Beyond Overall F1: Fine-grained Analysis of Polymer Property Tuple Extraction from Scientific Tables

Van-Thuy Phi[1], Dinh-Truong Do[1,2*], Hoang-An Trieu[1,2*], Yuji Matsumoto[1]
[1]RIKEN Center for Advanced Intelligence Project
[2]Japan Advanced Institute of Science and Technology
{thuy.phi, truong.do, an.trieu, yuji.matsumoto}@riken.jp
(* These authors contributed equally)

Abstract

Tables in polymer science documents contain a substantial portion of polymer property knowledge essential for materials informatics, yet reliably extracting structured records remains challenging. Our prior work (Phi et al., 2025) compared a two-stage hybrid pipeline, combining LLM-based table-to-text conversion with supervised named entity recognition and relation extraction, against direct end-to-end LLM extraction on a PoLyInfo-aligned benchmark. This paper presents a fine-grained analysis of tuple-level results to identify what makes extraction succeed or fail. Using 293 manually aligned 5-ary tuples from 37 tables across 29 papers, we analyze performance by tuple completeness and table complexity. Our analysis reveals that: (1) CONDITION entities represent the primary source of extraction errors; and (2) certain table structures such as merged headers favor the hybrid pipeline while subscript notation reverses this advantage. We conclude with practical recommendations for prompt design and structure-aware model selection.

1 Introduction

Polymer property data is essential for building knowledge bases and training predictive models in materials science. However, much of this information appears in tables within scientific publications, where structural complexity makes automated extraction challenging.

Recent advances in multimodal LLMs have improved direct extraction from semi-structured content, yet end-to-end approaches force a single model to simultaneously parse layouts, resolve alignments, and extract structured tuples, often resulting in incorrect entity associations. Our prior work (Phi et al., 2025) investigated a hybrid alternative using LLMs for table-to-text conversion followed by supervised named entity recognition (NER) and relation extraction models.

While overall benchmark scores indicate the hybrid strategy often achieves superior performance, aggregate F1 alone does not explain why certain tables remain difficult nor when direct extraction may be preferable. This paper addresses these gaps through fine-grained analysis across tuple characteristics and table structures. Our contributions include: (1) analysis of extraction performance by optional entity requirements (CONDITION, CHAR_METHOD); and (2) table-level difficulty patterns through DOI-level case studies.

2 Experimental Setup

2.1 Task and Evaluation

Following the schema established in our prior work (Phi et al., 2025), we extract 5-ary tuples: (POLYMER, PROP_NAME, PROP_VALUE, CONDITION, CHAR_METHOD), where POLYMER includes homopolymers, copolymers, blends, and composites; PROP_NAME and PROP_VALUE capture the measured property and its value; CONDITION specifies measurement conditions; and CHAR_METHOD indicates the characterization technique. This schema aligns with entries in the PoLyInfo database (Otsuka et al., 2011), the largest expert-curated database for polymer property information. CONDITION and CHAR_METHOD are optional, included only when present in PoLyInfo.

Our evaluation set comprises 293 golden tuples derived from the PoLyInfo database and manually

aligned to content in 37 tables drawn from 29 scientific papers. The ground truth for each tuple was sourced from expert-curated PoLyInfo entries, which domain experts had previously extracted and validated from the corresponding papers. This alignment process ensured that annotations reflect standardized, scientifically validated property information. Each tuple contains at minimum the three essential entities of POLYMER, PROP_NAME, and PROP_VALUE, with the optional entities included when available in PoLyInfo.

For evaluation, we employ a strict exact-match criterion at the tuple level. A predicted tuple is labeled as True if and only if all constituent entities exactly match the corresponding golden tuple; otherwise, the prediction is labeled as False. We report F1@PoLyInfo as the evaluation metric, following the protocol established in our prior work (Phi et al., 2025).

2.2 Systems Under Analysis

We analyze two distinct approaches for table extraction, each evaluated using five advanced vision-language models: Claude Sonnet 4.5, GPT-4.1, GPT-4o mini, Gemini 2.5 Flash, and Qwen2.5-VL 32B Instruct.

