding condition
Qwen3-14B	40,960	8,230	bfloat16, temperature 0, thinking disabled, YaRN disabled.
Qwen3-8B	40,960	8,230	bfloat16, temperature 0, thinking disabled, deterministic decoding.
Mistral-Nemo-Instruct-2407	131,072	69,736	bfloat16, temperature 0, no thinking mode.
Mistral-7B-Instruct-v0.3	32,768	4,387	bfloat16, temperature 0, no thinking mode.

Table 6: Local-model Main runtime and generation conditions. Context and output budgets are in tokens.

The local checkpoint identifiers are `Qwen/Qwen3-14B`, `Qwen/Qwen3-8B`, `mistralai/Mistral-Nemo-Instruct-2407`, and `mistralai/Mistral-7B-Instruct-v0.3`. The listed context limits correspond to the default checkpoint runtime used in Main rather than optional extended-context conditions. Immutable upstream revision identifiers were not uniformly retained, so exact bitwise replay is not claimed.

A.6 Parsing and Validity Controls

Classification is recovered only from dedicated `Label:` fields. After surrounding whitespace is removed, a value must exactly match one canonical task label. The parser performs no alias matching, fuzzy matching, semantic correction, case folding, label rewriting, or nearest-label repair. A `Label:` field need not be the last nonempty response line. Repeated identical canonical labels remain recoverable, whereas conflicting canonical labels make the prediction invalid.

A prediction is valid whenever the uniquely parsed final label belongs to the complete canonical task vocabulary. In particular, an SG-XICL prediction outside $T_m(q)$ remains legal if it belongs to the full task label space $\mathcal{Y}$. Parser-invalid outputs and terminal execution failures remain in the evaluation denominator and count as incorrect. Context-unsupported conditions receive no predictive score. Macro-F1 is computed over the complete canonical label vocabulary rather than only labels observed in predictions.

B Prompt Templates

This appendix records the prompt templates and their compositional rendering contract used in the experiments. Every Main online prompt combines a dataset-specific task fragment, the complete canonical label catalog, method-specific demonstrations or references, a dataset-specific query fragment, and the method-specific decision instruction. The formal query gold label is never inserted into the prompt.

B.1 Shared Prompt Rendering Contract

All Main online methods use the following exact system message:

```
Follow the provided task and examples. Ensure that your response ends with Label: followed
    by your final answer.
```

All offline explanation generation uses the following exact system message:

```
Follow the provided task and examples. Output only the requested reason text after Reason:.
```

The complete canonical label catalog is rendered under the literal header `Complete label space:`, with one exact canonical label per line in canonical label-ID order. The renderer does not lowercase labels, replace underscores, expand aliases, or reorder labels by query score. The exact block delimiter is a line containing only `###`. Dataset fragments and method templates are composed without paraphrasing or reordering.

All five Main methods retain the full task label space as the legal prediction space. For SG-XICL, $T_m(q)$ defines only the dedicated demonstration and reasoning scope; the canonical Main configuration uses $m = 8$.

Method	Model-visible evidence, query reasoning, and legal labels
ICL	Demonstration input + observed label; no explicit query reasoning request; final label from complete $\mathcal{Y}$.
Retrieval ICL	Retrieved demonstration input + observed label; no explicit query reasoning request; final label from complete $\mathcal{Y}$.
X-ICL	Demonstration input + own-label explanation + observed label; one concise query reasoning path; final label from complete $\mathcal{Y}$.
X^2-ICL	Demonstration input + full-label-conditioned explanations + observed label; query reasoning over complete $\mathcal{Y}$; final label from complete $\mathcal{Y}$.
SG-XICL	Label-aligned reference + own-label explanation + observed label; candidate-wise reasoning over $T_m(q)$; final label from complete $\mathcal{Y}$.

Table 7: Model-visible evidence and requested query-side reasoning in the five Main methods.

B.2 Main Online Prompt Templates

The templates below are exact composition templates. Placeholders and loops are expanded deterministically before model execution. They are not sent to the model in unexpanded form.

ICL.

```
{{ TASK_AND_EVIDENCE_FRAGMENT }}

Complete label space:
{% for label in labels %}
- {{ label }}
{% endfor %}

{% for demo in demonstrations %}
###
Example {{ loop.index }}:
{{ TASK_DEMO_INPUT(demo) }}
Label: {{ demo.label }}
{% endfor %}
###
Query:
{{ TASK_QUERY_INPUT(query) }}

Select one final label from the complete label space.
```

The demonstration order is preserved exactly as supplied by the active experiment, and no reason line is shown.

Retrieval ICL. Retrieval ICL uses the same model-visible natural-language template as ICL. Only the upstream demonstration selection differs.

X-ICL.

```
{{ TASK_AND_EVIDENCE_FRAGMENT }}

Complete label space:
{% for label in labels %}
- {{ label }}
{% endfor %}

{% for demo in demonstrations %}
###
Example {{ loop.index }}:
{{ TASK_DEMO_INPUT(demo) }}
Reason: {{ demo.own_label_reason }}
Label: {{ demo.label }}
{% endfor %}
###
Query:
{{ TASK_QUERY_INPUT(query) }}

Give one concise reason for the best-supported label, then select one final label from the
     complete label space.
```

Within an active seed, X-ICL uses the same demonstration IDs and ordering as ICL. Each demonstration exposes exactly one reason associated with its observed label.

X^2-ICL.

```
{{ TASK_AND_EVIDENCE_FRAGMENT }}

Complete label space:
{% for label in labels %}
- {{ label }}
{% endfor %}

{% for demo in demonstrations %}
###
Example {{ loop.index }}:
{{ TASK_DEMO_INPUT(demo) }}
Instruction: Explore the reasoning behind all the labels. Then, select the label that has
     the most valid reasoning.
{% for label in labels %}
Possible Reasoning for {{ label }}: {{ demo.label_conditioned_reasons[label] }}
{% endfor %}
Label: {{ demo.label }}
{% endfor %}
###
Query:
{{ TASK_QUERY_INPUT(query) }}

Instruction: Explore the reasoning behind all the labels. Then, select the label that has
     the most valid reasoning.
```

Within an active seed, X^2-ICL shares demonstration IDs and ordering with ICL and X-ICL. For every demonstration, label-conditioned explanations cover the complete canonical label catalog in canonical label-ID order. Query-side reasoning likewise ranges over the full task label space. The prompt is not reduced to a top-m or SG-style local scope.

SG-XICL.

```
{{ TASK_AND_EVIDENCE_FRAGMENT }}

Complete label space:
{% for label in labels %}
- {{ label }}
{% endfor %}

{% for reference in references %}
###
Example {{ loop.index }}:
{{ TASK_DEMO_INPUT(reference) }}
Reason: {{ reference.support_conditioned_own_label_reason }}
Label: {{ reference.label }}
{% endfor %}
###
Query:
{{ TASK_QUERY_INPUT(query) }}

Evaluate one concise possible reasoning sentence for each distinct label represented by the
     references, using the references and provided reasons as task-relevant evidence. Judge
     whether the query supports each label, rather than choosing by resemblance to a
     reference. Then select one final label from the complete label space.

Output exactly one reasoning line for each represented label in the following order:

{% for label in candidate_labels %}
Possible Reasoning for {{ label }}: <one concise sentence>
{% endfor %}
Label: <one canonical label from the complete label space>
```

Under the canonical Main configuration $m = 8$, the candidate labels are supplied upstream and rendered in canonical label-ID order. References are aligned one-to-one with those candidate labels in the same order. Each reference exposes exactly one support-conditioned own-label explanation, and the prompt requests one reasoning coordinate per candidate label.

B.3 Dataset-Specific Task and Input Fragments

The following fragments are shared across Main methods. Each dataset entry gives the exact task/evidence text, online demonstration/reference fields, online query fields, offline target/meta/support fields, full-meta expert instruction, and support-conditioned own-label objective. The complete label catalog is supplied separately by the shared rendering contract and is therefore not duplicated here.

FewRel. *Task and evidence.*

```
Task:
Classify the semantic relation expressed from the marked [HEAD] entity to the marked [TAIL]
     entity in the sentence.

Relevant evidence:
Use the sentence context and the ordered roles of the marked entities to determine which
     relation from HEAD to TAIL is supported by the sentence.
Do not classify the overall topic or source domain of the sentence.
Do not reverse the HEAD and TAIL entities.
```

Online demonstration/reference input.

```
Sentence: {{ item.text }}
Head entity: {{ item.head_entity }}
Tail entity: {{ item.tail_entity }}
```

Online query input.

```
Sentence: {{ query.text }}
Head entity: {{ query.head_entity }}
Tail entity: {{ query.tail_entity }}
```

Offline target/meta/support input.

```
Sentence: {{ item.text }}
Head entity: {{ item.head_entity }}
Tail entity: {{ item.tail_entity }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on FewRel-Public-All relation classification tasks.
     Given the task input and the corresponding label, please write a concise and precise
     reason to explain why the label is assigned to the example by following the provided
     examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the target sentence supports the assigned canonical relation from
     the marked HEAD entity to the marked TAIL entity.
Write one concise and precise reason explaining why the target example has the assigned
     label. Explain the target itself using task-relevant evidence.
```

Wiki80. *Task and evidence.*

```
Task:
Classify the semantic relation expressed from the marked [HEAD] entity to the marked [TAIL]
     entity in the sentence.

Relevant evidence:
Use the sentence context and the ordered roles of the marked entities to determine which
     relation from HEAD to TAIL is supported by the sentence.
Do not classify the overall topic of the sentence.
Do not reverse the HEAD and TAIL entities.
```

Online demonstration/reference input.

```
Sentence: {{ item.text }}
Head entity: {{ item.head_entity }}
Tail entity: {{ item.tail_entity }}
```

Online query input.

```
Sentence: {{ query.text }}
Head entity: {{ query.head_entity }}
Tail entity: {{ query.tail_entity }}
```

Offline target/meta/support input.

```
Sentence: {{ item.text }}
Head entity: {{ item.head_entity }}
Tail entity: {{ item.tail_entity }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on Wiki80 relation classification tasks. Given the task
    input and the corresponding label, please write a concise and precise reason to explain
    why the label is assigned to the example by following the provided examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the sentence supports the assigned relation from the marked HEAD
    entity to the marked TAIL entity.
Write one concise and precise reason explaining why the target example has the assigned
    label. Explain the target itself using task-relevant evidence.
```

CLINC150. *Task and evidence.*

```
Task:
Classify the single in-scope intent expressed by the user utterance.
Every utterance in this task belongs to exactly one intent in the complete label space.

Relevant evidence:
Use the user's requested action, information need, target object or service, stated
    condition or status, and relevant qualifiers to determine the intent.
Distinguish intents that concern the same topic but request different actions, information,
    or outcomes.
Classify what the user is asking or trying to accomplish. Do not choose an intent merely
    because the utterance contains words associated with that intent's general domain.
```

Online demonstration/reference input.

```
User utterance: {{ item.text }}
```

Online query input.

```
User utterance: {{ query.text }}
```

Offline target/meta/support input.

```
User utterance: {{ item.text }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on CLINC150 in-scope intent classification tasks. Given
    the task input and the corresponding label, please write a concise and precise reason
    to explain why the label is assigned to the example by following the provided examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the utterance expresses the assigned intent, identifying the
    user's requested action or information need and any relevant target, status, or
    qualifier.
Write one concise and precise reason explaining why the target example has the assigned
    label. Explain the target itself using task-relevant evidence.
```

Banking77. *Task and evidence.*

```
Task:
Classify the single banking customer-service intent expressed by the customer query.
Every query in this task belongs to exactly one intent in the complete label space.

Relevant evidence:
Use the customer's requested action or information need, the banking product or transaction
    channel involved, the stated transaction or service state, the reported problem, and
    the requested outcome to determine the intent.
Distinguish intents that concern the same banking object but differ in transaction channel,
    lifecycle state, problem type, direction, fee, recognition, timing, or desired
    resolution.
Classify the issue or request actually expressed by the customer. Do not choose an intent
    merely because the query contains words commonly associated with that intent.
```

Online demonstration/reference input.

```
Customer query: {{ item.text }}
```

Online query input.

```
Customer query: {{ query.text }}
```

Offline target/meta/support input.

```
Customer query: {{ item.text }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on BANKING77 intent classification tasks. Given the
    task input and the corresponding label, please write a concise and precise reason to
    explain why the label is assigned to the example by following the provided examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the customer query expresses the assigned banking intent,
    identifying the relevant banking product or transaction channel, the requested action
    or information need, and any stated status, problem, fee, timing, direction, or desired
    outcome.
Write one concise and precise reason explaining why the target example has the assigned
    label. Explain the target itself using task-relevant evidence.
```

