

CemLit-IE: Schema-Constrained LLM Extraction for Traceable Cement Strength Databases

Yanjie Sun¹, Mengze Hong^{1,2*}, Chen Jason Zhang¹, Haoyang Li¹

¹Hong Kong Polytechnic University

²Yuanbao Team, Tencent

{yanjie.sun, mengze.hong}@connect.polyu.hk

{jason-c.zhang, haoyang-comp.li}@polyu.edu.hk

Abstract

Cementitious materials teams need machine-readable databases of compressive strength records to support mix-design optimization. However, manual curation of the literature is too slow and hard to scale. We present CemLit-IE, a five-stage zero-shot pipeline, converting raw PDFs into fully attributed records with provenance. The core challenge is that the relevant information is distributed across tables, text, and figures. Even though table-only extraction produces plausible numbers, the results are not auditable. CemLit-IE addresses this through a series of methods, including keyword-weighted page scoring, structured table injection, value-provenance schema, and review queue. On a benchmark of 20 papers with 655 expert-annotated records, CemLit-IE achieves a recall of 76.3% with 99.2% precision and 100% exact-match accuracy in a text-only model. The recall can be raised to 83.5% when figure reading is enabled. The ablation study shows that each design choice contributes to improving precision. As for the cost, a separate 270-paper journal-batch run extracted 2,039 strength records at \$31.28 total API cost. All data and code are publicly released.¹

1 Introduction

Concrete is one of the most commonly used industrial materials on earth (Monteiro et al., 2017). Even though there is already a large and long-standing body of research in this field, the knowledge is scattered across tens of thousands of journal papers, each reporting heterogeneous tables, figures, and text in inconsistent formats. The need for reusable, machine-readable scientific data is consistent with open materials-data infrastructures such as the Materials Project (Jain et al., 2013). The traditional database development requires manually reading every paper, cross-referencing tables

against method sections and captions, and transcribing records with their units, mixture identities, curing conditions, and provenance. Even at only a few hours per paper, manually annotating 200 papers would require hundreds of researcher-hours.

There are three challenges for deploying large language models for this task. Firstly, PDFs are usually unstructured documents, where table extraction from such documents remains a dedicated document-understanding task (Smock et al., 2022). Figure information is not directly accessible to text-based systems, and critical contextual information is often scattered across multiple pages (Swain and Cole, 2016). Secondly, zero-error data quality is required for the material database. A fabricated value can corrupt downstream tasks, such as regression models, or lead to the wrong candidate formulations being recommended. This is unacceptable in industrial deployments, where not only high recall but also zero hallucinations are required (Gutierrez et al., 2022). The third challenge is operational costs and latency. A pipeline that takes a long time and has high expenses per paper will never be scheduled as a corpus-update job, even though the accuracy is high.

CemLit-IE addresses all three challenges through a five-stage pipeline. It does not apply the LLM to the full text. Instead, it narrows each paper to its 8–12 most evidence-rich pages before calling the LLM. This helps cut prompt tokens by $\approx 60\%$ compared to a full-document submission. Besides, it constrains model output to a domain-specific JSON Schema with four-tier value provenance (Willard and Louf, 2023). Combined with post-validation, this prevents unsupported values from entering the accepted database. It sends uncertain records to a lightweight review queue for further validation before accepting the record into the database. The system issues exactly one API call per paper with zero retries; in the 270-paper journal batch, the text-only pass averaged 55 seconds per

*Equal Contribution

¹<https://github.com/Yanjie-Bill/CemLit-IE>

Work	Primary target	Construction materials domain	Text	Tables	Figures	Numeric records	Value provenance	Review queue
ChemDataExtractor (Swain and Cole, 2016)	Chemical records	X	✓	✓	X	✓	X	X
Weston et al. (2019)	Materials entities	X	✓	X	X	X	X	X
Kononova et al. (2019)	Synthesis recipes	X	✓	X	X	✓	X	X
BatteryBERT (Huang and Cole, 2022)	Battery database	X	✓	X	X	X	X	X
MatSciBERT (Gupta et al., 2022)	Materials language model	X	✓	X	X	X	X	X
Dunn et al. (2022)	Structured scientific text IE	X	✓	X	X	X	X	X
Gutierrez et al. (2022)	Biomedical IE	X	✓	X	X	X	X	X
CemLit-IE (ours)	Cement strength database	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison with related scientific information-extraction and constrained-generation systems.

paper with a median latency of 45 seconds.

Our contributions are summarized as follows:

1. A keyword-weighted page scoring is introduced in this study. With such a score, the LLM prompt can focus on evidence-rich pages, helping save time and cost.
2. We combine text, tables, and images into a single context packet (priority from high to low: exact table $\succ$, explicit text $\succ$, approximate figure $\succ$, derived value). After this procedure, the machine-readable quality of the data is prepared for downstream filtering.
3. After the initial processing, a deterministic post-validator and review queue are introduced to ensure every accepted record is independently verifiable.
4. A comprehensive evaluation on 655 expert-annotated records across 20 papers, with controlled ablations and K -sensitivity analysis, is conducted to check the performance.

2 Related Work

As for the scientific information extraction, NER and relation-extraction pipelines have been applied to materials science to recover synthesis conditions (Kononova et al., 2019), battery properties (Huang and Cole, 2022), and material compositions (Weston et al., 2019). These systems mainly operate on plain text, ignoring the tabular and figure modalities (Table 1). ChemDataExtractor (Swain and Cole, 2016) extends rule-based extraction to chemistry documents. It includes the text and table extraction, but it cannot handle chart-embedded data. By comparison, CemLit-IE requires no labeled training data and addresses all modalities through structured table injection and optional page-image encoding.

Recent work has shown that GPT-family models can extract structured information from scientific text, including in low- or zero-shot settings (Gutierrez et al., 2022; Dunn et al., 2022). For database construction, however, the main concern

is not only whether a model can produce a plausible record, but whether that record can be traced back to source evidence. This is especially important in materials applications, where a hallucinated value or a misattributed condition can contaminate downstream modeling. Materials-domain models such as MatSciBERT (Gupta et al., 2022) and Mat2Vec (Tshitoyan et al., 2019) improves the representation of materials-science language, but they do not by themselves provide record-level provenance. MaTableGPT (Yi et al., 2025) moves closer to the tabular setting by using GPT-based extraction for materials tables, yet it does not address cross-page context linking or provide an explicit review queue. CemLit-IE therefore treats extraction as a database curation problem: outputs are constrained by a strict JSON Schema, linked to evidence IDs, checked by deterministic validators, and routed to human review when support is incomplete.