The hybrid pipeline approach decomposes the extraction task into two sequential stages. In the first stage, a multimodal LLM receives the table image along with its caption and footnotes, and is instructed to convert each data row into a descriptive natural language paragraph. The carefully engineered linearization prompt enforces a conditional grouping strategy that creates separate paragraphs for each material, with further subdivision by characterization method when explicitly stated. In the second stage, the generated text is processed by supervised models trained on the PolyNERE corpus (Phi et al., 2024). We employ a W2NER model (Li et al., 2022) with MatBERT encoder (Walker et al., 2021) for NER and an ATLOP model (Zhou et al., 2021) with DeBERTa-v3-large encoder (He et al., 2020) for RE. The extracted binary relations are then merged into 5-ary tuples following the relation schema defined in our prior work (Phi et al., 2024).

The direct LLM extraction approach follows a conventional end-to-end paradigm wherein the vision-enabled LLM directly analyzes the table image along with associated caption and footnote text to output complete property tuples in a single inference step. Both methodologies receive identical multimodal inputs consisting of high-fidelity table images extracted using the MinerU parser (Wang et al., 2024), ensuring fair comparison. All LLM inferences were performed with deterministic settings using temperature of zero to ensure reproducibility. Prompt templates for both approaches are provided in our prior work (Phi et al., 2025).

3 Results and Analysis

3.1 Overall Performance Pattern

Table 1 presents the overall F1@PoLyInfo scores for all five models across both methodologies, reproduced from our prior work (Phi et al., 2025) as a baseline for the fine-grained analysis that follows. The results reveal substantial variation in extraction performance, with the spread between the best configuration (Claude Sonnet 4.5 with hybrid pipeline at 67.92%) and the worst (Gemini 2.5 Flash with direct extraction at 24.91%) spanning 43 percentage points.

Model	Hybrid Pipeline	Direct Extraction	Difference
Claude Sonnet 4.5	67.92	56.66	+11.26
GPT-4.1	38.23	40.61	−2.38
GPT-4o mini	48.46	48.12	+0.34
Gemini 2.5 Flash	55.97	24.91	+31.06
Qwen2.5-VL 32B	53.92	29.35	+24.57

Table 1 Overall F1@PoLyInfo scores (%) for tuple extraction. The '*Difference*' column shows the performance change when using the hybrid pipeline relative to direct extraction.

Two patterns in these results motivate deeper analysis. First, the hybrid pipeline achieves superior performance for most models, but this advantage is not universal. GPT-4.1 performs slightly better with direct extraction (40.61%) compared to the hybrid pipeline (38.23%), suggesting that error propagation during linearization can sometimes outweigh the benefits of task decomposition. Second, the large performance variance across models indicates sensitivity to specific table characteristics and tuple requirements that are not visible from aggregate scores alone.

3.2 Impact of Optional Entities

A key feature of the PoLyInfo-aligned evaluation schema is that tuples may include additional fields beyond the essential triple of POLYMER, PROP_NAME, and PROP_VALUE. To understand how these factors affect extraction difficulty, we partitioned the 293 evaluation tuples into four groups based on whether CONDITION and CHAR_METHOD entities are non-empty in the golden annotation. Table 2 presents F1 scores for each group, reporting both the mean score across all five LLMs within each approach and the best single-LLM performance.

Required Entities	Count	Hybrid Avg	Direct Avg	Best Hybrid	Best Direct
Neither CONDITION nor CHAR_METHOD	140	67.05	45.18	74.1	71.94
CHAR_METHOD only	49	47.35	48.16	81.63	57.14
CONDITION only	96	36.25	24.37	65.62	33.33
Both CONDITION and CHAR_METHOD	8	5.00	40.00	25.00	100.00

Table 2 Extraction performance stratified by tuple completeness. '*Count'* represents the number of tuples in each group; '*Hybrid Avg*' and '*Direct Avg*' represent mean F1 scores across five LLMs; Best columns show the highest single-LLM score within each approach.

The results reveal that CONDITION constitutes the dominant bottleneck for extraction systems. When CONDITION is required, encompassing 104 tuples in total, average performance drops sharply for both approaches, with particularly severe decrease observed for direct extraction. This trend highlights the structural complexity of linking condition descriptors, such as temperature columns or footnote definitions, to specific measurement cells. The difficulty is further compounded when condition text is spatially distant from the corresponding values or distributed across merged headers.

The presence of CHAR_METHOD requirements exhibits asymmetric effects. When characterization methods are required but conditions are not, the best hybrid model achieves 81.63% F1, significantly outperforming the best direct model at 57.14%. This high ceiling for the hybrid approach suggests that when method cues from the table image and accompanying text (captions, footnotes) are correctly integrated during linearization, the supervised models can extract methods effectively. However, the Hybrid Avg (47.35%) is comparable to the Direct Avg (48.16%), indicating that the hybrid advantage is highly sensitive to the base LLM's capability: if the LLM fails to preserve method cues during linearization, performance drops sharply, whereas direct extraction offers lower but more consistent performance across models.