KBP37. *Task and evidence.*

```
Task:
Classify the exact directional relation expressed between the marked [HEAD] entity and the
    marked [TAIL] entity in the sentence.

Entity-role mapping:
[HEAD] corresponds to e1.
[TAIL] corresponds to e2.
A label ending in (e1,e2) uses HEAD/e1 as the first relation argument and TAIL/e2 as the
    second relation argument.
A label ending in (e2,e1) uses TAIL/e2 as the first relation argument and HEAD/e1 as the
    second relation argument.

Relevant evidence:
Use the sentence context and the marked entity roles to determine which relation is
    expressed between the two entities and whether its argument direction is (e1,e2) or
    (e2,e1).
Preserve the e1/HEAD and e2/TAIL mapping when interpreting the directional label suffix.
Use no_relation only when the sentence does not support any listed directional relation for
    the marked entity pair.
Do not classify the overall topic of the sentence, and do not infer a relation solely from
    the entity names or entity types.
```

Online demonstration/reference input.

```
Sentence: {{ item.text }}
Head entity (e1): {{ item.head_entity }}
Tail entity (e2): {{ item.tail_entity }}
```

Online query input.

```
Sentence: {{ query.text }}
Head entity (e1): {{ query.head_entity }}
Tail entity (e2): {{ query.tail_entity }}
```

Offline target/meta/support input.

```
Sentence: {{ item.text }}
Head entity (e1): {{ item.head_entity }}
Tail entity (e2): {{ item.tail_entity }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on KBP37 directional relation classification tasks.
    Given the task input and the corresponding label, please write a concise and precise
    reason to explain why the label is assigned to the example by following the provided
    examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the sentence supports the assigned relation type and exact
     argument direction for e1/HEAD and e2/TAIL. For no_relation, explain why the sentence
     does not establish any listed directional relation for the marked entity pair.
Write one concise and precise reason explaining why the target example has the assigned
     label. Explain the target itself using task-relevant evidence.
```

SemEval-2010 Task 8. *Task and evidence.*

```
Task:
Classify one canonical semantic relation label for the marked [HEAD] nominal and the marked
     [TAIL] nominal in the sentence.
For a directional relation label, preserve the exact semantic relation type and argument
     direction. Use Other when the sentence does not support any of the listed directional
     relations for the marked nominal pair.

Nominal-role mapping:
[HEAD] corresponds to e1.
[TAIL] corresponds to e2.
A label ending in (e1,e2) uses HEAD/e1 as the first relation argument and TAIL/e2 as the
     second relation argument.
A label ending in (e2,e1) uses TAIL/e2 as the first relation argument and HEAD/e1 as the
     second relation argument.

Relevant evidence:
Use the sentence context and the roles of the two marked nominals to determine which
     canonical relation label is supported.
Preserve the e1/HEAD and e2/TAIL mapping when interpreting a directional label suffix.
Treat a directional relation as supported only when it is asserted or entailed by the
     sentence. Do not treat a relation that is only negated, hypothetical, or speculative as
     present.
Select one canonical label from the complete label space. For a directional relation label,
     preserve its exact argument direction. Use Other when the sentence does not support any
     of the listed directional relations for the marked nominal pair.
Do not classify the overall topic of the sentence, and do not choose a relation solely from
     the nominal words or their general semantic types.
```

Online demonstration/reference input.

```
Sentence: {{ item.text }}
Head nominal (e1): {{ item.head_entity }}
Tail nominal (e2): {{ item.tail_entity }}
```

Online query input.

```
Sentence: {{ query.text }}
Head nominal (e1): {{ query.head_entity }}
Tail nominal (e2): {{ query.tail_entity }}
```

Offline target/meta/support input.

```
Sentence: {{ item.text }}
Head nominal (e1): {{ item.head_entity }}
Tail nominal (e2): {{ item.tail_entity }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on SemEval-2010 Task 8 directional semantic relation
     classification tasks. Given the task input and the corresponding label, please write a
     concise and precise reason to explain why the label is assigned to the example by
     following the provided examples:
```

Support-conditioned own-label objective.

```
For this task: For a directional relation label, explain why the sentence supports the
     assigned semantic relation type and exact argument direction for e1/HEAD and e2/TAIL.
     For Other, explain why the sentence does not support any of the listed directional
     relations for the marked nominal pair.
Write one concise and precise reason explaining why the target example has the assigned
     label. Explain the target itself using task-relevant evidence.
```

MAVEN-ETC. *Task and evidence.*

```
Task:
Classify the event type evoked by the marked [TRIGGER] span in the sentence.
Every target trigger in this task is a positive annotated event mention and belongs to
     exactly one event type in the complete label space.

Relevant evidence:
Use the meaning of the marked target trigger in its sentence context, including what
     happens, who or what is involved, and any relevant modifiers, to determine the event
     type.
Classify the event evoked by the marked trigger itself. Do not classify the overall topic of
     the sentence or an event evoked by another word.
Distinguish cases in which the same or a similar trigger expression has different event
     meanings in different contexts. Do not choose an event type only because a reference
     contains the same or a similar trigger word.
```

Online demonstration/reference input.

```
Sentence: {{ item.text }}
Target trigger: {{ item.target_trigger }}
```

Online query input.

```
Sentence: {{ query.text }}
Target trigger: {{ query.target_trigger }}
```

Offline target/meta/support input.

```
Sentence: {{ item.text }}
Target trigger: {{ item.target_trigger }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on MAVEN-ETC gold-trigger event type classification
     tasks. Given the task input and the corresponding label, please write a concise and
     precise reason to explain why the label is assigned to the example by following the
     provided examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the marked target trigger evokes the assigned event type in this
     sentence, using the contextual event meaning that distinguishes it from other possible
     event types.
Write one concise and precise reason explaining why the target example has the assigned
     label. Explain the target itself using task-relevant evidence.
```

DialogRE-SR. *Task and evidence.*

```
Task:
Classify the semantic relation expressed from the subject argument to the object argument in
     the complete dialogue.
Every retained argument pair in this task has exactly one positive relation in the complete
     label space.

Relevant evidence:
Use the complete dialogue, including the turn order, speaker identities, statements across
     multiple turns, and references to the two arguments, to determine the relation from the
     subject argument to the object argument.
Preserve the subject-object order. A relation supported in the reverse direction is not
     automatically the correct label for the given ordered pair.
Base the decision on what the dialogue supports. Do not infer a relation only from the
     argument names, argument types, general dialogue topic, or outside knowledge about the
     speakers or characters.
```

Online demonstration/reference input.

```
Dialogue:
{% for turn in item.dialogue_turns %}
  {{ turn }}
{% endfor %}

Subject argument: {{ item.subject_argument }}
Object argument: {{ item.object_argument }}
```

Online query input.

```
Dialogue:
{% for turn in query.dialogue_turns %}
  {{ turn }}
{% endfor %}

Subject argument: {{ query.subject_argument }}
Object argument: {{ query.object_argument }}
```

Offline target/meta/support input.

```
Dialogue:
{% for turn in item.dialogue_turns %}
  {{ turn }}
{% endfor %}

Subject argument: {{ item.subject_argument }}
Object argument: {{ item.object_argument }}
```

Full-meta expert instruction.

```
Assume that you're an expert working on DialogRE-SR dialogue relation classification tasks.
     Given the task input and the corresponding label, please write a concise and precise
     reason to explain why the label is assigned to the example by following the provided
     examples:
```

Support-conditioned own-label objective.

```
For this task: Explain why the complete dialogue supports the assigned relation from the
     subject argument to the object argument, using the relevant multi-turn statements,
     speaker identities, or references expressed in the dialogue.
Write one concise and precise reason explaining why the target example has the assigned
     label. Explain the target itself using task-relevant evidence.
```

B.4 Offline Explanation-Generation Templates

All offline families use the shared offline system message in Appendix B.1 and request a single `Reason:` field. Offline reason generation never receives the formal test-query gold label.

Support-conditioned own-label explanation. For a target reference, the support examples share its observed label, the target itself is excluded from the support set, and the generated reason is query-independent and reusable.

```
{{ TASK_AND_EVIDENCE_FRAGMENT }}

The following additional labeled examples share the same label as the target. Use them only
     as task-specific reference evidence for understanding the stable meaning of the label.
     Do not explain them and do not copy their details.

{% for support in supports %}
###
Support Example {{ loop.index }}:
{{ TASK_OFFLINE_INPUT(support) }}
Label: {{ support.label }}
{% endfor %}
###
Target labeled example:
{{ TASK_OFFLINE_INPUT(target) }}
Label: {{ target.label }}

{{ TASK_OWN_LABEL_REASON_OBJECTIVE }}

Reason:
```

The dataset-specific task text, offline field renderer, and own-label objective are the exact fragments listed in Appendix B.3.

Full per-label meta-reference bank. X-ICL and X^2-ICL explanation generation share a meta-reference bank containing exactly one meta reference for every canonical label, ordered by canonical label ID. Each meta reference carries the actual compatible own-label explanation for that specific example. The bank is not truncated to eight examples or to a query-specific local label set.

X-ICL observed-label explanation. The target identity is the demonstration ID, and only the demonstration's observed canonical label is used as the target label.

```
{{ TASK_FULL_META_EXPERT_INSTRUCTION }}

{% for meta in meta_examples %}
###
Meta Example {{ loop.index }}:
{{ TASK_OFFLINE_INPUT(meta) }}
Label: {{ meta.label }}
Reason: {{ meta.reason }}
{% endfor %}
###
Target example:
{{ TASK_OFFLINE_INPUT(target) }}
Label: {{ target.label }}
Reason:
```

No formal test-query text or gold label enters this offline prompt.

X^2-ICL label-conditioned explanation. The target identity is demonstration ID $\times$ canonical label ID. For a fixed target demonstration, the candidate label iterates over every label in $\mathcal{Y}$ in canonical label-ID order.

```
{{ TASK_FULL_META_EXPERT_INSTRUCTION }}

{% for meta in meta_examples %}
###
Meta Example {{ loop.index }}:
{{ TASK_OFFLINE_INPUT(meta) }}
Label: {{ meta.label }}
Reason: {{ meta.reason }}
{% endfor %}
###
Target example:
{{ TASK_OFFLINE_INPUT(target) }}
Label: {{ candidate_label }}
Reason:
```

The X^2-ICL renderer does not add hypothetical wording such as "if the label were" or reduce generation to the observed-label path.

B.5 Requested Output Structures and Validity Interface

The exact prompts request the following output structures.

ICL and Retrieval ICL.

```
Label: <canonical label>
```

X-ICL.

```
Reason: <one concise reason>
Label: <canonical label>
```

X^2-ICL.

```
Possible Reasoning for <canonical label 1>: <reasoning>
...
Possible Reasoning for <canonical label L>: <reasoning>
Label: <canonical label>
```

SG-XICL.

```
Possible Reasoning for <candidate label 1>: <one concise sentence>
...
Possible Reasoning for <candidate label m>: <one concise sentence>
Label: <canonical label>
```

These blocks specify the requested output structure and do not redefine evaluation validity. Classification validity follows the parsing and validity rules in Appendix A.6. Reasoning-format deviations are diagnostic unless they prevent recovery of one unique canonical final label.

C COMPLETE MAIN RESULTS

This section reports the complete cell-level Main results underlying the aggregate comparisons in the main paper. Point estimates use the frozen Main-500 evaluation scope. Randomized baselines are summarized over active seeds 1000 and 2000, whereas SG-XICL is a canonical single-run condition. Context-unsupported X^2-ICL cells are reported as `N/A_ctx` and are not assigned predictive scores.

C.1 COMPLETE MAIN PREDICTIVE RESULTS

The following six panels report all 48 model×dataset Main cells. Each entry is Accuracy / Macro-F1 in percent. Randomized methods show their active two-seed means. SG-XICL shows its canonical single-run result.