Developing a material database requires not only information extraction but also human intervention. Text mining can be used to reduce the search and triage burden (Yeh et al., 2003). After that, expert review plays an important role in validation, standardization, and provenance tracking (for Biocuration, 2018). This is particularly important for establishing a cement/concrete dataset: a strength value may be copied correctly but still become unusable if some important information is misread, such as the unit basis, curing age, and specimen type. Such data errors can propagate into downstream models and increase the inaccuracy (Sambasivan et al., 2021). CemLit-IE therefore treats human review as part of the database construction process: uncertain records are paired with evidence links and targeted questions before being promoted to accepted database entries.

3 Proposed Method

3.1 Task Formulation

Given the associated papers, this study aims to collect all reported compressive strengths, as well as

details about the mix design and curing methods. The record should include the identifier, strength value, unit, curing age, and property type. And the records should link to the source evidence. This approach is different from standard information extraction because the results are not just text snippets or likely structured answers. Instead, the extracted records will be used as database entries. For this reason, we set four requirements: every accepted value must be traceable to its source; unsupported values should be left out instead of being guessed; uncertain cases must be available for human review; and the process must have predictable cost and speed when applied to the entire dataset.

3.2 Pipeline Overview

CemLit-IE processes each paper through five stages, as shown in Figure 1. It first parses and renders the PDF (S1). After that, it scores pages based on domain keywords and expands the selected pages to include their neighbors (S2). The selected text, tables, and optional page images are packed into a single context packet (S3) for one schema-constrained LLM call (S4). Finally, the extracted records are checked by deterministic validators, and uncertain cases are written to a review queue (S5). Uncertain data will not be accepted into the database until it passes review. The pipeline components are chosen to support traceable database construction. Page scoring is used because strength evidence is sparse within long PDFs. If we send the full PDFs, it increases the cost and adds irrelevant context. Table injection is adopted since mixture strength often depends on a two-dimensional table structure. Each database record must contain certain important information for future applications, so Schema-constrained output is used. The review queue makes sure that uncertain or approximate records should be checked rather than accepted directly.

3.3 Page Scoring and Candidate Selection

Compressive-strength data occupy only a limited number of pages in a typical journal paper. We score each page by five keyword signal classes and select the top- K anchor pages:

$$\phi(p) = \sum_{k \in \mathcal{K}} w_k \cdot \min(c_k(p), 8) \quad (1)$$

where $c_k(p)$ is the count of signal class k on page p , and the cap of 8 prevents one dense page from dominating. Table 2 shows the signal weights. “com-

Signal k	Example patterns	w_k
Strength	compressive strength, MPa, N/mm ²	4.0
Mix design	kg/m ³ , water/binder, w/b ratio	3.0
Age	7 days, 28 d, 91 days	1.5
Unit	MPa, N/mm ²	1.5
Caption	Table 1, Figure 3	0.5

Table 2: Page-scoring signal weights (Eq. 1).

pressive strength” is the measured load-bearing strength. “mix design” refers to the material composition. “age” refers to the curing age at which the strength is measured. “unit” refers to the reported measurement unit. Each anchor expands to include its immediate neighbors to capture cross-page tables. This neighbor expansion is critical because cross-page tables may be shown in some papers, where column headers and data rows are separated by a page boundary. Our ablation study shows it accounts for 4.4 EM points.

3.4 Context Packet & S4 LLM Extraction

After the page selection, a structured JSON context packet is built for each paper. The packet contains the selected page text, the extracted table CSVs with page provenance, a rules block describing the extraction constraints, and a canonical SHA-256 hash used for caching. In vision mode, we also attach compressed page images at 1400 px resolution and JPEG quality 80. We treat figure-derived values separately because chart understanding requires visual and logical reasoning beyond plain text extraction (Masry et al., 2022), and values read from plots are usually approximate.

The model is asked to produce one JSON object that conforms to the output schema provided in the system message. To make the extracted values usable downstream, each result is assigned a provenance label, ordered from highest to lowest confidence: exact table $\succ$, explicit text $\succ$, approximate figure $\succ$, derived value. Records also include value_status, is_approximate, and estimated_error_mpa, so users can filter records by evidence quality without rereading the paper. This separation is especially important for figure-derived values. Table 3 gives a minimal example. Values read only from plots are not treated as exact records; they are assigned an image-estimate status, marked as approximate, and stored with an estimated uncertainty. In contrast, values in tables or explicit text are stored as non-approximate records. If the evidence for a field is missing or ambigu-

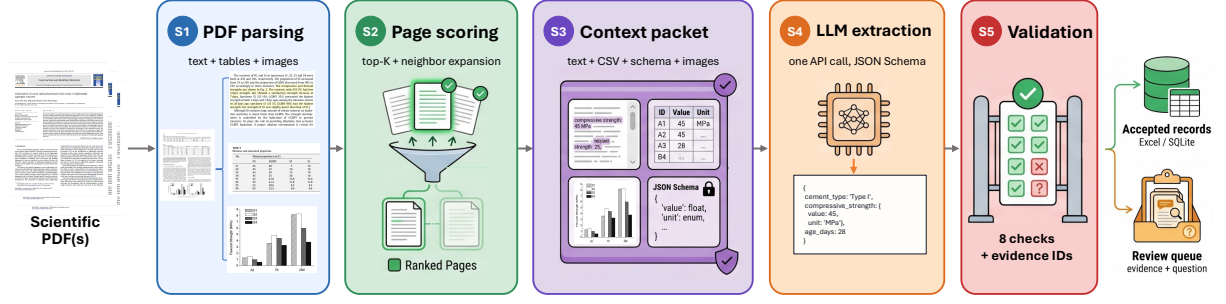


Figure 1: CemLit-IE system methodology. The pipeline narrows each PDF to evidence-rich pages, builds a multimodal context packet, performs schema-constrained LLM extraction, and validates each record.

recommended_value	final_value	value_status	is_approximate	estimated_error_mpa
63.0	63.0	image_estimate_only	yes	± 1.5 MPa
91.4	91.4	text_exact_anchor	no	

Table 3: Example schema fields for figure-derived and text-anchored values.