For the small subset requiring both entities (Count=8), direct extraction significantly outperforms the hybrid pipeline (Avg: 40% vs 5%; Best: 100% vs 25%). However, these results should be interpreted with caution due to the limited sample size. The tuples derive primarily from a single paper, and the perfect 100% score by one direct model may reflect a close match with that specific table's format rather than generalizable capability. Nevertheless, the pattern suggests that the two-stage pipeline struggles to preserve complex multi-entity dependencies during linearization.

3.3 Table-Level Difficulty Patterns

To understand which specific table characteristics most significantly challenge extraction systems, we examined difficulty patterns at the DOI level. Table 3 presents representative case studies that demonstrate significant performance variation between approaches, along with the count of tuples that remain unsolved by all system configurations.

The hybrid pipeline demonstrates its strongest advantage on structurally complex tables. In the table from DOI 10.1016/j.tca.2011.11.031, which contains multiple conditions and polymer blend ratios distributed across columns, the best hybrid system achieves perfect accuracy while direct extraction remains near failure at 12.50%. Similarly, when the same property name appears with different characterization methods specified in footnotes (DOI 10.1002/pi.2987), the hybrid approach successfully disambiguates these cases while direct extraction struggles.

However, direct extraction can dominate when preserving exact character sequences is critical and the linearization process introduces alterations. The table from DOI 10.1002/polb.23373 features material names with subscript notation such as PEO_{44} and PCL_{16}.

Difficulty pattern (annotated)	Example DOI	Count	Best Hybrid (%)	Best Direct (%)	Unsolved tuples (all systems)
Merged columns; multiple measurement conditions; material names ambiguous	10.1016/j.tca.2011.11.031	32	100.00	12.50	0
'Tg' with two CHAR_METHODs (same PROP_NAME; method in footnotes)	10.1002/pi.2987	7	100.00	57.14	0
Subscript in material names	10.1002/polb.23373	8	25.00	100.00	0
Extracted property names contain redundant units	10.1002/app.44462	48	62.50	83.33	0
Missing cells; complicated conditions	10.1021/acs.macromol.5b00096	19	36.84	5.26	11
Subscript measurement method; all LLMs extracted incorrectly (subset)	10.1016/j.polymer.2015.04.040	9	66.67	33.33	3

Table 3 Representative DOI-level case studies showing best single-model accuracy for each approach and count of tuples unsolved by all system variants. Table images for these case studies are provided in Appendix A.

For these cases, direct extraction achieves 100% accuracy while hybrid performance drops to 25%. This pattern suggests that single-stage multimodal models can sometimes copy complex surface forms more faithfully than two-stage pipelines that must preserve exact character sequences through generation and downstream extraction stages.

Unit placement emerges as a recurring failure mode. The table from DOI 10.1002/app.44462, containing property names with redundant embedded units, demonstrates that models frequently mishandle unit-value associations, either incorporating units into PROP_NAME or separating them from numerical values. Notably, direct extraction outperforms the hybrid approach on this table (83.33% vs 62.50%), likely because the linearization stage can introduce additional formatting inconsistencies such as "*water uptake [wt%] is 0*" that confuse downstream entity recognition.

A notable finding is that a small number of tables account for most unsolved tuples. Just two tables account for all 14 tuples that no system variant successfully extracts: the table from DOI 10.1021/acs.macromol.5b00096 contributes 11 unsolved tuples due to missing cells and complicated conditions, while the table from DOI 10.1016/j.polymer.2015.04.040 contributes the remaining 3 due to subscript-encoded measurement methods that all LLMs failed to extract correctly. This concentration indicates that the performance ceiling for current approaches is driven by specific structural challenges rather than uniform model limitations.

4 Practical Recommendations

Our analysis reveals predictable failure patterns suggesting practical improvements. Since CONDITION is the primary source of errors, condition normalization by standardizing symbols, units, and whitespace before matching could recover semantically correct tuples that fail due to formatting differences. Method-aware linearization should explicitly preserve characterization method names (e.g., DSC, DMA, TGA) from table headers, captions, and footnotes, since omissions cause unavoidable downstream failure.

Value-unit formatting should also be consistent: linearization prompts must emit PROP_VALUE as number immediately followed by unit. The subscript finding suggests structure-aware model selection: when tables contain subscript notation in material names, direct extraction should be preferred; a lightweight classifier could direct such cases appropriately.