Table 8: Complete Main point estimates for `DeepSeek-v4-pro`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	70.80 / 69.92	79.70 / 79.61	68.40 / 67.06	68.20 / 66.65	81.80 / 81.85
Wiki80	75.90 / 75.42	80.40 / 79.81	74.40 / 74.14	75.30 / 74.83	84.00 / 83.79
CLINC150	95.50 / 94.99	99.10 / 99.02	96.50 / 96.01	95.60 / 95.05	99.00 / 98.98
Banking77	82.20 / 81.43	91.10 / 90.91	82.60 / 81.93	82.70 / 81.89	93.60 / 93.19
KBP37	50.00 / 52.01	52.40 / 54.38	48.70 / 50.83	50.80 / 51.35	50.40 / 54.89
SemEval-2010 Task 8	82.30 / 83.40	87.40 / 88.30	82.70 / 83.13	83.00 / 82.90	89.20 / 89.67
MAVEN-ETC	55.90 / 53.74	63.00 / 61.46	55.00 / 53.22	57.10 / 55.18	66.80 / 66.06
DialogRE-SR	79.60 / 73.37	83.00 / 80.24	80.90 / 75.39	79.70 / 74.43	86.40 / 80.27

Table 9: Complete Main point estimates for `DeepSeek-v4-flash`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	67.30 / 68.12	75.20 / 77.18	70.40 / 69.06	71.70 / 70.62	82.60 / 82.36
Wiki80	74.60 / 75.66	78.80 / 79.93	75.50 / 74.99	78.50 / 78.03	84.80 / 84.70
CLINC150	91.20 / 92.84	95.90 / 97.10	96.00 / 95.62	96.80 / 96.43	98.80 / 98.81
Banking77	78.20 / 79.43	81.00 / 84.89	82.60 / 81.78	83.10 / 82.11	95.20 / 95.08
KBP37	33.90 / 38.54	40.40 / 46.69	42.00 / 44.39	44.10 / 46.39	48.40 / 54.33
SemEval-2010 Task 8	63.80 / 72.12	66.00 / 74.45	76.70 / 78.41	82.00 / 83.14	85.40 / 86.22
MAVEN-ETC	52.70 / 50.90	58.70 / 57.65	52.60 / 50.34	54.90 / 52.57	64.20 / 63.00
DialogRE-SR	81.60 / 76.48	82.40 / 79.88	81.70 / 76.72	82.00 / 76.09	86.20 / 82.18

Table 10: Complete Main point estimates for `Qwen3-14B`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	38.10 / 35.86	59.10 / 58.61	40.10 / 37.68	`N/A_ctx`	56.60 / 56.99
Wiki80	42.10 / 40.49	61.90 / 62.17	44.70 / 43.79	37.70 / 37.13	59.00 / 58.90
CLINC150	86.90 / 86.38	96.80 / 96.64	87.60 / 86.35	`N/A_ctx`	97.20 / 97.18
Banking77	75.30 / 73.51	88.30 / 88.01	75.10 / 73.98	72.60 / 71.04	89.00 / 88.47
KBP37	31.90 / 30.54	43.10 / 42.48	33.80 / 32.69	20.70 / 20.52	30.40 / 33.32
SemEval-2010 Task 8	41.90 / 38.82	59.50 / 59.38	48.80 / 48.02	44.30 / 41.46	53.20 / 53.77
MAVEN-ETC	39.70 / 38.66	57.70 / 57.02	36.70 / 35.30	`N/A_ctx`	61.40 / 59.99
DialogRE-SR	38.70 / 35.51	42.60 / 42.34	42.10 / 40.43	38.70 / 36.18	44.40 / 41.02

Table 11: Complete Main point estimates for `Qwen3-8B`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	30.10 / 28.27	51.10 / 52.24	32.00 / 30.96	N/A_ctx	55.40 / 54.69
Wiki80	40.20 / 40.29	58.90 / 58.56	42.00 / 41.93	31.50 / 32.90	55.00 / 56.04
CLINC150	83.60 / 82.95	96.80 / 96.58	85.00 / 84.02	N/A_ctx	93.60 / 93.28
Banking77	68.90 / 67.80	87.70 / 87.37	68.70 / 66.65	68.80 / 67.99	83.60 / 82.67
KBP37	23.70 / 22.29	31.30 / 32.72	27.80 / 27.17	18.70 / 19.37	30.40 / 31.76
SemEval-2010 Task 8	28.10 / 24.62	40.60 / 38.36	35.60 / 30.97	33.60 / 30.24	45.20 / 45.40
MAVEN-ETC	32.40 / 29.47	51.70 / 51.16	28.90 / 25.85	N/A_ctx	56.60 / 54.86
DialogRE-SR	32.50 / 32.83	34.30 / 36.58	33.90 / 31.45	31.60 / 32.40	37.20 / 35.19

Table 12: Complete Main point estimates for `Mistral-Nemo-Instruct-2407`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	31.20 / 29.25	53.50 / 54.31	33.60 / 32.02	0.00 / 0.00	43.00 / 41.78
Wiki80	41.20 / 40.62	59.00 / 59.71	41.00 / 40.87	3.20 / 4.53	46.60 / 46.91
CLINC150	83.30 / 81.54	96.70 / 96.70	82.80 / 81.11	3.40 / 4.83	94.00 / 93.75
Banking77	69.20 / 68.25	88.30 / 88.19	68.80 / 67.76	1.10 / 1.76	82.60 / 82.01
KBP37	28.20 / 24.08	34.60 / 32.58	33.90 / 32.04	5.60 / 6.75	28.60 / 27.39
SemEval-2010 Task 8	33.80 / 27.35	44.00 / 41.91	36.80 / 32.20	24.80 / 21.34	43.40 / 40.42
MAVEN-ETC	34.20 / 33.17	56.70 / 56.59	29.40 / 27.14	0.00 / 0.00	42.60 / 41.73
DialogRE-SR	41.30 / 39.41	42.50 / 42.80	36.60 / 34.35	18.50 / 20.67	30.80 / 29.63

Table 13: Complete Main point estimates for `Mistral-7B-Instruct-v0.3`. Entries are Accuracy / Macro-F1 (%).

Dataset	ICL	Retrieval ICL	X-ICL	X^2-ICL	SG-XICL
FewRel	15.70 / 15.96	41.50 / 44.57	22.10 / 22.17	N/A_ctx	31.40 / 31.42
Wiki80	24.10 / 23.66	39.40 / 42.69	31.90 / 32.50	N/A_ctx	34.00 / 37.01
CLINC150	38.60 / 42.83	9.50 / 12.75	77.00 / 75.78	N/A_ctx	79.80 / 81.02
Banking77	23.50 / 28.66	16.20 / 20.39	63.70 / 63.95	N/A_ctx	62.00 / 63.45
KBP37	9.00 / 9.47	13.90 / 15.99	18.50 / 16.67	12.10 / 11.62	13.40 / 13.52
SemEval-2010 Task 8	21.20 / 19.39	29.20 / 31.22	35.90 / 30.09	21.00 / 20.85	33.60 / 31.30
MAVEN-ETC	17.10 / 18.14	40.30 / 42.99	27.50 / 26.68	N/A_ctx	26.00 / 25.75
DialogRE-SR	1.20 / 2.49	1.10 / 1.82	31.50 / 31.69	22.80 / 21.78	21.80 / 24.40

C.2 Aggregate Main Comparisons

The aggregate comparisons in the main paper are computed from the cell-level Main results above. Let c index a model×dataset cell and let M denote either Accuracy or Macro-F1, measured in percentage points. For a randomized baseline B, its cell-level metric is first averaged over the two active seeds,

$$\overline{M}_B(c) = \frac{M_{B,1000}(c) + M_{B,2000}(c)}{2}.$$

SG-XICL contributes its single canonical value $M_{\mathrm{SG}}(c)$. The cell-level difference is then

$$\Delta_B^M(c) = M_{\mathrm{SG}}(c) - \overline{M}_B(c).$$

For each baseline, the reported aggregate is the unweighted mean over its common-supported model×dataset cells,

$$\overline{\Delta}_B^M = \frac{1}{|\mathcal{C}_B|} \sum_{c \in \mathcal{C}_B} \Delta_B^M(c),$$

where $\mathcal{C}_B$ is the baseline-specific set of cells for which both SG-XICL and the comparator are context-supported under the frozen Main runtime. This is a cell-balanced aggregation rather than a pooled-query metric. The two-seed sample standard deviations reported below describe run-to-run variability and are not used as aggregation weights.

Table 14: Aggregate SG-XICL differences relative to each Main baseline. Differences are SG-XICL minus baseline in percentage points. Aggregation is an unweighted mean over baseline-specific common-supported model×dataset cells after averaging the two active baseline seeds.

Baseline	Common-supported cells	Accuracy Δ	Macro-F1 Δ
ICL	48	+12.00	+12.49
Retrieval ICL	48	+2.42	+1.53
X-ICL	48	+7.46	+8.61
X^2-ICL	37	+16.44	+16.86

The common-support definition here concerns method-level context feasibility and is distinct from the query-level gold-label-evidence eligibility used in the conditional analysis. Cells marked `N/A_ctx` are excluded from both the numerator and denominator of the corresponding aggregate and are never converted to zero. By contrast, a context-supported but parser-invalid prediction remains in the formal Accuracy denominator and counts as incorrect.

C.3 Active-Seed Variability

For each supported randomized condition, the tables below report seed 1000, seed 2000, the active-seed mean, and the sample standard deviation. Context-unsupported X^2-ICL cells remain marked as `N/A_ctx`. Each supported metric cell has the form `seed1000 / seed2000; mean ± SD`, in percent. With only two active seeds, the SD is descriptive run-to-run variability rather than a precise estimate of a broader randomness distribution. SG-XICL is omitted because it has one canonical run rather than artificial seed replicates.

`DeepSeek-v4-pro`.

Table 15: Active-seed variability for `DeepSeek-v4-pro`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	72.00 / 69.60; 70.80 ± 1.70	71.03 / 68.80; 69.92 ± 1.57
FewRel	Retrieval ICL	78.60 / 80.80; 79.70 ± 1.56	78.56 / 80.66; 79.61 ± 1.48
FewRel	X-ICL	69.00 / 67.80; 68.40 ± 0.85	67.20 / 66.93; 67.06 ± 0.19
FewRel	X^2-ICL	67.40 / 69.00; 68.20 ± 1.13	65.65 / 67.66; 66.65 ± 1.43
Wiki80	ICL	76.00 / 75.80; 75.90 ± 0.14	75.53 / 75.30; 75.42 ± 0.16
Wiki80	Retrieval ICL	80.80 / 80.00; 80.40 ± 0.57	80.21 / 79.40; 79.81 ± 0.58
Wiki80	X-ICL	73.80 / 75.00; 74.40 ± 0.85	73.49 / 74.80; 74.14 ± 0.93
Wiki80	X^2-ICL	75.20 / 75.40; 75.30 ± 0.14	74.77 / 74.90; 74.83 ± 0.09
CLINC150	ICL	95.60 / 95.40; 95.50 ± 0.14	95.08 / 94.89; 94.99 ± 0.14
CLINC150	Retrieval ICL	99.00 / 99.20; 99.10 ± 0.14	98.90 / 99.14; 99.02 ± 0.17
CLINC150	X-ICL	96.80 / 96.20; 96.50 ± 0.42	96.42 / 95.60; 96.01 ± 0.58
CLINC150	X^2-ICL	95.40 / 95.80; 95.60 ± 0.28	94.75 / 95.36; 95.05 ± 0.43
Banking77	ICL	81.60 / 82.80; 82.20 ± 0.85	81.05 / 81.82; 81.43 ± 0.54
Banking77	Retrieval ICL	91.20 / 91.00; 91.10 ± 0.14	90.98 / 90.83; 90.91 ± 0.11
Banking77	X-ICL	83.40 / 81.80; 82.60 ± 1.13	82.60 / 81.26; 81.93 ± 0.95
Banking77	X^2-ICL	83.00 / 82.40; 82.70 ± 0.42	82.00 / 81.78; 81.89 ± 0.16
KBP37	ICL	50.40 / 49.60; 50.00 ± 0.57	52.44 / 51.58; 52.01 ± 0.60
KBP37	Retrieval ICL	52.40 / 52.40; 52.40 ± 0.00	54.77 / 53.99; 54.38 ± 0.55
KBP37	X-ICL	50.20 / 47.20; 48.70 ± 2.12	51.81 / 49.84; 50.83 ± 1.39
KBP37	X^2-ICL	51.00 / 50.60; 50.80 ± 0.28	51.07 / 51.62; 51.35 ± 0.39
SemEval-2010 Task 8	ICL	82.80 / 81.80; 82.30 ± 0.71	84.08 / 82.73; 83.40 ± 0.95
SemEval-2010 Task 8	Retrieval ICL	88.00 / 86.80; 87.40 ± 0.85	88.72 / 87.87; 88.30 ± 0.60
SemEval-2010 Task 8	X-ICL	84.20 / 81.20; 82.70 ± 2.12	84.75 / 81.51; 83.13 ± 2.29
SemEval-2010 Task 8	X^2-ICL	84.00 / 82.00; 83.00 ± 1.41	84.14 / 81.66; 82.90 ± 1.75

Dataset	Method	Accuracy	Macro-F1
MAVEN-ETC	ICL	56.20 / 55.60; 55.90 ± 0.42	53.73 / 53.75; 53.74 ± 0.02
MAVEN-ETC	Retrieval ICL	64.00 / 62.00; 63.00 ± 1.41	62.87 / 60.06; 61.46 ± 1.99
MAVEN-ETC	X-ICL	54.80 / 55.20; 55.00 ± 0.28	53.20 / 53.25; 53.22 ± 0.04
MAVEN-ETC	X^2-ICL	57.40 / 56.80; 57.10 ± 0.42	55.70 / 54.67; 55.18 ± 0.73
DialogRE-SR	ICL	79.60 / 79.60; 79.60 ± 0.00	73.30 / 73.44; 73.37 ± 0.10
DialogRE-SR	Retrieval ICL	83.20 / 82.80; 83.00 ± 0.28	80.48 / 79.99; 80.24 ± 0.34
DialogRE-SR	X-ICL	81.60 / 80.20; 80.90 ± 0.99	76.17 / 74.61; 75.39 ± 1.11
DialogRE-SR	X^2-ICL	81.00 / 78.40; 79.70 ± 1.84	75.23 / 73.63; 74.43 ± 1.13

`DeepSeek-v4-flash`.