Code	Condition	Pri.
UNKNOWN_MIXTURE	mixture_id undeclared	H
INVALID_EVIDENCE	Evidence ID not in list	H
IMPLAUSIBLE_VALUE	Value $\notin [0.1, 300]$ MPa	H
MISSING_AGE	age_days is null	H
INVALID_UNIT	Unit $\notin \{\text{MPa}, \text{N/mm}^2\}$	H
CONTEXT_CONFLICT	Evidence implies non-CS property	H
DUPLICATE	Same (mix, age, value) repeated	M
APPROX_OR_DERIVED	Routed to human review	M

Table 4: Deterministic post-validation rules (H = High, M = Medium priority).

ous, the model writes the issue to `unresolved_items` rather than filling the field by assumption. A summary of the prompt rules and required schema fields is provided in Appendix D.

3.5 Post-Validation and Review Queue

In this stage, each extracted record is passed through the rule-based checks in Table 4. Records are sent to the review queue if they fail a check, are marked as approximate, or are linked to an entry in `unresolved_items`. The queue stores the information and a specific question for the reviewer. With this processing, we can turn verification from rereading a full PDF into checking a small number of questions. The reviewer can focus on the exact source of uncertainty, such as a missing curing age or a value read from a figure. Representative review-queue items are shown in Appendix E.

4 Experimental Setup

We evaluate on a benchmark CBM-20. It comprises 20 papers from academic journals with expert annotations. Censored inequality records (e.g., “<1 MPa”) are excluded, and finally yield 655 gold

Variant	Definition
Full (Text-only)	Selected page text + extracted table CSVs.
Full (Vision+Fig)	Full (Text-only) + selected page images.
–Table	Remove extracted table CSV injection.
–Neighbor	Remove adjacent-page expansion around selected pages.
–Scored	Replace keyword-scored page selection with first- K pages.
–Rules	Replace structured extraction rules with a short unconstrained prompt.
Full Doc	Send all pages without keyword scoring.

Table 5: System variants used in the main comparison and ablation study.

records. Of these, 77.7% carry exact table or explicit text values; 22.3% are figure estimates. Full per-paper statistics appear in Appendix A (Table 9).

System variants. As for the setting, all LLM variants use the same model setting (gpt-5.4, max 32K tokens), and we use the same JSON Schema and extraction prompt. The variants differ only in the input context or in one removed pipeline component. Table 5 summarizes the setting of each variant. The main setting is Full (Text-only), providing the model with selected page text and table CSVs. Full (Vision+Fig) uses the same packet but also includes images. The ablations remove one part of this context at a time. “–Table” removes table CSVs; “–Neighbor” removes neighbor expansion; “–Scored” replaces keyword-scored selection with sequential first- K pages; “–Rules” replaces the structured 10-rule prompt with a 3-line minimal prompt; and “Full Doc” sends all pages without scoring. Appendix B reports the sensitivity analysis for $K \in \{4, 6, 8, 12\}$. Vision+Fig is used as a targeted recovery setting rather than as the default database-building mode. It is applied to the six papers where text-only extraction is expected to miss figure-derived or low-recall records (Appendix F).

Evaluation protocol. Gold and predicted records are matched in two passes. Pass 1 (strict): exact join on (mixture_original_id, age_days). Pass 2 (soft): for unmatched gold records, accept a prediction with the same age and value within ± 0.5 MPa.

System	<i>R</i>	<i>P</i>	F_1	EM	NM	MAE
<i>Rule-based (no LLM)</i>						
pdfplumber regex	.421	.883	.569	.948	.971	0.19
<i>Schema-free LLM</i>						
Full-doc, unstructured	.752	.941	.836	.887	.921	0.89
Text-only, unstructured	.793	.912	.848	.891	.926	0.71
<i>CemLit-IE (this work)</i>						
Full (Text-only)	.763	.992	.864	1.000	1.000	0.00
Full (Vision+Fig)	.835	.977	.900	.951	.989	0.08

Table 6: Comparison with alternative strategies on CBM-20. **Bold** = best per column.

Source of recall gap	Recs	Pts
FIG : figure-only values (CBM0057, CBM0140)	40	6.1
PARSE : merged-cell table failure (CBM0083)	70	10.7
PAGE : page selection miss (CBM0134)	16	2.4
PART : partial multi-age coverage (CBM0177, CBM0192)	29	4.4
Total	155	23.7

Table 7: The recall gap (text-only mode).

Metrics: Recall (*R*), Precision (*P*), F_1 , Exact Match ($|\Delta| \leq 0.1$ MPa, EM), Near Match ($|\Delta| \leq 1.0$ MPa, NM), and MAE.

Comparison baselines. To place CemLit-IE in context, we also include three reference systems that can be run without gold annotations. “Naive Full-Doc” sends all pages to the model, with no page scoring and no structured schema. This is the zero-investment LLM baseline. “Text-only Unstructured” keeps page scoring and table injection, but replaces the JSON Schema constraint with a free-text response prompt. It tests how much is lost when the extraction is schema-free. “Rule-based (pdfplumber)” uses deterministic regex patterns over extracted table CSV rows and makes no LLM call. Together, these reference systems complement the five ablations by giving separate baselines for full-document prompting, schema-free extraction, and rule-only extraction.

5 Results and Discussion

5.1 Main Comparison

Table 6 presents the comparison between CemLit-IE and alternative extraction strategies. Figure 2 shows the predicted versus gold scatter for text-only and figure-included models.

