Literature Mining with Large Language Models to Assist the Development of Sustainable Building Materials

Yifei Duan (MIT), Yixi Tian (MIT), Soumya Ghosh (IBM), Richard Goodwin (IBM), Vineeth Venugopal (MIT), Jeremy Gregory (MIT), Jie Chen (IBM), Elsa Olivetti (MIT)

MIT IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY
DMSE DEPARTMENT OF MATERIALS SCIENCE & ENGINEERING
MIT-IBM Watson AI Lab

TL;DR: well-designed instruction schemes and fine-tuning strategies yield SOTA performance on complex entity inference using small LLMs; extracted information provide both an extensive summary and novel insights on the beneficial uses of alternative raw materials in concrete

Background & motivation

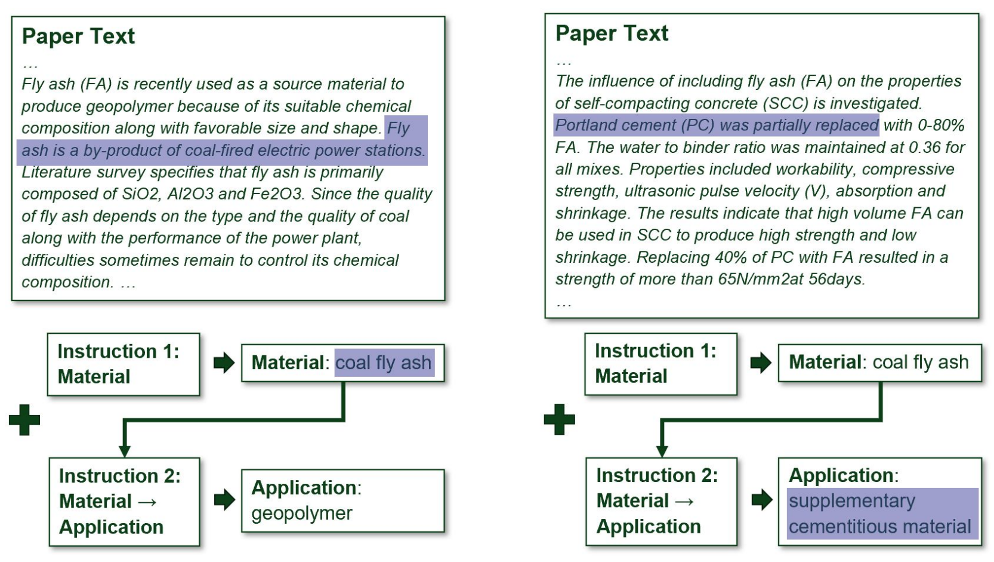
- **Concrete production** is one of the largest industrial sources of global CO_2 emissions, constituting **8~9% of annual emissions**
- Previous studies have extensively investigated using **alternative raw materials** including processed natural minerals, recycled demolition and construction waste, industrial residues and agricultural and municipal waste residues **to substitute the constituents of concrete**
- The **lack of a systematic summary** impedes the advancement of commercially viable climate-friendly concrete production
- An **exhaustive literature review** to enrich the knowledge base of the **applications of alternative materials** in concrete production is necessary and made possible with **recent advancement of LLMs**

Data

- **Hierarchical keyword-based retrieval:** 51,295 papers on concrete → 6,995 papers on alternative raw materials
- **Data preparation:** 102 papers divided into a training corpus of 82 and testing corpus of 20. Annotation was performed for (1) material, (2) application, (3) product. Examples are derived and augmented to form the training (7080) and testing (1200) sets

Methods

Instruction-based Entity Inference Schemes



Paper Text

…
Fly ash (FA) is recently used as a source material to produce geopolymer because of its suitable chemical composition along with favorable size and shape. Fly ash is a by-product of coal-fired electric power stations. Literature survey specifies that fly ash is primarily composed of SiO2, Al2O3 and Fe2O3. Since the quality of fly ash depends on the type and the quality of coal along with the performance of the power plant, difficulties sometimes remain to control its chemical composition. …

Instruction 1: Material → **Material**: coal fly ash

Instruction 2: Material → Application → **Application**: geopolymer

Paper Text

…
The influence of including fly ash (FA) on the properties of self-compacting concrete (SCC) is investigated. Portland cement (PC) was partially replaced with 0-80% FA. The water to binder ratio was maintained at 0.36 for all mixes. Properties included workability, compressive strength, ultrasonic pulse velocity (V), absorption and shrinkage. The results indicate that high volume FA can be used in SCC to produce high strength and low shrinkage. Replacing 40% of PC with FA resulted in a strength of more than 65N/mm2at 56days.
…