Table 16: Active-seed variability for `DeepSeek-v4-flash`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	67.40 / 67.20; 67.30 ± 0.14	68.18 / 68.06; 68.12 ± 0.09
FewRel	Retrieval ICL	73.80 / 76.60; 75.20 ± 1.98	75.90 / 78.46; 77.18 ± 1.81
FewRel	X-ICL	71.60 / 69.20; 70.40 ± 1.70	70.42 / 67.71; 69.06 ± 1.91
FewRel	X^2-ICL	72.80 / 70.60; 71.70 ± 1.56	71.42 / 69.82; 70.62 ± 1.13
Wiki80	ICL	74.00 / 75.20; 74.60 ± 0.85	74.83 / 76.50; 75.66 ± 1.18
Wiki80	Retrieval ICL	78.80 / 78.80; 78.80 ± 0.00	79.88 / 79.98; 79.93 ± 0.07
Wiki80	X-ICL	75.40 / 75.60; 75.50 ± 0.14	74.82 / 75.16; 74.99 ± 0.24
Wiki80	X^2-ICL	78.40 / 78.60; 78.50 ± 0.14	77.88 / 78.19; 78.03 ± 0.22
CLINC150	ICL	93.40 / 89.00; 91.20 ± 3.11	94.43 / 91.25; 92.84 ± 2.25
CLINC150	Retrieval ICL	95.40 / 96.40; 95.90 ± 0.71	96.54 / 97.66; 97.10 ± 0.79
CLINC150	X-ICL	96.40 / 95.60; 96.00 ± 0.57	95.97 / 95.28; 95.62 ± 0.49
CLINC150	X^2-ICL	97.00 / 96.60; 96.80 ± 0.28	96.70 / 96.16; 96.43 ± 0.38
Banking77	ICL	78.80 / 77.60; 78.20 ± 0.85	80.05 / 78.82; 79.43 ± 0.87
Banking77	Retrieval ICL	81.20 / 80.80; 81.00 ± 0.28	84.77 / 85.02; 84.89 ± 0.18
Banking77	X-ICL	82.60 / 82.60; 82.60 ± 0.00	81.80 / 81.76; 81.78 ± 0.03
Banking77	X^2-ICL	83.00 / 83.20; 83.10 ± 0.14	81.95 / 82.27; 82.11 ± 0.23
KBP37	ICL	35.40 / 32.40; 33.90 ± 2.12	39.93 / 37.15; 38.54 ± 1.96
KBP37	Retrieval ICL	40.00 / 40.80; 40.40 ± 0.57	46.75 / 46.63; 46.69 ± 0.09
KBP37	X-ICL	44.40 / 39.60; 42.00 ± 3.39	46.25 / 42.53; 44.39 ± 2.63
KBP37	X^2-ICL	44.40 / 43.80; 44.10 ± 0.42	46.10 / 46.68; 46.39 ± 0.41
SemEval-2010 Task 8	ICL	61.80 / 65.80; 63.80 ± 2.83	70.69 / 73.55; 72.12 ± 2.02
SemEval-2010 Task 8	Retrieval ICL	67.20 / 64.80; 66.00 ± 1.70	75.75 / 73.15; 74.45 ± 1.84
SemEval-2010 Task 8	X-ICL	76.20 / 77.20; 76.70 ± 0.71	78.01 / 78.82; 78.41 ± 0.57
SemEval-2010 Task 8	X^2-ICL	82.00 / 82.00; 82.00 ± 0.00	83.14 / 83.15; 83.14 ± 0.01
MAVEN-ETC	ICL	53.20 / 52.20; 52.70 ± 0.71	51.42 / 50.38; 50.90 ± 0.74
MAVEN-ETC	Retrieval ICL	61.20 / 56.20; 58.70 ± 3.54	60.22 / 55.09; 57.65 ± 3.62
MAVEN-ETC	X-ICL	52.60 / 52.60; 52.60 ± 0.00	50.20 / 50.48; 50.34 ± 0.20
MAVEN-ETC	X^2-ICL	55.20 / 54.60; 54.90 ± 0.42	53.04 / 52.10; 52.57 ± 0.67
DialogRE-SR	ICL	81.80 / 81.40; 81.60 ± 0.28	77.42 / 75.54; 76.48 ± 1.33
DialogRE-SR	Retrieval ICL	81.60 / 83.20; 82.40 ± 1.13	79.57 / 80.19; 79.88 ± 0.44
DialogRE-SR	X-ICL	80.40 / 83.00; 81.70 ± 1.84	76.68 / 76.77; 76.72 ± 0.06
DialogRE-SR	X^2-ICL	82.40 / 81.60; 82.00 ± 0.57	78.23 / 73.95; 76.09 ± 3.02

`Qwen3-14B`.

Table 17: Active-seed variability for `Qwen3-14B`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	37.60 / 38.60; 38.10 ± 0.71	35.28 / 36.44; 35.86 ± 0.82
FewRel	Retrieval ICL	58.00 / 60.20; 59.10 ± 1.56	57.47 / 59.75; 58.61 ± 1.61
FewRel	X-ICL	43.00 / 37.20; 40.10 ± 4.10	40.14 / 35.22; 37.68 ± 3.48
FewRel	X^2-ICL	`N/A_ctx`	`N/A_ctx`
Wiki80	ICL	41.40 / 42.80; 42.10 ± 0.99	39.77 / 41.22; 40.49 ± 1.03
Wiki80	Retrieval ICL	62.00 / 61.80; 61.90 ± 0.14	62.37 / 61.98; 62.17 ± 0.27
Wiki80	X-ICL	44.40 / 45.00; 44.70 ± 0.42	43.89 / 43.69; 43.79 ± 0.14
Wiki80	X^2-ICL	37.40 / 38.00; 37.70 ± 0.42	37.68 / 36.58; 37.13 ± 0.78
CLINC150	ICL	86.60 / 87.20; 86.90 ± 0.42	86.07 / 86.69; 86.38 ± 0.44
CLINC150	Retrieval ICL	96.80 / 96.80; 96.80 ± 0.00	96.61 / 96.67; 96.64 ± 0.04
CLINC150	X-ICL	88.60 / 86.60; 87.60 ± 1.41	87.33 / 85.37; 86.35 ± 1.39
CLINC150	X^2-ICL	`N/A_ctx`	`N/A_ctx`
Banking77	ICL	75.20 / 75.40; 75.30 ± 0.14	73.24 / 73.78; 73.51 ± 0.38
Banking77	Retrieval ICL	88.80 / 87.80; 88.30 ± 0.71	88.44 / 87.58; 88.01 ± 0.61
Banking77	X-ICL	74.80 / 75.40; 75.10 ± 0.42	73.51 / 74.44; 73.98 ± 0.66
Banking77	X^2-ICL	74.00 / 71.20; 72.60 ± 1.98	72.19 / 69.89; 71.04 ± 1.63
KBP37	ICL	33.80 / 30.00; 31.90 ± 2.69	30.78 / 30.29; 30.54 ± 0.34
KBP37	Retrieval ICL	41.40 / 44.80; 43.10 ± 2.40	41.47 / 43.50; 42.48 ± 1.43
KBP37	X-ICL	35.00 / 32.60; 33.80 ± 1.70	32.67 / 32.71; 32.69 ± 0.03
KBP37	X^2-ICL	22.80 / 18.60; 20.70 ± 2.97	22.05 / 19.00; 20.52 ± 2.16
SemEval-2010 Task 8	ICL	40.40 / 43.40; 41.90 ± 2.12	36.97 / 40.68; 38.82 ± 2.62
SemEval-2010 Task 8	Retrieval ICL	60.00 / 59.00; 59.50 ± 0.71	60.18 / 58.59; 59.38 ± 1.12
SemEval-2010 Task 8	X-ICL	48.20 / 49.40; 48.80 ± 0.85	48.28 / 47.76; 48.02 ± 0.37
SemEval-2010 Task 8	X^2-ICL	41.40 / 47.20; 44.30 ± 4.10	38.42 / 44.51; 41.46 ± 4.31
MAVEN-ETC	ICL	39.80 / 39.60; 39.70 ± 0.14	38.66 / 38.65; 38.66 ± 0.01
MAVEN-ETC	Retrieval ICL	58.00 / 57.40; 57.70 ± 0.42	56.90 / 57.15; 57.02 ± 0.18
MAVEN-ETC	X-ICL	38.00 / 35.40; 36.70 ± 1.84	36.08 / 34.52; 35.30 ± 1.10
MAVEN-ETC	X^2-ICL	`N/A_ctx`	`N/A_ctx`
DialogRE-SR	ICL	38.60 / 38.80; 38.70 ± 0.14	35.92 / 35.10; 35.51 ± 0.58
DialogRE-SR	Retrieval ICL	42.00 / 43.20; 42.60 ± 0.85	41.63 / 43.06; 42.34 ± 1.01
DialogRE-SR	X-ICL	44.20 / 40.00; 42.10 ± 2.97	40.37 / 40.49; 40.43 ± 0.09
DialogRE-SR	X^2-ICL	37.40 / 40.00; 38.70 ± 1.84	35.23 / 37.13; 36.18 ± 1.34

`Qwen3-8B`.

Table 18: Active-seed variability for `Qwen3-8B`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	34.20 / 26.00; 30.10 ± 5.80	31.63 / 24.92; 28.27 ± 4.74
FewRel	Retrieval ICL	51.00 / 51.20; 51.10 ± 0.14	52.45 / 52.03; 52.24 ± 0.30
FewRel	X-ICL	35.80 / 28.20; 32.00 ± 5.37	33.96 / 27.95; 30.96 ± 4.26
FewRel	X^2-ICL	`N/A_ctx`	`N/A_ctx`
Wiki80	ICL	40.20 / 40.20; 40.20 ± 0.00	40.86 / 39.71; 40.29 ± 0.82
Wiki80	Retrieval ICL	59.60 / 58.20; 58.90 ± 0.99	58.63 / 58.48; 58.56 ± 0.11
Wiki80	X-ICL	43.20 / 40.80; 42.00 ± 1.70	43.08 / 40.78; 41.93 ± 1.63
Wiki80	X^2-ICL	28.00 / 35.00; 31.50 ± 4.95	30.23 / 35.56; 32.90 ± 3.77
CLINC150	ICL	84.20 / 83.00; 83.60 ± 0.85	83.41 / 82.48; 82.95 ± 0.65
CLINC150	Retrieval ICL	96.20 / 97.40; 96.80 ± 0.85	95.90 / 97.25; 96.58 ± 0.95
CLINC150	X-ICL	85.00 / 85.00; 85.00 ± 0.00	83.88 / 84.16; 84.02 ± 0.19
CLINC150	X^2-ICL	`N/A_ctx`	`N/A_ctx`

Dataset	Method	Accuracy	Macro-F1
Banking77	ICL	69.00 / 68.80; 68.90 ± 0.14	68.10 / 67.50; 67.80 ± 0.43
Banking77	Retrieval ICL	87.80 / 87.60; 87.70 ± 0.14	87.34 / 87.40; 87.37 ± 0.05
Banking77	X-ICL	69.60 / 67.80; 68.70 ± 1.27	67.36 / 65.93; 66.65 ± 1.01
Banking77	X^2-ICL	69.20 / 68.40; 68.80 ± 0.57	68.32 / 67.66; 67.99 ± 0.46
KBP37	ICL	28.20 / 19.20; 23.70 ± 6.36	25.55 / 19.03; 22.29 ± 4.61
KBP37	Retrieval ICL	31.60 / 31.00; 31.30 ± 0.42	33.40 / 32.04; 32.72 ± 0.97
KBP37	X-ICL	32.80 / 22.80; 27.80 ± 7.07	32.33 / 22.02; 27.17 ± 7.29
KBP37	X^2-ICL	22.00 / 15.40; 18.70 ± 4.67	21.60 / 17.13; 19.37 ± 3.17
SemEval-2010 Task 8	ICL	30.60 / 25.60; 28.10 ± 3.54	26.51 / 22.74; 24.62 ± 2.67
SemEval-2010 Task 8	Retrieval ICL	41.00 / 40.20; 40.60 ± 0.57	38.86 / 37.86; 38.36 ± 0.71
SemEval-2010 Task 8	X-ICL	35.40 / 35.80; 35.60 ± 0.28	31.18 / 30.75; 30.97 ± 0.30
SemEval-2010 Task 8	X^2-ICL	31.60 / 35.60; 33.60 ± 2.83	28.69 / 31.79; 30.24 ± 2.19
MAVEN-ETC	ICL	32.40 / 32.40; 32.40 ± 0.00	28.70 / 30.24; 29.47 ± 1.09
MAVEN-ETC	Retrieval ICL	51.80 / 51.60; 51.70 ± 0.14	51.08 / 51.23; 51.16 ± 0.11
MAVEN-ETC	X-ICL	30.60 / 27.20; 28.90 ± 2.40	27.62 / 24.08; 25.85 ± 2.50
MAVEN-ETC	X^2-ICL	N/A_ctx	N/A_ctx
DialogRE-SR	ICL	33.80 / 31.20; 32.50 ± 1.84	33.65 / 32.00; 32.83 ± 1.17
DialogRE-SR	Retrieval ICL	33.60 / 35.00; 34.30 ± 0.99	36.66 / 36.50; 36.58 ± 0.12
DialogRE-SR	X-ICL	33.40 / 34.40; 33.90 ± 0.71	31.59 / 31.31; 31.45 ± 0.19
DialogRE-SR	X^2-ICL	32.40 / 30.80; 31.60 ± 1.13	33.58 / 31.22; 32.40 ± 1.67

`Mistral-Nemo-Instruct-2407`.