The rule-based pipeline reaches 42.1% recall, indicating that table extraction alone is insufficient. As for schema-free LLM systems, the recall rises to $\approx 79\%$, but the EM drops. There are some error cases, such as treating flexural strength as compressive strength or assigning a value to an unsupported

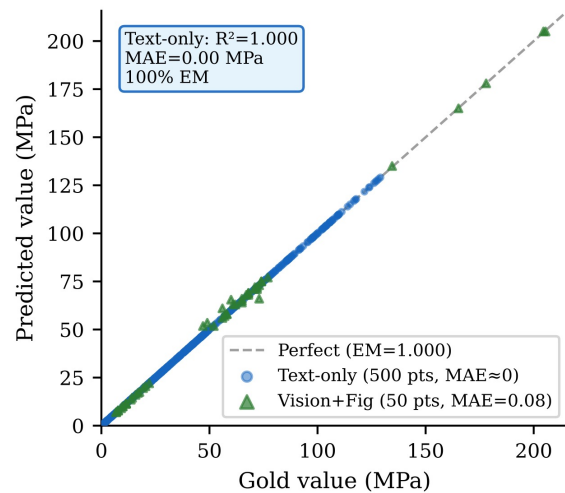


Figure 2: Predicted vs. gold compressive strength (MPa). Text-only matches ($n = 500$) lie exactly on the diagonal. Vision+Fig matches extend to figure-derived values with a wider spread, all tagged “is_approximate” = true.

mix ID. For a material database, EM = 88% is still a serious problem, since roughly 1 in 10 matched records would require correction before it could be trusted in downstream regression models.

CemLit-IE (Text-only) achieves EM = 100% on all 500 matched records, but with a low recall 76.3%. As shown in Table 7, the recall gap is due to four categories: figure-only results, the merged-cell table, a page-selection miss, and partial multi-age coverage. Detailed case illustrations are provided in Appendix G. Adding the figure extraction function raises the recall to 83.5%. This gain should be interpreted as coverage recovery rather than a replacement for text-only extraction. The text-only setting is more conservative and gives exactly matched values for records supported by tables or explicit text. Vision+Fig recovers additional figure-derived records, but these values are marked as approximate and remain distinguishable from exact records in the database. This recall gap should be interpreted as a precision and coverage trade-off rather than as unconstrained extraction failure. For materials database construction, we prefer to omit uncertain records rather than introduce unsupported or incorrect values into the accepted database.

5.2 Precision–Recall Frontier and Industrial Significance

Figure 3 plots all seven system variants in precision–recall space. The full system is the only point achieving both high precision and EM = 1.000. In industrial material databases, EM is a key metric.

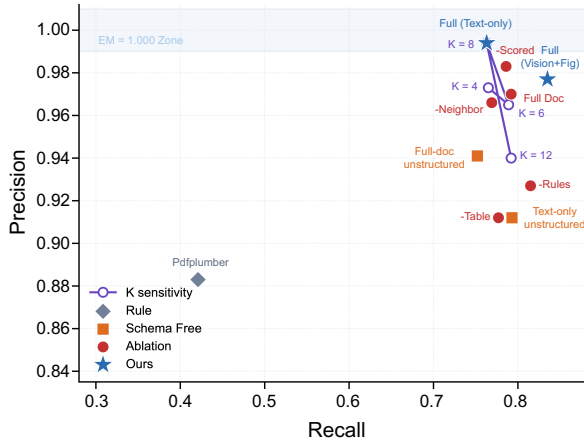


Figure 3: Precision–recall scatter for all system variants. CemLit-IE (Text-only, $K=8$) is the optimal point. Removing any component lowers precision.

Variant	R	P	EM	MAE
Full (Text-only, <i>ours</i>)	.763	.992	1.000	0.00
–Rules (simple prompt)	.815	.927	.936	0.53
–Table (no CSV inject)	.777	.912	.943	0.27
–Neighbor (no ± 1 exp.)	.769	.966	.956	0.20
–Scored (sequential K)	.786	.983	.946	0.20
Full Doc (all pages)	.792	.970	.940	0.29

Table 8: Ablation results: removing any single component degrades EM or MAE. $K=8$ throughout; K -sensitivity results in Appendix B.

If a record contains an error, it can change a mix-design decision, especially when tolerances are tight. The ablation pattern in Figure 3 shows that each part of CemLit-IE helps reduce hallucination instead of restricting coverage.

5.3 Ablation Analysis

Table 8 reports an ablation study. The ablation results should not be read as a simple recall ranking. Some removed components increase recall because the model becomes less conservative, but this also introduces unsupported or numerically incorrect records. For database construction, the preferred setting is therefore the one that preserves EM and precision, even when it leaves more records for later recovery or review. The uncertain records should be withheld or reviewed.

Prompt rules. Replacing the structured prompt with 3-line instructions yields the largest recall gain (+5.2 pts) but the worst EM (0.936) and highest MAE (0.53 MPa). Without explicit prohibitions, the model extracts values from flexural or tensile contexts.

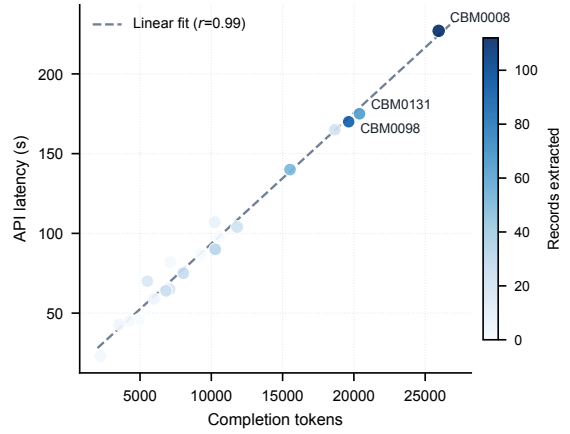


Figure 4: Per-paper API latency vs. completion tokens (text-only mode). Latency scales linearly with output length ($r=0.99$); color encodes the number of records.

Table injection. Removing table CSVs causes the sharpest precision drop (from 0.992 to 0.912) and a 5.7 pts EM collapse. That means without the 2D cell structure, values are extracted from text, which are often approximate.

Neighbor expansion. Neighbor expansion captures cross-page tables (headers on one page, data rows on the next). Disabling it drops 4.4 pts EM.

Scoring vs. sequential vs. full document. Sequential selection introduces abstract and introduction content. Full Doc uses more tokens yet achieves lower EM than the scored system, confirming that scoring is a strict improvement across both quality and cost.