Finally, the concentration of unsolved tuples in specific table types indicates that future work should target subscript-aware vision features and missing-cell inference mechanisms.

5 Conclusion

This paper analyzed polymer property tuple extraction from scientific tables, demonstrating that CONDITION requirements cause most errors and that table structures systematically favor different approaches. Improved systems should incorporate normalization, method-aware linearization, and structure-aware model selection.

References

[1] He, P., Liu, X., Gao, J., and Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. **arXiv preprint** arXiv:2006.03654.

[2] Li, J., Fei, H., Liu, J., Wu, S., Zhang, M., Teng, C., Ji, D., and Li, F. (2022). Unified named entity recognition as word-word relation classification. In **Proceedings of the AAAI Conference on Artificial Intelligence**, volume 36, pages 10965–10973.

[3] Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., and Yamazaki, M. (2011). PoLyInfo: Polymer database for polymeric materials design. In **2011 International Conference on Emerging Intelligent Data and Web Technologies**, pages 22–29. IEEE.

[4] Phi, V. T., Teranishi, H., Matsumoto, Y., Oka, H., and Ishii, M. (2024). PolyNERE: A novel ontology and corpus for named entity recognition and relation extraction in polymer science domain. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pages 12856–12866.

[5] Phi, V. T., Do, D. T., Trieu, H. A., and Matsumoto, Y. (2025). A hybrid LLM and supervised model pipeline for polymer property extraction from tables in scientific literature. In **Third Workshop on Artificial Intelligence for Scientific Publications** (WASP@IJCNLP-AACL 2025).

[6] Walker, N., Trewartha, A., Huo, H., Lee, S., Cruse, K., Dagdelen, J., Dunn, A., Persson, K., Ceder, G., and Jain, A. (2021). The impact of domain-specific pre-training on named entity recognition tasks in materials science. **Available at SSRN 3950755**.

[7] Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., and Zhang, B. (2024). MinerU: An open-source solution for precise document content extraction. **arXiv preprint** arXiv:2409.18839.

[8] Zhou, W., Huang, K., Ma, T., and Huang, J. (2021). Document-level relation extraction with adaptive thresholding and localized context pooling. In **Proceedings of the AAAI Conference on Artificial Intelligence**, volume 35, pages 14612–14620.

A Sample Tables from Case Studies

DOI: 10.1016/j.tca.2011.11.031 (Table 1)

Table 1
Isothermal crystallization kinetics of PBS at 95 and 100 °C in the blends.

PBS/PES	95 (°C)		100 (°C)	
	n	k (min^{-n})	n	k (min^{-n})
100/0	2.19	2.94×10^{-2}	2.32	2.45×10^{-3}
80/20	2.20	1.15×10^{-2}	2.43	2.71×10^{-4}
60/40	2.20	1.39×10^{-2}	2.40	3.73×10^{-4}
40/60	1.94	1.40×10^{-1}	2.09	7.97×10^{-3}

DOI: 10.1002/pi.2987 (Table 1)

Table 1. Thermal properties of LPEEK/HPEEK blends with various HPEEK contents

LPEEK/HPEEK (w/w)	T_g (°C)[a]	T_g (°C)[b]	T_c (°C)	T_m (°C)	X_c (%)	T_d^5 (°C)
100/0	145	167	283	334	15.4	560
99/1	145	170	288	337	15.5	566
98/2	145	171	291	337	17.2	570
97/3	144	172	292	338	17.2	570
96/4	144	173	293	338	17.6	570
95/5	144	174	294	338	19.5	565

[a] Measured using DSC.
[b] Measured using DMA.

DOI: 10.1002/polb.23373 (Table 2)

TABLE 2 Glass Transition Temperature T_g, Melting Temperature T_m, Specific Enthalpy of Melting Δh_m, and Degree of Crystallinity X_{DSC} of the Samples[a]