Table 19: Active-seed variability for `Mistral-Nemo-Instruct-2407`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	31.80 / 30.60; 31.20 ± 0.85	30.28 / 28.21; 29.25 ± 1.46
FewRel	Retrieval ICL	54.20 / 52.80; 53.50 ± 0.99	55.19 / 53.44; 54.31 ± 1.24
FewRel	X-ICL	36.40 / 30.80; 33.60 ± 3.96	34.74 / 29.31; 32.02 ± 3.84
FewRel	X^2-ICL	0.00 / 0.00; 0.00 ± 0.00	0.00 / 0.00; 0.00 ± 0.00
Wiki80	ICL	41.60 / 40.80; 41.20 ± 0.57	40.92 / 40.33; 40.62 ± 0.42
Wiki80	Retrieval ICL	59.40 / 58.60; 59.00 ± 0.57	59.99 / 59.42; 59.71 ± 0.41
Wiki80	X-ICL	43.40 / 38.60; 41.00 ± 3.39	43.58 / 38.16; 40.87 ± 3.83
Wiki80	X^2-ICL	2.40 / 4.00; 3.20 ± 1.13	2.98 / 6.08; 4.53 ± 2.19
CLINC150	ICL	83.40 / 83.20; 83.30 ± 0.14	81.75 / 81.32; 81.54 ± 0.30
CLINC150	Retrieval ICL	97.40 / 96.00; 96.70 ± 0.99	97.44 / 95.96; 96.70 ± 1.05
CLINC150	X-ICL	83.80 / 81.80; 82.80 ± 1.41	81.74 / 80.47; 81.11 ± 0.90
CLINC150	X^2-ICL	0.00 / 6.80; 3.40 ± 4.81	0.00 / 9.65; 4.83 ± 6.82
Banking77	ICL	68.80 / 69.60; 69.20 ± 0.57	68.09 / 68.41; 68.25 ± 0.22
Banking77	Retrieval ICL	88.20 / 88.40; 88.30 ± 0.14	87.88 / 88.50; 88.19 ± 0.44
Banking77	X-ICL	68.20 / 69.40; 68.80 ± 0.85	66.86 / 68.67; 67.76 ± 1.28
Banking77	X^2-ICL	1.00 / 1.20; 1.10 ± 0.14	1.68 / 1.85; 1.76 ± 0.12
KBP37	ICL	29.40 / 27.00; 28.20 ± 1.70	22.54 / 25.62; 24.08 ± 2.18
KBP37	Retrieval ICL	34.80 / 34.40; 34.60 ± 0.28	32.15 / 33.01; 32.58 ± 0.61
KBP37	X-ICL	35.60 / 32.20; 33.90 ± 2.40	32.71 / 31.38; 32.04 ± 0.94
KBP37	X^2-ICL	5.00 / 6.20; 5.60 ± 0.85	6.30 / 7.20; 6.75 ± 0.64
SemEval-2010 Task 8	ICL	35.00 / 32.60; 33.80 ± 1.70	29.29 / 25.40; 27.35 ± 2.75
SemEval-2010 Task 8	Retrieval ICL	44.00 / 44.00; 44.00 ± 0.00	42.26 / 41.55; 41.91 ± 0.50
SemEval-2010 Task 8	X-ICL	36.40 / 37.20; 36.80 ± 0.57	30.94 / 33.46; 32.20 ± 1.78
SemEval-2010 Task 8	X^2-ICL	25.20 / 24.40; 24.80 ± 0.57	22.49 / 20.19; 21.34 ± 1.63
MAVEN-ETC	ICL	32.80 / 35.60; 34.20 ± 1.98	31.16 / 35.17; 33.17 ± 2.84
MAVEN-ETC	Retrieval ICL	56.20 / 57.20; 56.70 ± 0.71	55.49 / 57.68; 56.59 ± 1.54

Dataset	Method	Accuracy	Macro-F1
MAVEN-ETC	X-ICL	31.00 / 27.80; 29.40 ± 2.26	28.93 / 25.36; 27.14 ± 2.52
MAVEN-ETC	X^2-ICL	0.00 / 0.00; 0.00 ± 0.00	0.00 / 0.00; 0.00 ± 0.00
DialogRE-SR	ICL	43.00 / 39.60; 41.30 ± 2.40	41.07 / 37.75; 39.41 ± 2.35
DialogRE-SR	Retrieval ICL	43.20 / 41.80; 42.50 ± 0.99	42.82 / 42.79; 42.80 ± 0.02
DialogRE-SR	X-ICL	38.40 / 34.80; 36.60 ± 2.55	36.46 / 32.24; 34.35 ± 2.98
DialogRE-SR	X^2-ICL	17.40 / 19.60; 18.50 ± 1.56	20.74 / 20.60; 20.67 ± 0.09

`Mistral-7B-Instruct-v0.3`.

Table 20: Active-seed variability for `Mistral-7B-Instruct-v0.3`.

Dataset	Method	Accuracy	Macro-F1
FewRel	ICL	12.20 / 19.20; 15.70 ± 4.95	12.88 / 19.05; 15.96 ± 4.37
FewRel	Retrieval ICL	41.80 / 41.20; 41.50 ± 0.42	45.09 / 44.05; 44.57 ± 0.73
FewRel	X-ICL	23.00 / 21.20; 22.10 ± 1.27	23.81 / 20.52; 22.17 ± 2.32
FewRel	X^2-ICL	`N/A_ctx`	`N/A_ctx`
Wiki80	ICL	23.20 / 25.00; 24.10 ± 1.27	23.64 / 23.68; 23.66 ± 0.03
Wiki80	Retrieval ICL	37.20 / 41.60; 39.40 ± 3.11	40.20 / 45.17; 42.69 ± 3.51
Wiki80	X-ICL	30.00 / 33.80; 31.90 ± 2.69	30.51 / 34.50; 32.50 ± 2.82
Wiki80	X^2-ICL	`N/A_ctx`	`N/A_ctx`
CLINC150	ICL	37.40 / 39.80; 38.60 ± 1.70	42.41 / 43.24; 42.83 ± 0.59
CLINC150	Retrieval ICL	8.40 / 10.60; 9.50 ± 1.56	11.24 / 14.27; 12.75 ± 2.14
CLINC150	X-ICL	74.00 / 80.00; 77.00 ± 4.24	72.36 / 79.19; 75.78 ± 4.83
CLINC150	X^2-ICL	`N/A_ctx`	`N/A_ctx`
Banking77	ICL	21.40 / 25.60; 23.50 ± 2.97	25.72 / 31.60; 28.66 ± 4.16
Banking77	Retrieval ICL	18.40 / 14.00; 16.20 ± 3.11	23.33 / 17.44; 20.39 ± 4.16
Banking77	X-ICL	62.80 / 64.60; 63.70 ± 1.27	63.14 / 64.76; 63.95 ± 1.15
Banking77	X^2-ICL	`N/A_ctx`	`N/A_ctx`
KBP37	ICL	8.40 / 9.60; 9.00 ± 0.85	8.70 / 10.25; 9.47 ± 1.10
KBP37	Retrieval ICL	14.00 / 13.80; 13.90 ± 0.14	16.88 / 15.11; 15.99 ± 1.26
KBP37	X-ICL	22.80 / 14.20; 18.50 ± 6.08	20.50 / 12.83; 16.67 ± 5.42
KBP37	X^2-ICL	14.00 / 10.20; 12.10 ± 2.69	13.40 / 9.83; 11.62 ± 2.53
SemEval-2010 Task 8	ICL	21.40 / 21.00; 21.20 ± 0.28	20.21 / 18.57; 19.39 ± 1.16
SemEval-2010 Task 8	Retrieval ICL	30.00 / 28.40; 29.20 ± 1.13	32.42 / 30.03; 31.22 ± 1.69
SemEval-2010 Task 8	X-ICL	33.80 / 38.00; 35.90 ± 2.97	28.39 / 31.79; 30.09 ± 2.40
SemEval-2010 Task 8	X^2-ICL	26.20 / 15.80; 21.00 ± 7.35	24.83 / 16.87; 20.85 ± 5.63
MAVEN-ETC	ICL	26.00 / 8.20; 17.10 ± 12.59	25.37 / 10.90; 18.14 ± 10.23
MAVEN-ETC	Retrieval ICL	40.40 / 40.20; 40.30 ± 0.14	43.47 / 42.50; 42.99 ± 0.69
MAVEN-ETC	X-ICL	29.60 / 25.40; 27.50 ± 2.97	28.43 / 24.92; 26.68 ± 2.48
MAVEN-ETC	X^2-ICL	`N/A_ctx`	`N/A_ctx`
DialogRE-SR	ICL	1.20 / 1.20; 1.20 ± 0.00	2.94 / 2.04; 2.49 ± 0.63
DialogRE-SR	Retrieval ICL	1.20 / 1.00; 1.10 ± 0.14	1.73 / 1.91; 1.82 ± 0.13
DialogRE-SR	X-ICL	30.20 / 32.80; 31.50 ± 1.84	31.15 / 32.24; 31.69 ± 0.77
DialogRE-SR	X^2-ICL	20.40 / 25.20; 22.80 ± 3.39	20.01 / 23.55; 21.78 ± 2.50

D Evidence, Scope, and Token Details

This section provides the full numerical detail underlying the main paper's analyses of evidence access, conditional predictive behavior, evaluated context feasibility, explicit-processing breadth, and online token burden.

D.1 GOLD-LABEL EVIDENCE COVERAGE

Table 21 reports how often Retrieval ICL and canonical SG-XICL expose at least one demonstration or reference from the query's gold label. Retrieval coverage is seed-specific because the retrieval pool is randomized by active seed. The pooled Retrieval columns therefore use 500 frozen queries $\times$ two active seeds $= 1000$ query-seed coordinates per dataset, whereas SG-XICL uses one canonical selection over 500 queries per dataset. The final two columns condition on query-seed coordinates for which Retrieval ICL misses the gold label. The canonical SG-XICL coverage status for a query is aligned to each corresponding Retrieval seed coordinate only for this conditional calculation; this does not create additional SG-XICL runs.

Table 21: Gold-label evidence coverage. Coverage columns are percentages. "Ret. miss N" is the pooled number of Retrieval-miss query-seed coordinates across seeds 1000 and 2000. The final column reports SG-XICL gold-label coverage on those Retrieval-miss coordinates as covered/miss (percentage).

Dataset	$\lvert\mathcal{Y}\rvert$	Ret. 1000	Ret. 2000	Ret. pooled	SG	Ret. miss N	SG on Ret. miss
SemEval-2010 Task 8	18	71.00	67.80	69.40	94.20	306	263/306 (85.95%)
DialogRE-SR	30	56.60	60.40	58.50	80.20	415	287/415 (69.16%)
KBP37	37	45.60	48.60	47.10	69.00	529	301/529 (56.90%)
Banking77	77	99.00	98.80	98.90	99.80	11	10/11 (90.91%)
Wiki80	80	71.80	71.20	71.50	94.00	285	234/285 (82.11%)
FewRel	124	68.80	68.40	68.60	90.80	314	243/314 (77.39%)
CLINC150	150	99.00	98.80	98.90	99.80	11	9/11 (81.82%)
MAVEN-ETC	165	62.80	62.00	62.40	82.40	376	248/376 (65.96%)

The conditional percentages for Banking77 and CLINC150 are based on only 11 pooled Retrieval-miss coordinates in each dataset, so their exact numerators and denominators are shown explicitly rather than interpreting the percentages in isolation. No custom coverage or recovery metric is introduced here.

D.2 ACCURACY UNDER GOLD-LABEL EVIDENCE COVERAGE

For each dataset and active seed, a query is eligible for this descriptive comparison only when both (i) the seed-specific fixed ICL-family demonstration set used by X-ICL and X^2-ICL contains at least one gold-label example and (ii) the canonical SG-XICL selected references contain at least one gold-label example. Eligibility is defined before prediction and does not depend on correctness, output validity, model identity, or explanation content. This restriction matches only binary gold-label evidence availability; it does not match demonstration quality, semantic similarity, explanation content, candidate identities, prompt length, ordering, or reasoning structure.