5.4 Operational Efficiency

Figure 4 shows a near-linear relationship between completion length and API latency ($r=0.99$), making batch runtime easier to estimate. The system issues exactly one API call per paper with zero retries. At the upper cost estimate (\$0.42/paper), a 200-paper corpus refresh costs \$84 and completes in under 3 hours at concurrency = 2, a nightly batch job replacing weeks of manual curation. A separate unannotated journal-batch run over 270 papers from nine construction materials journals produced 2,039 strength records at \$31.28 total API cost. Text-only extraction yielded 1,749 records from 129 papers; vision recovery added 290 records and recovered 17 additional zero-result papers. Detailed efficiency metrics appear in Appendix C.

5.5 Discussion

In a materials database, a record is useful only when all associated information is correct. The text-only configuration is conservative for collecting data. It accepts records supported by tables or explicit text, but leaves figure-only cases unresolved. Although this design choice lowers recall, it prevents unsupported values from entering the accepted database. To extract more data, Vision+Fig, therefore, provides a complementary recovery mode. After the extraction from this model, the values from the plots are tagged as approximate. They are therefore treated as non-exact records, rather than as the same type of evidence as table or text supported values. And, downstream users can include, exclude, or send them for review. The ablation results show that higher recall is not always better for database construction. In several variants, recall increases as the model becomes less restrictive. However, it is accompanied by larger numerical errors. The review queue provides a practical compromise between automation and expert control. This design keeps the database expandable while preserving a boundary between accepted and unresolved records. Operationally, the one-call design and predictable latency make CemLit-IE suitable for repeated corpus updates rather than one-time extraction. The current schema is cement-specific by design because it encodes the fields that make compressive-strength records reusable: mixture identifier, curing age, unit, property type, value status, and evidence provenance. This domain specialization is intentional for an Industry Track application, where practical deployment requires a concrete schema. However, the pipeline pattern is not limited only to cement strength. Page scoring, table injection, schema-constrained output, deterministic validation, and review routing can be reused for other material properties. We therefore view cement strength as a high-value industrial test for a broader material-associated application.

6 Conclusion

CemLit-IE converts cement and concrete papers into a searchable, quality-annotated database with machine-readable provenance. On 655 benchmark records, it reaches 76.3% recall with 100% exact match in text-only mode, increasing to 83.5% when figure reading is turned on. All three main engineering choices (page scoring, neighbor expansion, and table injection) help improve output quality.

The complete system is the only configuration with EM = 1.000 while maintaining the highest precision. Every accepted record includes a provenance tag. The system uses one API call per paper, has predictable latency, and costs less than \$84 for 200 papers. In a separate 270-paper journal-batch run, CemLit-IE extracted 2,039 strength records across nine construction-materials journals at \$31.28 total API cost. This makes it practical to run at scale as a nightly batch job, removing the main bottleneck in building materials databases.

Limitations

This study has several limitations. First, CBM-20 contains 20 English-language papers. Although it includes 655 expert-annotated gold records and supports record-level evaluation, broader coverage across journals, languages, document styles, and publication years is needed. The separate 270-paper journal batch still needs to be fully expert-annotated. Future work will expand the gold dataset through broader paper collection and systematic domain-expert validation. The scoring and extraction methods focus on cement and concrete compressive strength, but more data are needed on other key properties like durability, shrinkage, carbonation depth, and chloride resistance. The current process also relies on how well PDFs are parsed. If papers have merged cells, scanned pages, or unusual table layouts, some records may be missed. This is an area that still needs improvement. Finally, while the review queue helps reduce the workload for experts, expert judgment is still necessary.

Ethical Considerations

The pipeline extracts structured information from published scientific articles for research database construction. No personal data is involved.

Acknowledgments

This work was partially supported by NSFC/RGC Joint Research Scheme (N_PolyU5179/25), Hong Kong RGC (PolyU25600624), Innovation Technology Fund (ITS/052/23MX, PRP/009/22FX), industrial sponsors and RMGS (P0045948, P0048183, P0048191, P0046453, P0060272).

References

- Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.
- International Society for Biocuration. 2018. Biocuration: Distilling data into knowledge. *PLoS Biology*, 16(4):e2002846.
- Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.
- Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512.
- Shu Huang and Jacqueline M Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *Journal of chemical information and modeling*, 62(24):6365–6377.
- Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).
- Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):203.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.
- Paulo JM Monteiro, Sabbie A Miller, and Arpad Horvath. 2017. Towards sustainable concrete. *Nature materials*, 16(7):698–699.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “everyone wants to do the model work, not the data work”: Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.
- Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.
- Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.
- Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.
- Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.
- Alexander S Yeh, Lynette Hirschman, and Alexander A Morgan. 2003. Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *arXiv preprint cs/0308032*.
- Gyeong Hoon Yi, Jiwoo Choi, Hyeongyun Song, Olivia Miano, Jaewoong Choi, Kihoon Bang, Byungju Lee, Seok Su Sohn, David Buttler, Anna Hiszpanski, Sang Soo Han, and Donghun Kim. 2025. Matablegpt: Gpt-based table data extractor from materials science literature. *Advanced Science*, 12(16):2408221.

A Per-Paper Results (Text-only Mode)

Table 9 reports per-paper statistics for the text-only baseline. Papers with $R = 0$ are attributable to known pipeline limitations, not model hallucination.

B K Sensitivity

K	R	P	EM	MAE
4	.765	.973	.954	0.15
6	.789	.965	.942	0.21
8 (default)	.763	.992	1.000	0.00
12	.792	.940	.931	0.28

Table 10: K sensitivity (text-only mode, CBM-20).