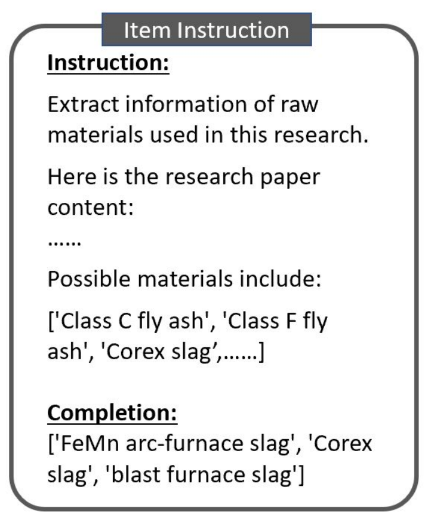
Instruction 1: Material → **Material**: coal fly ash

Instruction 2: Material → Application → **Application**: supplementary cementitious material

- Complex logical inference from **non-N**oun**P**hrase **sources** is required for entity extraction
- Conventional **N**amed **E**ntity **R**ecognition not applicable
- Devised **separate instructions** to guide the entity inference

Problem Setting: Multiple Choice

- **Item Instruction** scheme: choices provided as lists of items (names)
- **Multiple Choice** scheme: choices denoted with double-digit notations



Item Instruction

Instruction:

Extract information of raw materials used in this research.

Here is the research paper content:

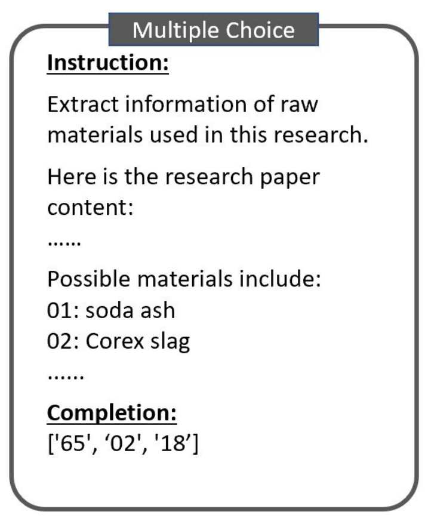
……

Possible materials include:

['Class C fly ash', 'Class F fly ash', 'Corex slag',……]

Completion:

['FeMn arc-furnace slag', 'Corex slag', 'blast furnace slag']



Multiple Choice

Instruction:

Extract information of raw materials used in this research.

Here is the research paper content:

……

Possible materials include:
01: soda ash
02: Corex slag
……

Completion:

['65', '02', '18']

Choice Permutation:

- To avoid random guessing and LLM sensitivity to ordering
- To augment dataset
- Augmented examples from the same paper either all in training or all in testing set, to avoid data leakage

LLMs and Adaptation Methods

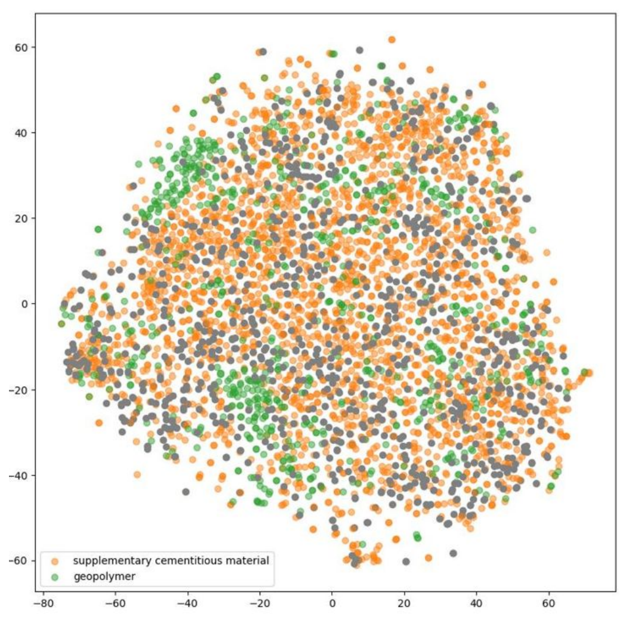
- **Fine-tuning:** On small open-source LLMs, **pythia-2.8B** and **dolly-3B** (pythia instruction-tuned on common sense questions), using both instruction schemes.
- **Baseline In-context Learning:** On state-of-the-art large models (175B), **GPT-3.5**, using both instruction schemes