Polymer		T_m (°C)	T_g (°C)	Δh_m (J g^{-1})	X_{DSC}
Linear	PGA_{17}	–	−20	–	–
	PEO_{44}	54.0	−67.0[d]	179.8	0.913
	PCL_{16}	51.8	−66.9[e]	97.2	0.619
	PCL_{25}	54.2	−66.9[e]	95.4	0.608
	PCL_{16}-*b*-PEO_{44}	52.0[c]	−63.0[f]	85.3	0.480
	PCL_{25}-*b*-PEO_{44}	54.7[c]	−63.0[f]	40.6	0.234
Grafted	PGA_{17}-*g*-PEO_{44}	54.2	−22.8[g]	127.5	0.715
	PGA_{17}-*g*-PCL_{15}	51.3	−24.8[g]	49.1	0.347
	PGA_{17}-*g*-PCL_{24}	54.0	−26.2[g]	39.6	0.270
	PGA_{17}-*g*-(PCL_{15}-*b*-PEO_{44})[b]	48.3	−20.5[g]	73.7	0.461
	PGA_{17}-*g*-(PCL_{24}-*b*-PEO_{44})	50.4[c]	−21.0[g]	34.9	0.219

[a] Values were determined by DSC measurements on isothermally crystallized samples (T_c = 40 °C) at a heating rate of 10 K min^{-1}.
[b] This sample was crystallized at 34 °C for 24 h.
[c] As T_m of PEO and PCL are very close, it was impossible to resolve individual melting temperatures by means of DSC measurements.
[d] Value taken from ref. 21.
[e] Value taken from ref. 22 for a PCL of M_n = 2900 g mol^{-1}.
[f] T_g of a PCL_{24}-*b*-PEO_{17} diblock copolymer as reported in ref. 23.
[g] Data represent the T_g of the PGA backbone. The T_g of PEO and PCL segments could not be quantitatively determined since the step in the heat flow is too small and hard to detect.

DOI: 10.1002/app.44462 (Table III)

Table III. Summary of Power Law Index (n) and Dynamic Viscosity (η^*) of POSS Filled Nanocomposites of PPS (at T = 305 °C) and PEEK (at T = 365 °C)

Sample	n	η^* (Pa s)
Neat PPS	0.10	3.60E + 06
PPS1	0.63	1.06E + 05
PPS3	0.75	4.62E + 04
PPS10	0.82	3.61E + 04
Neat PEEK	0.61	3.57E + 05
PEEK1	0.69	1.64E + 05
PEEK3	0.58	1.42E + 05
PEEK10	0.37	3.16E + 06

DOI: 10.1016/j.polymer.2015.04.040 (Table 3)

Table 3
Comparison between the values of Tg evaluated from $C_{p,DSC}$ and the peak temperatures from G″.

	Tg_{DSC}/°C	$T_{peak\ G''}$/°C
SBR41/19	−45	−29
SBR20/60	−44	−24
SBR48/19	−25	−9

DOI: 10.1021/acs.macromol.5b00096 (Table 2)

Table 2. Thermal Properties of Copolymers of Glycidyl Phenyl Ether (GPE) and Tetrahydrofuran (THF) and Their Corresponding Homopolymers

entry	composition	onset T_g (°C)	medium T_g (°C)	ΔC_p (J/g°C)	T_c^a (°C)	T_m (°C)	χ_c^b (%)
1	P(GPE)	10.7	13.4	0.385			
2	P($GPE_{0.80}$-*co*-$THF_{0.20}$)	−3.8	−0.7	0.354			
3	P($GPE_{0.63}$-*co*-$THF_{0.37}$)	−23.2	−19.1	0.368			
4	P($GPE_{0.53}$-*co*-$THF_{0.47}$)	−39.1	−30.2	0.372			
5	P($GPE_{0.35}$-*co*-$THF_{0.65}$)	−60.4	−45.7	0.612			
6	P($GPE_{0.40}$-*co*-$THF_{0.60}$)	−55.6	−39.0	0.785			
7	P($GPE_{0.42}$-*co*-$THF_{0.58}$)	−54.4	−38.3	0.70			
8	P($GPE_{0.34}$-*co*-$THF_{0.66}$)	−63.5[c]	−45.9[c]	0.717	broad (−15 to −3)[d]	9.8[d]	0.4
9	P($GPE_{0.28}$-*co*-$THF_{0.72}$)	−72.5[c]	−68.9[c]	0.60	−24.8[c]	16.1[c]	16
10	P($GPE_{0.03}$-*co*-$THF_{0.97}$)	−83.9[c]	−82.5[c]	0.21	−66.9[c]	27.5[c]	27
11	PTHF[e]	−85.6[c]	−83.0[c]	0.38	−63.0[c]	19.9[c]	25

[a]Cold-crystallization temperature (T_c). [b]Degree of crystallinity obtained from samples cooled at 5 °C/min. [c]Obtained from samples quenched in liquid nitrogen. [d]Obtained from sample cooled at 5 °C/min. [e]Purchased from Sigma-Aldrich, M_n = 20 kg/mol.