Paper ID	Article	Gold	Pred	Match	R	P	MAE	RMSE	R^2
CBM0008	Arellano Aguilar et al. (2010), lightweight metakaolin-fly ash concrete	112	112	112	1.00	1.00	0.00	0.00	1.000
CBM0018	Papayianni and Anastasiou (2010), high-strength concrete with industrial by-products	18	18	18	1.00	1.00	0.00	0.00	1.000
CBM0038	Naji Givi et al. (2010), rice husk ash particle size effects	27	27	27	1.00	1.00	0.00	0.00	1.000
CBM0047	Son et al. (2011), waste tyre rubber-filled reinforced concrete columns	6	6	6	1.00	1.00	0.00	0.00	1.000
CBM0048	Al-Jabri et al. (2011), copper slag as fine aggregate	51	51	51	1.00	1.00	0.00	0.00	1.000
CBM0050	Mazaheripour et al. (2011), polypropylene fibers in lightweight SCC	8	8	8	1.00	1.00	0.00	0.00	1.000
CBM0052	Gunasekaran et al. (2011), coconut shell concrete	20	22	20	1.00	0.91	0.00	0.00	1.000
CBM0053	Chakradhara Rao et al. (2011), recycled aggregate concrete under impact load	16	16	16	1.00	1.00	0.00	0.00	1.000
CBM0057	Kim et al. (2011), bottom ash aggregates in high-strength concrete	22	0	0	0.00	—	—	—	—
CBM0083	Nguyen Van Tuan et al. (2011), rice husk ash UHPC	70	2	0	0.00	0.00	—	—	—
CBM0085	Shafagh et al. (2011), oil palm shell lightweight aggregate	32	32	32	1.00	1.00	0.00	0.00	1.000
CBM0098	Megat Johari et al. (2011), supplementary cementitious materials in high-strength concrete	91	91	91	1.00	1.00	0.00	0.00	1.000
CBM0099	Sahin et al. (2011), matrix and steel fibre tensile strengths	21	21	21	1.00	1.00	0.00	0.00	1.000
CBM0129	Abosrra et al. (2011), reinforcement corrosion in concretes of different strengths	3	3	3	1.00	1.00	0.00	0.00	1.000
CBM0131	Uysal et al. (2011), SCC with mineral admixtures	64	64	64	1.00	1.00	0.00	0.00	1.000
CBM0134	Limeira et al. (2011), concrete with dredged marine sand	16	0	0	0.00	—	—	—	—
CBM0140	Deo and Neithalath (2011), pervious concrete porosity and compressive response	18	0	0	0.00	—	—	—	—
CBM0148	Pelletier-Chaignat et al. (2012), limestone filler with calcium sulphoaluminate cement	24	24	24	1.00	1.00	0.00	0.00	1.000
CBM0177	Meng et al. (2012), nano-TiO ₂ effects on cement mortar	20	3	3	0.15	1.00	0.03	0.06	1.000
CBM0192	Megat Johari et al. (2012), high-volume ultrafine palm oil fuel ash concrete	16	4	4	0.25	1.00	0.00	0.00	1.000
Total	—	655	504	500	.763	.992	0.00	0.00	1.000

Table 9: Per-paper results in text-only mode. Article information is given as the first author, year, and shortened title. MAE and RMSE are in MPa. $R=0$ papers are attributable to known pipeline limitations rather than model hallucination: CBM0057 and CBM0140 are figure-only; CBM0083 has merged-cell tables; CBM0134 has a page-selection miss.

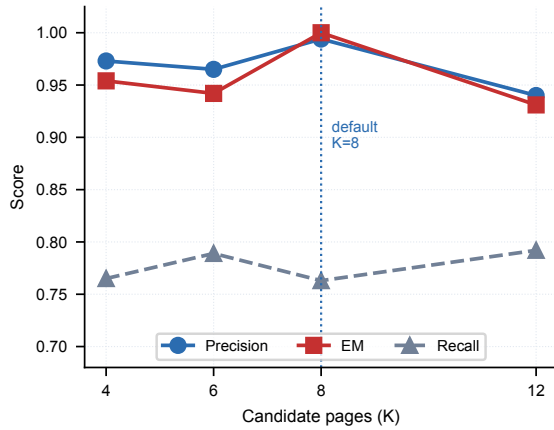


Figure 5: Sensitivity of EM, precision, and recall to candidate page count K . Token cost (bars) is nearly flat; EM peaks uniquely at $K=8$.

Table 10 shows results across $K \in \{4, 6, 8, 12\}$. Figure 5 visualizes the Precision–EM curve. Token cost is nearly flat across all K values (≈ 21.5 k tokens/paper) because the 30k-character text budget is the binding constraint. $K=8$ is uniquely the only configuration achieving EM=1.000.

C Operational Efficiency Metrics

Table 11 summarizes the operational footprint on CBM-20.

Metric	Text-only	Vision+Fig
Latency / paper (s)	87	$\approx 60^\dagger$
API calls / paper	1	1 (0 retries)
Prompt tokens / paper	11,077	$\approx 15,600$ (+41%)
Completion tokens / paper	10,390	$\approx 5,600$ (−46%)
Total tokens (20 papers)	429,330	—
Est. cost / paper (\$)	0.11–0.42	+30–40%
200-paper corpus (\$)	22–84	—
200-paper wall time	<3 h	—

Table 11: Operational efficiency on CBM-20.

† Vision+Fig: single-threaded, 6 papers only.

D Prompt and Schema Summary

As shown in Table 12, the prompt instructs the model to return one JSON object and to leave unsupported fields unresolved rather than infer values from general context. The output schema enforces five top-level arrays: mixtures, results, conditions, evidence, and unresolved_items. Each result requires 12 fields including value_status, is_approximate, estimated_error_mpa, and evidence_ids. Setting additionalProperties: false throughout prevents the model from inserting invented fields. The schema is embedded verbatim in the system message at inference time; no constrained decoding is required.

Prompt rule	Purpose
Use source evidence only	Prevent values from being inferred without support.
Keep table, text, and figure values separate	Preserve the provenance tier for downstream filtering.
Do not merge different mixtures	avoid silently combining records across mix identifiers.
Report missing fields in unresolved_items	Convert uncertainty into reviewable questions.
Mark figure-derived values as approximate	Separate chart readings from exact table or text values.

Table 12: Summary of the main prompt rules used with the output schema.

E Review Queue Examples

Table 13 shows representative review-queue items created by the post-validator. The queue is intended to make human review specific: instead of rereading the full paper, the reviewer receives the candidate record, the evidence location, and the unresolved question.