Table 22 reports the complete six-model, eight-dataset, two-seed Appendix projection. Each paired entry is shown in seed order 1000 / 2000. "Eligible N" is the formal Accuracy denominator for that dataset-seed subset. Supported invalid outputs remain in this denominator and count as incorrect. $\mathrm{N/A}_{\mathrm{ctx}}$ means the method-faithful X^2-ICL condition is context-unsupported under the frozen runtime and is not converted to zero.

SG-XICL itself has one canonical Main prediction vector. Its two displayed Accuracy values for a model-dataset pair evaluate that same prediction vector on the two seed-specific eligible subsets; they are not independent SG-XICL runs.

Table 22: Accuracy under gold-label evidence coverage. Paired entries are seed 1000 / seed 2000. Accuracy entries are percentages. The table preserves all six Main models and both active seeds; $\mathrm{N/A}_{\mathrm{ctx}}$ denotes a context-unsupported X^2-ICL condition rather than a zero score.

Model	Dataset	Eligible N	SG-XICL	X-ICL	X^2-ICL
DeepSeek-v4-pro	SemEval-2010 Task 8	166 / 113	87.35 / 87.61	80.12 / 84.07	76.51 / 81.42
DeepSeek-v4-pro	DialogRE-SR	89 / 63	96.63 / 92.06	92.13 / 90.48	93.26 / 88.89
DeepSeek-v4-pro	KBP37	78 / 65	62.82 / 55.38	73.08 / 44.62	67.95 / 47.69
DeepSeek-v4-pro	Banking77	51 / 56	92.16 / 98.21	82.35 / 91.07	82.35 / 91.07
DeepSeek-v4-pro	Wiki80	34 / 46	82.35 / 91.30	82.35 / 82.61	79.41 / 80.43
DeepSeek-v4-pro	FewRel	31 / 28	83.87 / 85.71	80.65 / 78.57	70.97 / 82.14
DeepSeek-v4-pro	CLINC150	28 / 26	100.00 / 100.00	100.00 / 100.00	100.00 / 100.00

Model	Dataset	Eligible N	SG-XICL	X-ICL	X^2-ICL
DeepSeek-v4-pro	MAVEN-ETC	25 / 24	68.00 / 70.83	56.00 / 45.83	56.00 / 45.83
DeepSeek-v4-flash	SemEval-2010 Task 8	166 / 113	84.34 / 87.61	78.92 / 85.84	83.73 / 84.07
DeepSeek-v4-flash	DialogRE-SR	89 / 63	88.76 / 87.30	87.64 / 90.48	89.89 / 90.48
DeepSeek-v4-flash	KBP37	78 / 65	62.82 / 49.23	66.67 / 38.46	65.38 / 38.46
DeepSeek-v4-flash	Banking77	51 / 56	96.08 / 96.43	82.35 / 87.50	80.39 / 87.50
DeepSeek-v4-flash	Wiki80	34 / 46	85.29 / 89.13	79.41 / 82.61	76.47 / 84.78
DeepSeek-v4-flash	FewRel	31 / 28	90.32 / 82.14	83.87 / 75.00	80.65 / 82.14
DeepSeek-v4-flash	CLINC150	28 / 26	100.00 / 100.00	100.00 / 96.15	100.00 / 100.00
DeepSeek-v4-flash	MAVEN-ETC	25 / 24	80.00 / 75.00	64.00 / 45.83	64.00 / 45.83
Qwen3-14B	SemEval-2010 Task 8	166 / 113	55.42 / 68.14	54.82 / 60.18	49.40 / 56.64
Qwen3-14B	DialogRE-SR	89 / 63	48.31 / 61.90	68.54 / 76.19	56.18 / 53.97
Qwen3-14B	KBP37	78 / 65	58.97 / 40.00	73.08 / 41.54	52.56 / 29.23
Qwen3-14B	Banking77	51 / 56	88.24 / 94.64	78.43 / 89.29	80.39 / 85.71
Qwen3-14B	Wiki80	34 / 46	67.65 / 65.22	58.82 / 65.22	58.82 / 39.13
Qwen3-14B	FewRel	31 / 28	58.06 / 60.71	74.19 / 71.43	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Qwen3-14B	CLINC150	28 / 26	100.00 / 96.15	100.00 / 96.15	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Qwen3-14B	MAVEN-ETC	25 / 24	76.00 / 66.67	68.00 / 66.67	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Qwen3-8B	SemEval-2010 Task 8	166 / 113	50.00 / 62.83	55.42 / 66.37	40.96 / 58.41
Qwen3-8B	DialogRE-SR	89 / 63	53.93 / 44.44	66.29 / 65.08	51.69 / 36.51
Qwen3-8B	KBP37	78 / 65	56.41 / 50.77	69.23 / 36.92	50.00 / 30.77
Qwen3-8B	Banking77	51 / 56	86.27 / 87.50	72.55 / 83.93	68.63 / 76.79
Qwen3-8B	Wiki80	34 / 46	67.65 / 65.22	52.94 / 60.87	55.88 / 36.96
Qwen3-8B	FewRel	31 / 28	64.52 / 60.71	54.84 / 50.00	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Qwen3-8B	CLINC150	28 / 26	100.00 / 96.15	100.00 / 92.31	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Qwen3-8B	MAVEN-ETC	25 / 24	72.00 / 70.83	64.00 / 62.50	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Mistral-Nemo-Instruct-2407	SemEval-2010 Task 8	166 / 113	33.73 / 34.51	45.78 / 46.02	30.12 / 37.17
Mistral-Nemo-Instruct-2407	DialogRE-SR	89 / 63	31.46 / 25.40	50.56 / 65.08	31.46 / 39.68
Mistral-Nemo-Instruct-2407	KBP37	78 / 65	55.13 / 30.77	73.08 / 35.38	7.69 / 12.31
Mistral-Nemo-Instruct-2407	Banking77	51 / 56	84.31 / 80.36	74.51 / 73.21	0.00 / 0.00
Mistral-Nemo-Instruct-2407	Wiki80	34 / 46	55.88 / 58.70	55.88 / 60.87	8.82 / 10.87
Mistral-Nemo-Instruct-2407	FewRel	31 / 28	45.16 / 57.14	74.19 / 64.29	0.00 / 0.00
Mistral-Nemo-Instruct-2407	CLINC150	28 / 26	100.00 / 92.31	100.00 / 92.31	0.00 / 11.54
Mistral-Nemo-Instruct-2407	MAVEN-ETC	25 / 24	52.00 / 37.50	44.00 / 50.00	0.00 / 0.00
Mistral-7B-Instruct-v0.3	SemEval-2010 Task 8	166 / 113	30.12 / 36.28	39.16 / 60.18	43.98 / 29.20
Mistral-7B-Instruct-v0.3	DialogRE-SR	89 / 63	13.48 / 22.22	44.94 / 58.73	20.22 / 23.81
Mistral-7B-Instruct-v0.3	KBP37	78 / 65	19.23 / 30.77	57.69 / 27.69	30.77 / 26.15
Mistral-7B-Instruct-v0.3	Banking77	51 / 56	72.55 / 50.00	70.59 / 83.93	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Mistral-7B-Instruct-v0.3	Wiki80	34 / 46	41.18 / 41.30	41.18 / 43.48	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Mistral-7B-Instruct-v0.3	FewRel	31 / 28	35.48 / 28.57	58.06 / 53.57	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Mistral-7B-Instruct-v0.3	CLINC150	28 / 26	92.86 / 92.31	92.86 / 96.15	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$
Mistral-7B-Instruct-v0.3	MAVEN-ETC	25 / 24	20.00 / 25.00	52.00 / 37.50	$\mathrm{N/A_{ctx}}$ / $\mathrm{N/A_{ctx}}$

A measured Accuracy of 0.00% is therefore different from $\mathrm{N/A_{ctx}}$. In particular, the zero-valued Mistral-Nemo-Instruct-2407 X^2-ICL entries on several supported conditions are observed outcomes with supported requests, not missing values. The purpose of this table is to expose model- and dataset-dependent behavior after binary gold-label evidence availability is partially matched; no pooled conditional effect, win/loss summary, or causal estimand is introduced.

D.3 Evaluated Context Feasibility

ICL, Retrieval ICL, X-ICL, and SG-XICL are context-supported in all 48 model×dataset cells. X^2-ICL is supported in 37 of 48 cells under the same fixed model conditions. Table 23 shows how the X^2-ICL support footprint varies across datasets. The unsupported X^2-ICL cells are Qwen3-14B on CLINC150, FewRel, and MAVEN-ETC; Qwen3-8B on the same three datasets; and Mistral-7B-Instruct-v0.3 on Banking77, CLINC150, FewRel, MAVEN-ETC, and Wiki80. At the seed-condition level these correspond to 22 unsupported conditions because X^2-ICL has two active randomized seeds. Unsupported conditions have no predictive score and are not assigned zero.

Table 23: Evaluated X^2-ICL context support by dataset. The other four Main methods are supported for all six models on every dataset.

Dataset	$\lvert\mathcal{Y}\rvert$	X^2-ICL supported models
SemEval-2010 Task 8	18	6/6
DialogRE-SR	30	6/6
KBP37	37	6/6
Banking77	77	5/6
Wiki80	80	5/6
FewRel	124	3/6
CLINC150	150	3/6
MAVEN-ETC	165	3/6

D.4 Explicit Reasoning and Reference-Explanation Breadth

Table 24 reports a descriptive prompt-structure comparison between canonical SG-XICL and X^2-ICL. Under canonical SG-XICL, the number of query-side candidate-wise reasoning coordinates and visible own-label reference explanations remains eight. X^2-ICL expands query-side reasoning over the full task label space and exposes eight demonstrations with one label-conditioned reference explanation per task label. These counts describe explicit prompt structure, not whole-system asymptotic complexity.

Table 24: Explicit reasoning and reference-explanation breadth. Final prediction remains over the full task label space for both methods.

Dataset	$\lvert\mathcal{Y}\rvert$	SG reason coord.	X^2 reason labels	SG ref. expl.	X^2 ref. expl.
SemEval-2010 Task 8	18	8	18	8	144
DialogRE-SR	30	8	30	8	240
KBP37	37	8	37	8	296
Banking77	77	8	77	8	616
Wiki80	80	8	80	8	640
FewRel	124	8	124	8	992
CLINC150	150	8	150	8	1200
MAVEN-ETC	165	8	165	8	1320

D.5 Online Token Burden

For each SG-XICL versus randomized-baseline comparison, token effects are first computed separately for seeds 1000 and 2000. The two seed-specific effects are then averaged within the same model$\times$dataset setting. Dataset-level summaries take the median across supported model conditions. Model-level summaries use the corresponding median across supported datasets. The Overall summary takes the median directly across all common-supported model$\times$dataset effects and is not recomputed from the displayed dataset or model medians.

The primary quantities are median input-token and total-token relative changes for SG-XICL against each baseline. Positive values indicate greater SG-XICL token use and negative values indicate lower SG-XICL token use.

Table 25: Dataset-level online token changes for SG-XICL relative to each baseline. Each entry is Input / Total relative change. X^2-ICL summaries use only context-supported model conditions. The Overall row is computed directly across common-supported model$\times$dataset effects.

Dataset	$\lvert\mathcal{Y}\rvert$	ICL	Retrieval ICL	X-ICL	X^2-ICL
SemEval-2010 Task 8	18	+39.07% / +62.71%	+45.34% / +70.10%	+6.41% / +21.51%	-82.83% / -81.51%
DialogRE-SR	30	+39.20% / +48.21%	+22.15% / +29.90%	+22.32% / +28.50%	-77.59% / -77.65%
KBP37	37	+39.40% / +59.56%	+42.96% / +63.63%	+9.30% / +20.75%	-90.88% / -89.73%
Banking77	77	+57.10% / +89.76%	+56.66% / +89.02%	+18.21% / +37.44%	-95.67% / -95.03%
Wiki80	80	+31.30% / +53.07%	+35.68% / +58.08%	+6.65% / +20.86%	-94.79% / -94.31%
FewRel	124	+33.47% / +52.27%	+31.94% / +50.28%	+10.89% / +23.96%	-95.96% / -91.37%
CLINC150	150	+40.78% / +64.30%	+41.17% / +65.08%	+13.30% / +29.67%	-97.25% / -96.11%
MAVEN-ETC	165	+26.44% / +43.86%	+27.60% / +45.52%	+6.20% / +18.81%	-96.80% / -93.62%
Overall	–	+38.21% / +62.71%	+37.11% / +65.08%	+9.88% / +28.09%	-94.58% / -90.32%

Table 26: Model-level online token changes for SG-XICL relative to each baseline. Each entry is Input / Total relative change and is the median across supported datasets after averaging the two baseline-seed effects within each model×dataset setting. The final column gives the X^2-ICL dataset denominator.