Results & Analysis

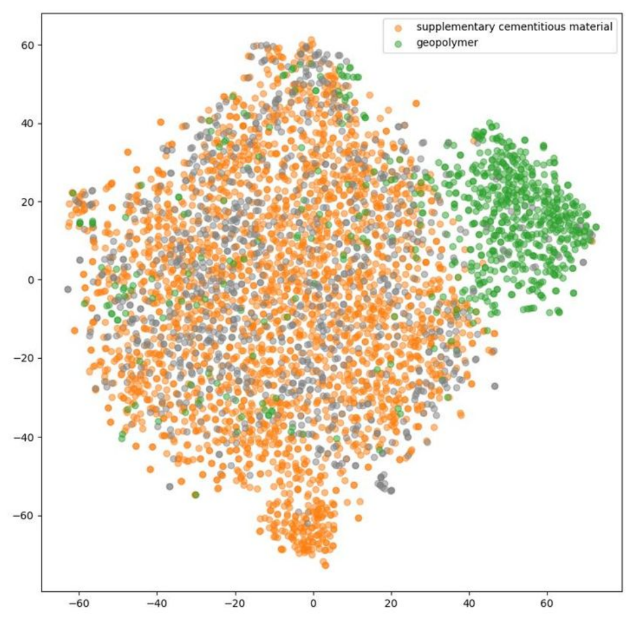
Model Performance

Model	Instruction Scheme	F1 Score	Precision	Recall
pythia-2.8B	Item Instruction	77.0	78.2	75.7
pythia-2.8B	Multiple Choice	**79.0**	**81.2**	**77.0**
pythia-2.8B	Without Options	30.5	33.3	28.1
dolly-3B	Multiple Choice	69.9	71.0	68.9
dolly-3B	Item Instruction	60.4	61.3	59.5
dolly-3B	Without Options	20.3	20.8	19.8
gpt-3.5 @4-shot	Item Instruction	57.2	62.8	52.6
gpt-3.5 @4-shot	Multiple Choice	51.9	46.8	58.1

Pythia fine-tuned with the **Multiple Choice** scheme outperforms all other model-scheme combinations, highlighting:
(1) the **effectiveness of multiple-choice instruction** design
(2) the potential of **small, free models** to attain **SOTA** performance (vs GPT-3.5)
(3) the importance of **considering domains** in model selection (vs dolly)



