Supplementary Material

August 5, 2025

1 Full Evaluation Results on 20-example Subset

Table 1 presents precision and coverage metrics for various LLM models and baseline pipelines evaluated on a 20-example subset. Results are shown for both Abstract-only and Full-text settings.

Table 1: Full Evaluation Results on 20-example subset: Precision and Coverage for LLM Models and Baselines.

Method	**Precision**	**Coverage**
Abstract-only		
Gemini 2.0 Flash	0.68	0.55
Llama 3 70B	0.62	0.48
GPT-4 Turbo	0.71	0.57
Full-text		
Gemini 2.0 Flash	0.59	0.62
Llama 3 70B	0.57	0.61
GPT-4 Turbo	0.63	0.65

2 Human Validation Protocol

To validate the quality of our claim extraction and matching, two independent human annotators assessed a subset of the data using the following guidelines.

Annotation Instructions

Annotators rated claim–citance semantic alignment on a 10-point scale:

- 10: Perfect match – claim and citance express the same scientific contribution with full semantic alignment.
- 9: Near-perfect match – almost identical with negligible differences in wording or emphasis.
- 8: Strong match – highly aligned, only minor contextual differences.
- 7: Clear match – substantially aligned and conveying the same core contribution, though with minor omissions or rephrasing. (Threshold for strong matches)

- 6: Moderate match – partial overlap with noticeable differences in scope or emphasis.
- 5: Weak partial match – some overlap but with significant missing or conflicting elements.
- 4: Weak match – minimal overlap, vague relation.
- 3: Very weak match – only loosely related, little semantic similarity.
- 2: Almost no match – barely related content.
- 1: No match – completely unrelated claim and citance.

Pairs rated ≥ 7 were considered strong matches in our experiments.

Quality Control and Agreement Summary

We validated the LLM-based evaluation by having two independent annotators assess 100 claim–citance pairs. The LLM's scores closely matched human ratings, with 92% within ±1 point and a strong inter-annotator agreement ($\kappa = 0.85$). An embedding-based cosine similarity metric was also evaluated, showing 78% of scores within ±1 point and lower consistency ($\kappa = 0.79$). We set a matching threshold of 7 for both methods to balance precision and recall, using LLM scores to compute coverage and precision.

3 Sensitivity to Degree-of-Match Threshold

We analyze how varying the degree-of-match threshold DM_{th} affects evaluation metrics for both the *Unsupervised* and *Weakly Supervised* pipelines under *Abstract-only* and *Full-text* settings. Lower thresholds (e.g., 6) increase coverage but reduce precision, whereas higher thresholds (e.g., 8) improve precision at the expense of coverage. Threshold 7 strikes a balance, as highlighted in Table 2.

Table 2: Effect of threshold variation (DM_{th}) on average precision and coverage for Unsupervised and Weakly Supervised pipelines, evaluated with both LLM-based and cosine similarity metrics.

Setting	**Method**	**LLM-Prec.**	**LLM-Cov.**	**Cosine-Prec.**	**Cosine-Cov.**
Abstract-only					
Unsupervised	6	0.64	0.83	0.48	0.76
	7	0.70	0.58	0.53	0.59
	8	0.74	0.48	0.58	0.45
Weakly Supervised	6	0.68	0.85	0.50	0.78
	7	0.74	0.62	0.56	0.63
	8	0.78	0.52	0.61	0.48
Full-text					
Unsupervised	6	0.55	0.82	0.44	0.75
	7	0.56	0.78	0.47	0.68
	8	0.60	0.68	0.52	0.55
Weakly Supervised	6	0.60	0.84	0.46	0.76
	7	0.63	0.79	0.49	0.70
	8	0.67	0.70	0.54	0.56