Model	ICL	Retrieval ICL	X-ICL	X^2-ICL	X^2 support
DeepSeek-v4-pro	+36.91% / +119.63%	+35.18% / +127.79%	+9.49% / +83.43%	-94.98% / -89.55%	8/8
DeepSeek-v4-flash	+36.91% / +113.99%	+35.18% / +122.51%	+9.49% / +88.98%	-94.98% / -89.81%	8/8
Qwen3-14B	+39.73% / +55.85%	+39.20% / +60.07%	+10.16% / +21.57%	-90.96% / -90.32%	5/8
Qwen3-8B	+39.73% / +54.87%	+39.20% / +58.97%	+10.16% / +21.17%	-90.96% / -89.76%	5/8
Mistral-Nemo-Instruct-2407	+38.79% / +52.14%	+37.96% / +55.13%	+10.04% / +20.10%	-95.23% / -94.67%	8/8
Mistral-7B-Instruct-v0.3	+39.24% / +56.69%	+39.95% / +61.05%	+11.08% / +24.08%	-82.82% / -81.28%	3/8

Across the 74 common-supported seed-specific SG-XICL versus X^2-ICL comparisons, SG-XICL has lower median input tokens in 74/74 comparisons. The input-token reduction ranges from 76.01% to 97.45%, with a median of 94.45%. Total-token reduction ranges from 50.35% to 97.13%, with a median of 90.24%. These results characterize online token burden rather than latency, throughput, monetary cost, or end-to-end efficiency.

E Evidence-Depth and Explanation-Scope Sensitivity

This appendix expands the targeted sensitivity analyses summarized in the main paper. The panel contains DeepSeek-v4-pro, Qwen3-14B, and Mistral-Nemo-Instruct-2407 on FewRel and MAVEN-ETC. It is targeted diagnostic evidence selected after Main outcomes were available, rather than a preregistered or population-representative evaluation panel. FewRel (124 labels) and MAVEN-ETC (165 labels) provide two large-label settings from different task families and exhibited different Main evidence-coverage behavior. The three-model panel combines one API model with two locally executed model families that showed distinct Main behavior; this diagnostic rationale does not make the panel representative of a broader task or model population.

E.1 Targeted Supplementary Scope and Support Contract

Each dataset uses a deterministic 200-query projection from the already frozen Main-500 scope,

$$D_d^{200,\mathrm{sup}} \subset D_d^{500,\mathrm{Main}}.$$

The Supplementary-200 builder first assigns one query to every canonical label and then allocates the remaining quota by residual label capacity with deterministic largest-remainder allocation and canonical-label-ID tie breaking. Within each label, query selection uses a fixed SHA256 ordering derived from the dataset and query identifiers. No query outside Main-500 can enter the supplementary panel. Gold labels are used only for this one-time scope construction and later evaluation. Candidate construction, reference ranking, explanation resolution, prompt construction, support decisions, and inference remain gold-blind after the scope is frozen.

Supplement A varies the number of selected references per candidate label over $b \in \{1, 2, 3, 4\}$ while keeping $r = 10$, $m = 8$, and $h = 8$ fixed. Thus the selected-reference count is $|S(q)| = mb = 8b$. The $b = 1$ condition is the canonical Main SG-XICL condition evaluated on the same Supplementary-200 queries. Conditions $b > 1$ are sensitivity manipulations and do not redefine the Main method. All four depths are supported on 200/200 queries in every model×dataset cell.

Supplement B fixes $b = 1$ and compares canonical own-label reference explanations with all-candidate reference explanations. Its predictive and token effects use the exact 188/200 paired common-support queries in every model×dataset cell. The 12 unsupported queries per cell are excluded from the paired estimand and are not assigned incorrect predictions or zero token values. Unless otherwise stated, fixed-panel predictive aggregates in this appendix are unweighted arithmetic means over the six model×dataset cells. Token-burden summaries instead use the median of cell-level ratio-of-medians defined below.

E.2 Evidence-Depth Predictive Sensitivity

Table 27 reports all 24 model×dataset×depth predictive conditions. Each entry is Accuracy / Macro-F1 in percent. Across the fixed six-cell panel, aggregate Accuracy is 58.33%, 61.67%, 63.25%, and

62.42% for $b = 1, 2, 3, 4$, respectively. The corresponding Macro-F1 values are 52.79%, 55.92%, 58.31%, and 57.34%. These point estimates show early gains followed by diminishing returns, but the cell-level responses are heterogeneous.

Table 27: Complete evidence-depth predictive results on the targeted Supplementary-200 panel. Each condition entry is Accuracy / Macro-F1 (%).

Model	Dataset	$b = 1$	$b = 2$	$b = 3$	$b = 4$
DeepSeek-v4-pro	FewRel	84.00 / 79.20	88.00 / 84.06	89.00 / 85.05	90.00 / 86.91
DeepSeek-v4-pro	MAVEN-ETC	63.00 / 56.69	66.00 / 59.62	66.50 / 59.56	64.50 / 59.07
Qwen3-14B	FewRel	58.50 / 52.85	66.50 / 62.37	66.50 / 63.48	67.00 / 62.91
Qwen3-14B	MAVEN-ETC	57.00 / 50.61	63.50 / 56.39	64.50 / 58.67	61.50 / 54.95
Mistral-Nemo-Instruct-2407	FewRel	46.00 / 41.94	42.00 / 35.77	48.50 / 44.18	48.50 / 42.39
Mistral-Nemo-Instruct-2407	MAVEN-ETC	41.50 / 35.43	44.00 / 37.30	44.50 / 38.91	43.00 / 37.85

Table 28 gives the corresponding cell-level point contrasts. Each entry is Accuracy difference / Macro-F1 difference in percentage points, and every contrast uses the same $N = 200$ Supplementary queries within its model×dataset cell.

Table 28: Cell-level evidence-depth point contrasts on the targeted Supplementary-200 panel. Each entry is Accuracy Δ / Macro-F1 Δ (pp); $N = 200$ for every comparison.

Model	Dataset	$b2 - b1$	$b3 - b2$	$b4 - b3$	$b4 - b1$
DeepSeek-v4-pro	FewRel	+4.00 / +4.87	+1.00 / +0.99	+1.00 / +1.85	+6.00 / +7.71
DeepSeek-v4-pro	MAVEN-ETC	+3.00 / +2.92	+0.50 / -0.05	-2.00 / -0.49	+1.50 / +2.38
Qwen3-14B	FewRel	+8.00 / +9.52	+0.00 / +1.11	+0.50 / -0.57	+8.50 / +10.05
Qwen3-14B	MAVEN-ETC	+6.50 / +5.77	+1.00 / +2.28	-3.00 / -3.72	+4.50 / +4.33
Mistral-Nemo-Instruct-2407	FewRel	-4.00 / -6.17	+6.50 / +8.41	+0.00 / -1.79	+2.50 / +0.44
Mistral-Nemo-Instruct-2407	MAVEN-ETC	+2.50 / +1.87	+0.50 / +1.61	-1.50 / -1.06	+1.50 / +2.42

Cell-level responses are heterogeneous. Mistral-Nemo-Instruct-2407 on FewRel decreases from $b = 1$ to $b = 2$ before recovering, whereas Qwen3-14B on FewRel improves sharply at $b = 2$. Several MAVEN-ETC cells decline from $b = 3$ to $b = 4$, so the aggregate $b = 3$ peak does not define a new canonical or universally optimal setting.

E.3 Evidence-Depth Same-Scope Token Burden

The token comparison uses the same 200 frozen queries for all four evidence depths. Canonical $b = 1$ token measurements are evaluated on those exact Supplementary-200 queries. For token-count quantity τ and model×dataset cell c, the cell-level change is

$$\Delta_\tau(c) = 100 \left(\frac{\text{median}(\tau_{\text{new}}, c)}{\text{median}(\tau_{b=1}, c)} - 1 \right).$$

The fixed-panel summary is the median of the six cell-level ratio-of-medians. Input and total tokens are the reported computational quantities.

Table 29 gives the complete same-scope token conditions. Table 30 gives all 18 $b > 1$ versus $b = 1$ cell-level contrasts. The fixed-panel median input-token changes are +41.59%, +84.32%, and +127.42% for $b = 2, 3, 4$ relative to $b = 1$. The corresponding total-token changes are +33.41%, +67.33%, and +102.07%. Thus token burden increases progressively with evidence depth even though predictive gains diminish.

Table 29: Complete evidence-depth same-scope token conditions. Token columns are medians on the same 200 Supplementary queries.

Model	Dataset	b	N	Input	Total
DeepSeek-v4-pro	FewRel	1	200	1897.0	4157.5
DeepSeek-v4-pro	FewRel	2	200	2741.5	5245.5
DeepSeek-v4-pro	FewRel	3	200	3626.5	5802.5
DeepSeek-v4-pro	FewRel	4	200	4521.0	6682.0
DeepSeek-v4-pro	MAVEN-ETC	1	200	1889.0	4120.0

Model	Dataset	b	N	Input	Total
DeepSeek-v4-pro	MAVEN-ETC	2	200	2520.5	4969.0
DeepSeek-v4-pro	MAVEN-ETC	3	200	3202.0	5547.0
DeepSeek-v4-pro	MAVEN-ETC	4	200	3887.5	6331.0
Qwen3-14B	FewRel	1	200	1837.0	2088.0
Qwen3-14B	FewRel	2	200	2735.5	2978.0
Qwen3-14B	FewRel	3	200	3617.0	3872.0
Qwen3-14B	FewRel	4	200	4530.0	4767.0
Qwen3-14B	MAVEN-ETC	1	200	1793.0	2056.5
Qwen3-14B	MAVEN-ETC	2	200	2472.5	2729.0
Qwen3-14B	MAVEN-ETC	3	200	3156.5	3415.5
Qwen3-14B	MAVEN-ETC	4	200	3842.5	4104.5
Mistral-Nemo-Instruct-2407	FewRel	1	200	1877.0	2087.0
Mistral-Nemo-Instruct-2407	FewRel	2	200	2824.0	3030.5
Mistral-Nemo-Instruct-2407	FewRel	3	200	3769.5	3973.5
Mistral-Nemo-Instruct-2407	FewRel	4	200	4708.0	4923.0
Mistral-Nemo-Instruct-2407	MAVEN-ETC	1	200	1877.5	2105.0
Mistral-Nemo-Instruct-2407	MAVEN-ETC	2	200	2603.5	2823.0
Mistral-Nemo-Instruct-2407	MAVEN-ETC	3	200	3332.0	3548.5
Mistral-Nemo-Instruct-2407	MAVEN-ETC	4	200	4065.0	4306.0

Table 30: Complete evidence-depth token contrasts relative to canonical $b = 1$. Relative changes are computed from within-cell token medians.

Model	Dataset	Contrast	N	Input Δ	Total Δ
DeepSeek-v4-pro	FewRel	b2/b1	200	+44.52%	+26.17%
DeepSeek-v4-pro	FewRel	b3/b1	200	+91.17%	+39.57%
DeepSeek-v4-pro	FewRel	b4/b1	200	+138.32%	+60.72%
DeepSeek-v4-pro	MAVEN-ETC	b2/b1	200	+33.43%	+20.61%
DeepSeek-v4-pro	MAVEN-ETC	b3/b1	200	+69.51%	+34.64%
DeepSeek-v4-pro	MAVEN-ETC	b4/b1	200	+105.80%	+53.67%
Qwen3-14B	FewRel	b2/b1	200	+48.91%	+42.62%
Qwen3-14B	FewRel	b3/b1	200	+96.90%	+85.44%
Qwen3-14B	FewRel	b4/b1	200	+146.60%	+128.30%
Qwen3-14B	MAVEN-ETC	b2/b1	200	+37.90%	+32.70%
Qwen3-14B	MAVEN-ETC	b3/b1	200	+76.05%	+66.08%
Qwen3-14B	MAVEN-ETC	b4/b1	200	+114.31%	+99.59%
Mistral-Nemo-Instruct-2407	FewRel	b2/b1	200	+50.45%	+45.21%
Mistral-Nemo-Instruct-2407	FewRel	b3/b1	200	+100.83%	+90.39%
Mistral-Nemo-Instruct-2407	FewRel	b4/b1	200	+150.83%	+135.89%
Mistral-Nemo-Instruct-2407	MAVEN-ETC	b2/b1	200	+38.67%	+34.11%
Mistral-Nemo-Instruct-2407	MAVEN-ETC	b3/b1	200	+77.47%	+68.57%
Mistral-Nemo-Instruct-2407	MAVEN-ETC	b4/b1	200	+116.51%	+104.56%

E.4 Reference-Explanation-Scope Sensitivity

The reference-explanation comparison keeps the query, candidate label subspace, selected references, candidate-wise query reasoning targets, and full task label space fixed. The own-label condition exposes one explanation for each selected reference's observed label. The all-candidate condition exposes explanation slots for every label in the candidate label subspace for each selected reference. Only the exact 188-query common-support set in each model$\times$dataset cell enters the paired comparison.

Table 31 reports both absolute paired-support scores and the all-candidate minus own-label effects. Across the six fixed cells, the aggregate effects are -0.80 pp in Accuracy and -0.71 pp in Macro-F1. The near-zero aggregate masks substantial cell heterogeneity and provides no evidence that broader reference explanations improve prediction on this targeted panel.