Case	Reason	Reviewer question
Missing age	Age not stated near value	What curing age does this value use?
Unknown mix	Mixture ID not defined	Which mixture does this value belong to?
Approx. figure	Value read from plot	Should this approximate value be accepted?
Unit conflict	Unit basis unclear	Is the reported unit MPa or N/mm ² ?
Duplicate record	Same mix-age-value repeated	Is this a duplicate or a separate specimen?

Table 13: Examples of review-queue items generated by deterministic checks.

F Vision+Figure Per-Paper Breakdown

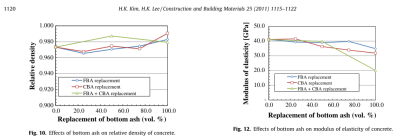
Table 14 reports the six papers processed in Vision+Figure mode.

Paper	Text <i>R</i>	Vis <i>R</i>	ΔR	MAE
CBM0057	0.00	1.00	+1.00	0.86
CBM0140	0.00	0.83	+0.83	0.15
CBM0177	0.15	0.40	+0.25	2.51
CBM0083	0.00	0.07	+0.07	—
CBM0134	0.00	0.00	0.00	—
CBM0192	0.25	0.25	0.00	0.00

Table 14: Vision+Fig per-paper results (MAE in MPa). CBM0057 fully recovered. CBM0083 remains low: bottleneck is table parsing, not figures.

G Representative Error Cases

This appendix gives source excerpts for the four main categories in the text-only recall gap. Each case illustrates a different failure mode and, therefore, a different route for improvement: figure recovery, table reconstruction, page selection, or partial multi-age extraction.



40 MPa with use of 50% bottom ash as a fine aggregate [27]. Kurma and Kaya reported that the compressive strength of concrete was decreased from 41 MPa to 29 MPa in accordance with the replacement of bottom ash as fine aggregate up to 50% [28]. On the contrary, Kayali reported an increase of compressive strength and indirect tensile strength of concrete with fly ash aggregate (FA), corresponding with bottom ash in this study [5]. The compressive strength of concrete was increased from 58 MPa to 67 MPa when granite natural aggregate were replaced up to 100% with FA and coarse aggregates, even though the VFC ratio, dosage of silica fume, and superplasticizer were fixed (Kayali, 2008). In this study, the proportion of cement paste was higher than that in other preceding works. Approximately 522 L/m³ of cement paste (water, binder materials, and superplasticizer) was applied in this study, whereas the amount ranged from 294 L/m³ to 368 L/m³ in Kurama and Kaya and Kayali [2,28]. Therefore, the effects of bottom ash aggregates were smaller than in the previous studies. To clarify the effect of bottom ash aggregates on the compressive strength of concrete, further research is needed with respect to aggregate proportions, water absorption condition of the aggregate, and the S/a ratio, important factors in lightweight aggregate concrete.

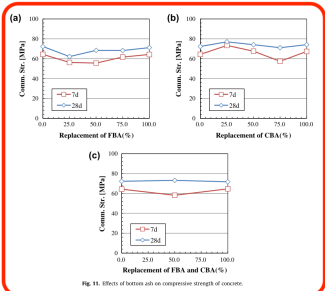


Figure 6: **FIG case.** Source excerpt from CBM0057, page 6. The compressive-strength values are reported as plotted curves rather than as a machine-readable table. The text-only pipeline, therefore, abstains instead of estimating unsupported values from the figure.

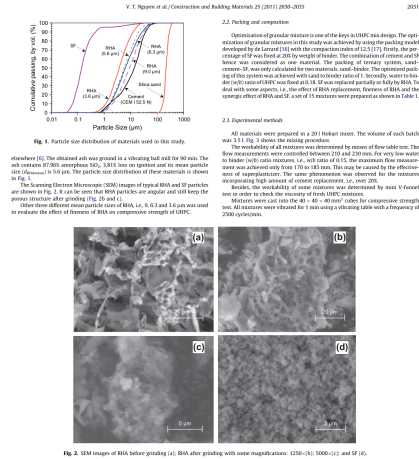


Table 1
SFA: composition used in this study

Water to binder ratio (by weight)	Sand to binder ratio (by weight)	BFA (S by weight)	SF (S by weight)	The mean particle size of BFA (mm)
0.18	1	0-10-20	1	5.6
0.18	1	10-20-30	1	5.6
0.18	1	30	1	5.6
0.18	1	5	15	5.6
0.18	1	15	5	5.6
0.15-0.18-0.20-0.23	1	00	50	5.6

Figure 7: **PARSE case.** Source excerpt from CBM0083, page 2. The composition table contains range-style entries and visually grouped cells, which makes reliable reconstruction of the mixture-condition alignment difficult after PDF table extraction.

Table 10
Fresh properties by slump test for concrete.

Type of concrete	BC	CA15S	CB25S	CB35S	CB45S	CC25S	CC35S	CC45S
Slump (mm)	1.0	0.0	1.0	3.0	1.5	2.0	1.5	1.5

Table 11
Physical properties of hardened concrete.

	BC	CB25S	CB35S	CB45S
Apparent density (kg/m^3)	2.50	2.51	2.46	2.52
Dry density (kg/m^3)	2.37	2.34	2.28	2.37
Dry mass density (kg/m^3)	1.84	1.81	1.76	1.81
Water absorption capacity (%)	3.0	3.0	2.7	3.0

Table 12
Mechanical properties of hardened concrete (MPa).

Concrete	Compressive strength		Tensile strength	Splitting tensile strength	Adhesion (N/mm ²)
	7 days	28 days			
BC	32	35	41,000	5.0	3.2
CA15S	30	33	-	4.5	-
CB25S	30	30	1.7	4.5	1.6
CB35S	31	30	30,000	5.5	5.4
CB45S	37	39	40,000	5.0	5.4
CC25S	30	30	40,000	4.9	5.1
CC35S	31	30	40,000	5.2	5.0

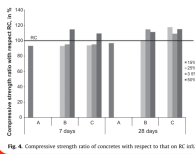


Table 13
Compressive strength evolution from 7 to 28 days of curing.