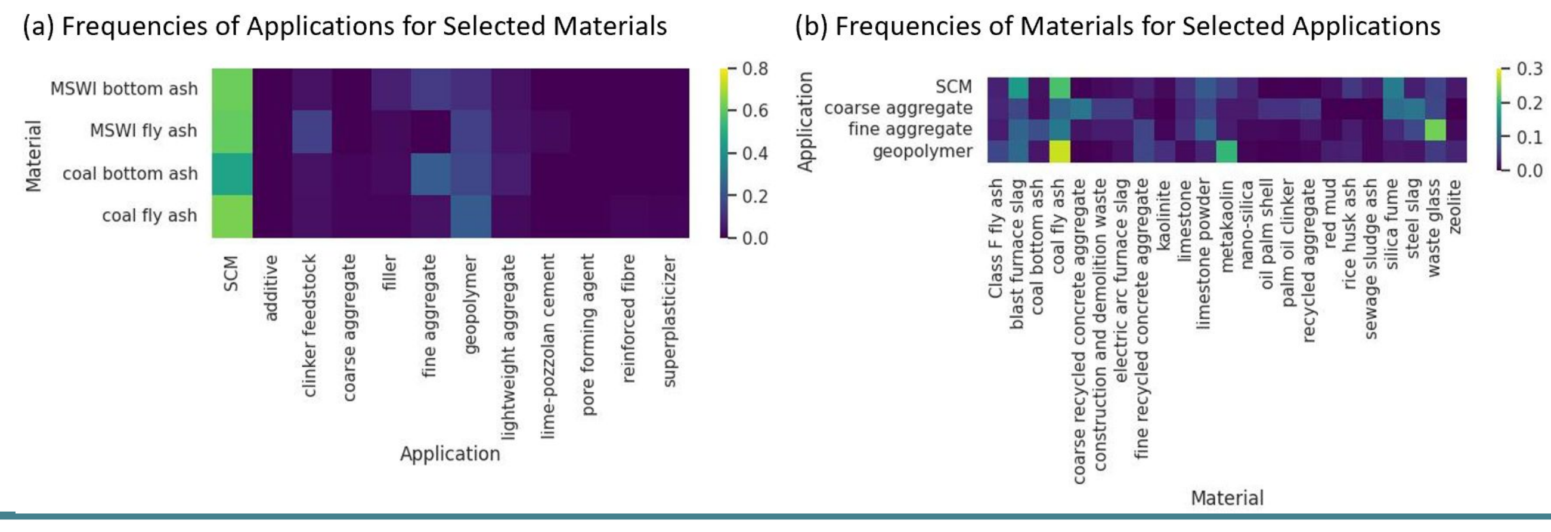
Pre-trained pythia-2.8B



Fine-tuned pythia-2.8B

Changes of **clustering effects** in the plots show that **fine-tuning** effectively improves information extraction performances through **adaptation on the embedding level**

Extracted Information Analysis

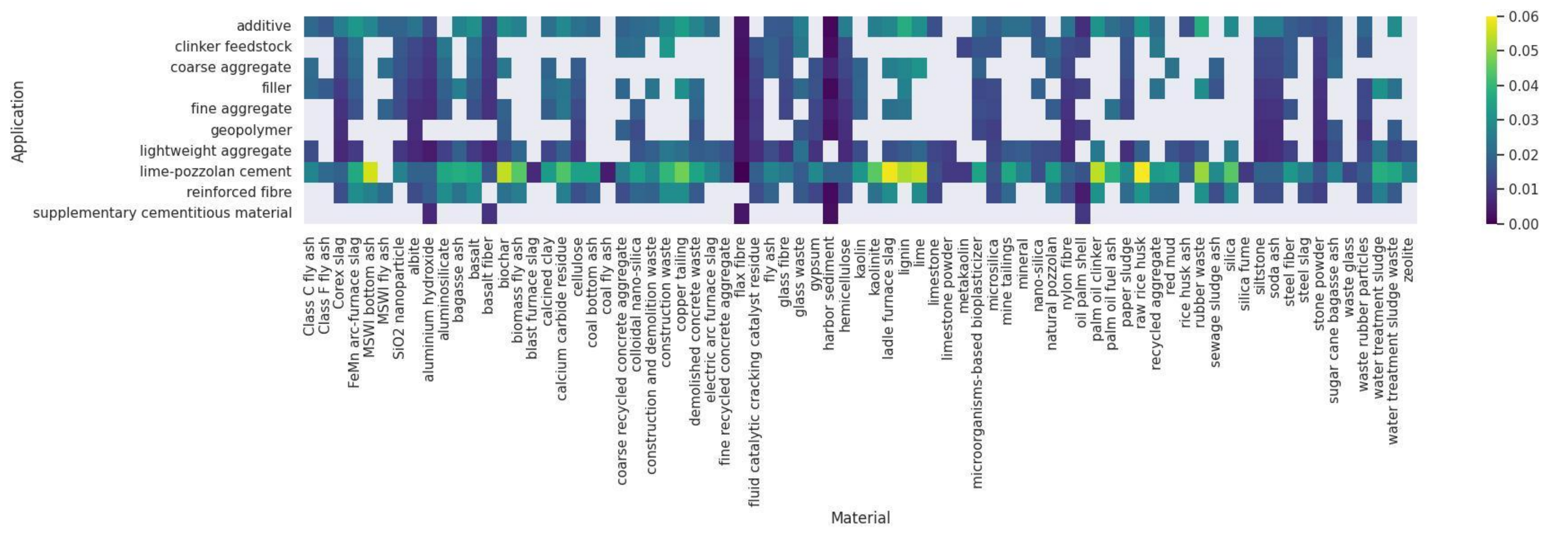

(a) Frequencies of Applications for Selected Materials

(b) Frequencies of Materials for Selected Applications

Information was extracted from an extensive **un-annotated corpus**, and utilized to construct a **knowledge graph**.

- SCMs most common application, followed by geopolymer and aggregates
- Identified promising applications of materials for future industrial deployment (e.g. SCMs: limestone powder, rice husk ash, waste glass. geopolymer: metakaolin, coal fly ash; fine aggregate: waste glass; coarse aggregate: recycled concrete aggregate)

Graph Insights - Link Prediction



Link prediction to guide further studies:

- Notable potential links including a range of industrial, municipal and agricultural residues as lime-pozzolan cement, construction waste as clinker feedstock, copper tailing and rubber waste as fillers, etc.
- Graph structural insights partly verified by domain knowledge (e.g. MSWI bottom ash as lime-pozzolan cement)