Table 31: Complete paired predictive comparison for reference-explanation scope. Absolute entries are Accuracy / Macro-F1 (%) on the exact 188-query common-support set. Δ is all-candidate minus own-label in percentage points.

Model	Dataset	N	Own-label	All-candidate	Δ
DeepSeek-v4-pro	FewRel	188	84.57 / 77.85	86.17 / 80.89	+1.60 / +3.04
DeepSeek-v4-pro	MAVEN-ETC	188	63.30 / 52.65	60.64 / 51.64	-2.66 / -1.01
Qwen3-14B	FewRel	188	59.04 / 51.42	62.23 / 54.26	+3.19 / +2.84
Qwen3-14B	MAVEN-ETC	188	56.91 / 46.94	56.91 / 47.11	+0.00 / +0.18
Mistral-Nemo-Instruct-2407	FewRel	188	46.28 / 40.60	44.68 / 36.20	-1.60 / -4.39
Mistral-Nemo-Instruct-2407	MAVEN-ETC	188	41.49 / 32.73	36.17 / 27.81	-5.32 / -4.92

The same support universe is used for token comparison: canonical own-label token measurements are restricted to the same 188 paired queries before comparison with all-candidate. Table 32 gives the complete token conditions and Table 33 gives all six paired contrasts. The fixed-panel median increases are +168.00% for input tokens and +150.38% for total tokens. The corresponding cell ranges are [+156.40%, +172.07%] and [+102.21%, +155.49%].

Table 32: Complete paired common-support token conditions for reference-explanation scope. Token columns are medians over the exact 188-query common-support set.

Model	Dataset	Condition	N	Input	Total
DeepSeek-v4-pro	FewRel	own-label	188	1897.0	4125.5
DeepSeek-v4-pro	FewRel	all-cand.	188	4955.0	8342.0
DeepSeek-v4-pro	MAVEN-ETC	own-label	188	1886.5	4162.0
DeepSeek-v4-pro	MAVEN-ETC	all-cand.	188	4837.0	8894.5
Qwen3-14B	FewRel	own-label	188	1837.0	2088.5
Qwen3-14B	FewRel	all-cand.	188	4998.0	5336.0
Qwen3-14B	MAVEN-ETC	own-label	188	1792.5	2055.0
Qwen3-14B	MAVEN-ETC	all-cand.	188	4858.0	5187.0
Mistral-Nemo-Instruct-2407	FewRel	own-label	188	1877.0	2084.5
Mistral-Nemo-Instruct-2407	FewRel	all-cand.	188	5066.0	5307.0
Mistral-Nemo-Instruct-2407	MAVEN-ETC	own-label	188	1873.5	2102.5
Mistral-Nemo-Instruct-2407	MAVEN-ETC	all-cand.	188	4985.5	5221.5

Table 33: Complete all-candidate versus own-label token contrasts on paired common support.

Model	Dataset	N	Input Δ	Total Δ
DeepSeek-v4-pro	FewRel	188	+161.20%	+102.21%
DeepSeek-v4-pro	MAVEN-ETC	188	+156.40%	+113.71%
Qwen3-14B	FewRel	188	+172.07%	+155.49%
Qwen3-14B	MAVEN-ETC	188	+171.02%	+152.41%
Mistral-Nemo-Instruct-2407	FewRel	188	+169.90%	+154.59%
Mistral-Nemo-Instruct-2407	MAVEN-ETC	188	+166.11%	+148.35%

E.5 Support and Interpretation Boundaries

The explanation-scope comparison is restricted to the same 188-query common-support set in every model$\times$dataset cell. Unsupported queries are excluded before predictive and token comparisons and are not assigned zero values. The evidence-depth and explanation-scope results are point-estimate sensitivity analyses on the fixed panel. They do not establish $b = 3$ as a global or newly canonical setting, do not imply monotonic improvement with evidence depth, and do not show that all-candidate explanations are universally harmful. Token results characterize online input and total-token burden rather than latency or end-to-end system efficiency.

F Ablation Details

This appendix reports the controlled conditions, complete condition-level results, and predefined structural contrasts for the ablation analysis in the main paper. The panel uses DeepSeek-v4-flash on MAVEN-ETC, SemEval-2010 Task 8, and DialogRE-SR, with the same 500 fixed Main queries per dataset. C0–C4 are predefined structural alternatives rather than a cumulative component-addition sequence.

F.1 Controlled Ablation Conditions

All five conditions use the complete canonical legal prediction space $\mathcal{Y}$. Their candidate label subspaces organize dedicated evidence and reasoning but do not restrict the final output space. The table below states the model-visible structural differences among conditions.

Table 34: Controlled ablation conditions. "Top-10 mean" is the canonical SG-XICL candidate construction, and "single-nearest" is the C3 alternative. "Own-label expl." denotes model-visible reference-side own-label explanation text. C2 retains the same selected references as the aligned canonical condition.

Cond.	Candidate rule	Refs.	Own-label expl.	Candidate list	Query reasoning
C0	top-10 mean	aligned	Yes	No	generic
C1	top-10 mean	aligned	Yes	Yes	generic
C2	top-10 mean	aligned	No	No	candidate-wise
C3	single-nearest	C3-aligned	Yes	No	candidate-wise
C4	top-10 mean	aligned	Yes	No	candidate-wise

For canonical-candidate conditions C0, C1, C2, and C4, candidate membership is identical for a fixed query. Selected-reference identities are aligned whenever required by the contrast, and compatible own-label explanation assets are shared across explanation-bearing aligned conditions. C1 differs from C0 by adding an explicit consolidated candidate-label list. C2 retains the selected references but removes their model-visible own-label explanation text. C3 changes the upstream candidate-construction rule to a single-nearest alternative. C4 is canonical SG-XICL.

F.2 Ablation Setting

All five conditions use DeepSeek-v4-flash, the same 500 fixed queries per dataset, and the same generation, parsing, validity, and evaluation semantics as the corresponding Main model condition in Appendix A. C4 is canonical SG-XICL. C0–C3 are the controlled structural alternatives defined in Table 34.

F.3 Complete Condition-Level Results

Table 35 reports all 15 condition-level predictive results. Each entry is Accuracy / Macro-F1 in percent on the same 500-query denominator. Means are unweighted across the three datasets and are computed from the underlying full-precision values.

Table 35: Complete condition-level ablation results on DeepSeek-v4-flash. Each entry is Accuracy / Macro-F1 (%).

Condition	MAVEN-ETC	SemEval	DialogRE-SR	Mean
C0	62.60/61.2686	80.40/80.5649	63.00/57.9815	68.6667/66.6050
C1	61.40/60.4636	79.80/79.3206	64.20/59.0667	68.4667/66.2836
C2	61.60/60.4196	60.40/59.7732	63.80/58.7261	61.9333/59.6397
C3	62.80/62.0850	67.00/67.0740	63.80/59.7988	64.5333/62.9859
C4	64.20/63.0039	85.40/86.2202	86.20/82.1770	78.6000/77.1337

These condition-level values are descriptive results for the fixed panel. Because C0–C4 are not successive component additions, their ordering is not an additive decomposition of SG-XICL.

F.4 Predefined Structural Contrasts

The four paper-facing comparisons are predefined structural contrasts. Each cell in Table 36 reports Accuracy / Macro-F1 Δ in percentage points in the direction shown in the row label. The mean is the unweighted three-dataset mean computed from the full-precision values.

Table 36: Full-precision structural contrasts on the fixed Ablation panel. Each entry is Accuracy / Macro-F1 Δ (pp) in the indicated direction.

Contrast	MAVEN-ETC	SemEval	DialogRE-SR	Mean
Label salience (C1–C0)	-1.20/-0.8050	-0.60/-1.2443	+1.20/+1.0853	-0.2000/-0.3214
Reasoning structure (C4–C1)	+2.80/+2.5403	+5.60/+6.8996	+22.00/+23.1102	+10.1333/+10.8500
Own-label evidence (C4–C2)	+2.60/+2.5843	+25.00/+26.4469	+22.40/+23.4509	+16.6667/+17.4940
Candidate construction (C4–C3)	+1.40/+0.9189	+18.40/+19.1462	+22.40/+22.3782	+14.0667/+14.1478

C1–C0 compares the explicit consolidated candidate-label list against C0 while retaining canonical candidates, aligned reference evidence, visible own-label explanations, and generic query reasoning. C4–C1 compares canonical candidate-wise reasoning with the tested candidate-salience/generic-reasoning alternative. The contrast is not matched for token count, format, position, output structure, or compute. C4–C2 compares visible own-label reference explanations with their removal while retaining canonical candidates, the same selected references, and candidate-wise reasoning. C4–C3 compares canonical top-10-mean candidate construction with the single-nearest alternative. Changing candidate membership can deterministically change selected references, explanation bindings, reasoning-label identities, and rendered context.

These contrasts test structural alternatives rather than independent additive component effects. In particular, "own-label evidence" denotes model-visible reference-side own-label explanation evidence and does not mean presence versus absence of the raw selected references.

F.5 Interpretation Boundaries

The three-dataset panel is not presented as a preregistered sample or as a random sample from a wider task or model population.

The Ablation uses point-estimate structural contrasts without active uncertainty or significance claims. C1–C0 is mixed and near neutral on this fixed panel, whereas C4–C1, C4–C2, and C4–C3 are positive on all three datasets for both Accuracy and Macro-F1. These results are panel-specific and do not establish additive contributions, universal causal effects, universal necessity of any single design choice, or generalization to all Main models and datasets.

G QUERY-LEVEL LOCAL DISCRIMINATION

Figure 5 complements the aggregate analyses with one success case and one failure case of query-level local discrimination. The pair is illustrative rather than statistically representative. The visible reasoning is explicit model-generated query-side content and is not interpreted as hidden chain-of-thought or as a causal account of the final prediction.

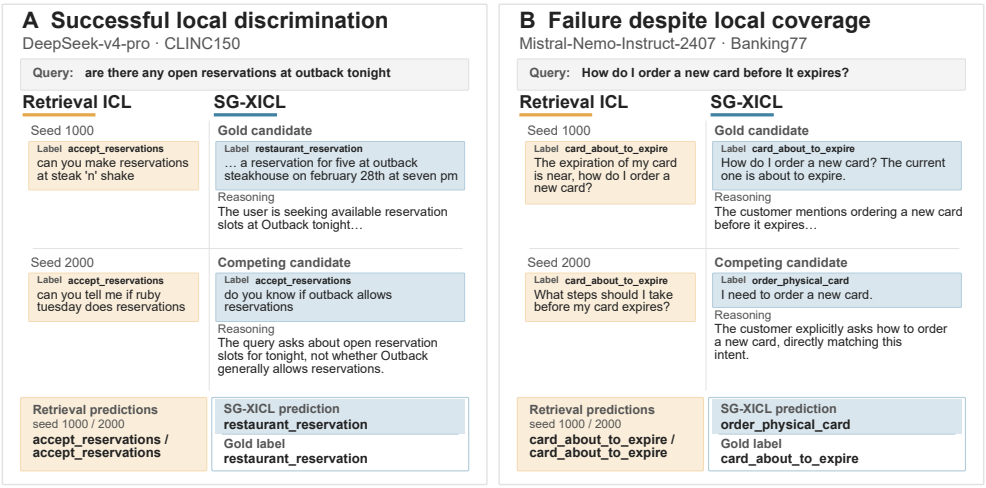


Figure 5: Illustrative query-level outcomes of local label discrimination. **(A)** CLINC150 success: both Retrieval ICL seeds predict `accept_reservations`, whereas SG-XICL predicts the gold label `restaurant_reservation`. **(B)** Banking77 failure despite local coverage: both Retrieval seeds predict `card_about_to_expire`, whereas SG-XICL predicts `order_physical_card`. For readability, only the highest-ranked Retrieval example from each displayed seed and two decision-relevant SG-XICL candidates are expanded. The actual Retrieval context contains eight demonstrations per seed, and canonical SG-XICL uses $m = 8$ labels in the candidate label subspace $T_m(q)$ with corresponding reference-side explanation structure. The candidate label subspace organizes explicit evidence and reasoning, while the final prediction remains over the full task label space $\mathcal{Y}$. These cases are illustrative rather than aggregate evidence.

In the success case, both Retrieval ICL seeds predict `accept_reservations`. The displayed top-ranked Retrieval examples emphasize whether a restaurant generally accepts reservations, whereas the query asks whether a reservation is available that night. SG-XICL places both `restaurant_reservation` and `accept_reservations` in its candidate label subspace $T_m(q)$, making the distinction between current booking availability and general reservation policy explicit. It predicts the gold label `restaurant_reservation` in this case.

In the failure case, both Retrieval ICL seeds predict the gold label `card_about_to_expire`. SG-XICL also includes this gold label in its candidate label subspace $T_m(q)$ together with the competing label `order_physical_card`. The query contains cues for both ordering a replacement card and an approaching expiry. SG-XICL predicts `order_physical_card` in this case. Because the gold label is already represented locally, this is a failure despite local gold-label coverage rather than a failure to include the gold label in the candidate label subspace.