Concrete	BC	CA15S	CB25S	CB35S	CB45S	CC25S	CC35S	CC45S
7 days	32	30	30	31	37	30	31	30
28 days	35	33	30	30	39	30	30	30

3.2.3. Durability properties

The following section discusses in detail each one of the tests performed in order to evaluate the influence of DMS material on the durability of concrete, and also the influence of partial substitution of raw crushed sand. Table 14 summarizes the corresponding durability properties of all concretes produced after 28 days of moist curing. Each of the presented values is the average of three data.

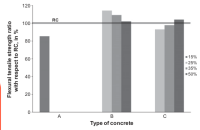


Fig. 5. Thermal shrinkage ratio of concrete at 28 days with respect to that on BC, BC.

3.2.3.1. Accessible pores. The replacement of FAI 0/4 mm by DMS-B and DMS-C did not produce any significant change in the accessible pores with respect to BC, regardless of the replacement rate. Comparing the values obtained in this research work with those collected in Linares et al. (2011), a smaller volume of accessible pores was observed in the present work, probably due to the lower effective water adopted (0.44 against 0.49) and, consequently, the more denser cement paste obtained.

The use of DMS material (which was finer than crushed limestone sand FAI) in substitution of FAI, improved the particle packing density of the cementitious system. Also, probably due to the irregularity tendency of FAI and DMS-B, entrapment of very small capillary pores might have occurred and consequently concrete made with DMS-C achieved better properties, as observed by Lehigh (1982). The CC25S concrete obtained values of accessible pores an 8.5% lower than BC.

3.2.3.2. Sorptivity. Fig. 6 presents the volume of water absorbed per unit surface area versus the square root of time. From this figure a linear trend was obtained for each type of concrete, thus revealing sorptivity from the slope of these trend lines. It was observed that the slope of all the trend lines is sharper at the beginning of the process. All the concretes presented a similar evolution during the test. The values found in the literature [31,39] revealed that the initial surface absorption results showed no significant difference between mixtures with sea sand and the control mixtures made with natural sand.

Concrete	Accessibility		Sorptivity (mm ² /s)		Water penetration (mm)	LPV (mm/s)
	BC	CA15S	CB25S	CB35S		
BC	7.1	0.008	2.7	2.5	4400	4400
CA15S	7.1	0.008	1.8	2.1	4400	4400
CB25S	7.1	0.008	1.9	2.1	4400	4400
CB35S	7.1	0.008	1.9	2.1	4400	4400
CC25S	7.1	0.008	1.9	2.1	4400	4400
CC35S	7.1	0.008	1.9	2.2	4400	4400

Figure 8: **PAGE case**. Source excerpt from CBM0134, page 7. The page contains the relevant 7-day and 28-day compressive-strength table, but it was not selected by the keyword-scored candidate-page step. This is a page-selection miss rather than an extraction hallucination.

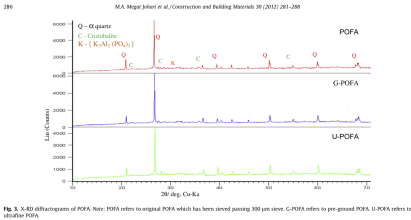


Fig. 3. XRD diffractograms of POFA. Note: POFA refers to original POFA, which has been sieved passing 300 µm sieve. C-POFA refers to pre-ground POFA, U-POFA refers to ultrafine POFA.

Concrete	Slump (mm)	Initial setting time (min)	Final setting time (min)
OPC	190	140	285
POFA20	190	140	285
POFA40	225	270	460
POFA60	230	330	515

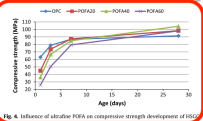


Fig. 4. Influence of ultrafine POFA on compressive strength development of HSC. Note: The results of these earlier studies somehow portrayed the limitation of DMS, and also, the use of ultrafine POFA (maximum strength obtained was 85.0 MPa at a replacement level of 20%). The reason for using ultrafine POFA is that the ultrafine POFA is a much more efficient pozzolanic material, which not only leads to greater strength enhancement, but also allows the POFA to be used at higher replacement levels of 40% and yet achieving higher 28 days compressive strength than the OPC-HSC.

5.4. Porosity and initial surface absorption

The test results for porosity as shown in Fig. 6 clearly demonstrate that the use of the ultrafine POFA reduces the porosity of

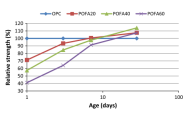


Fig. 5. Relative strength of HSC with different ultrafine POFA content.

the HSC at all of the testing ages. At 3 days, the recorded porosity is 8.4%, 7.5%, 7.8% and 7.9% for OPC, POFA20, POFA40 and POFA60, respectively. The porosity reduces to 7.5%, 7.2%, 6.5% and 6.1% at 7 days and subsequently to 7.8%, 6.9%, 6.0% and 5.8% at 28 days for OPC, POFA20, POFA40 and POFA60, respectively. In addition, it is also evident that at 3 days, the increase of POFA content from 20% to 60% is to slightly increase porosity from 7.5% to 7.9%. Nonetheless, at 7 and 28 days, the opposite trends could be observed, i.e. the porosity reduces with increasing POFA content. The reduction in porosity for all concrete mixes from 3 to 28 days indicates that the OPC, POFA20, POFA40 and POFA60 record a reduction in porosity of 7%, 8%, 21% and 27%, respectively. Hence, the higher the POFA content, the greater the reduction in porosity with longer period of water curing.

In the case of initial surface absorption test (SAT), the test results which were performed on 3, 7 and 28 days are shown in Fig. 7. The results for SAT values determined for time interval of two hours clearly show that the inclusion of the ultrafine POFA reduces the SAT values of the POFA-HSC, in comparison to the OPC-HSC. In particular, at higher POFA content of 40% and 60%, as well as at longer curing period of 28 days. In comparison to the OPC-HSC, the POFA-HSC experiences a reduction in SAT values of 43%, 43%, and 67% for POFA20, POFA40 and POFA60, respectively. The observed overall reduction in porosity and initial

Figure 9: **PART case**. Source excerpt from CBM0192, page 6. The paper reports compressive-strength values across multiple curing ages, but only part of the age series is recovered by the text-only system. The remaining values require either better evidence linking or review.