



FireSeek-Sci: A large language model with fire science knowledge for smart building safety

Yifei Ding^a, Hongtao Zheng^a, Tong Lu^a, Yanfu Zeng^b, Rong Deng^a, Yuxin Zhang^{a,*}, Xinyan Huang^{a,**}

^a Research Centre for Smart Urban Resilience and Firefighting, Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China

^b Arup Australia Services Pty Ltd., Sydney, Australia

ARTICLE INFO

Keywords:

Large language model
Fire knowledge search
Retrieval augmented generation
Fire safety science
Safety management
Smart building

ABSTRACT

Fire science knowledge is fundamental to fire science education and the subsequent learning and practicing fire protection engineering. The fragmented knowledge and time-consuming information retrieval make it difficult to obtain reliable answers in time, which seriously affects learning and practical application. To address this challenge, this study builds an expert-level fire domain knowledge model, FireSeek-Sci, based on the advanced pretrained large language model (LLM). Firstly, a specific fire domain-knowledge (FDK) database was established covering key areas, including fundamentals of combustion, fire dynamics, fire protection engineering, etc., organized in 6750 Q&A pairs. Secondly, the base LLM was fine-tuned using FDK and achieved significant improvements with over 30 points in generative semantic accuracy. Thirdly, retrieval augmented generation (RAG) technology was further embedded with the fine-tuned fire-specific LLM, further enhancing the accuracy and richness of the generated corpus over 69 %. Based on these developments, an LLM-powered fire science knowledge chatbot named FireSeek-Sci was deployed within a user-friendly web interface. Internal testing with participants demonstrated positive evaluations, with an overall satisfaction score of 3.3 (1–5 scale) and 83 % of participants expressing willingness to use the chatbot again. To further evaluate its engineering applicability, FireSeek-Sci was applied in illustrative case studies to interpret AI-generated simulation outputs into natural-language safety assessments and recommendations, which showed good quantitative grounding, consistent reasoning, and responses that were closer to practical fire engineering judgment. In conclusion, FireSeek-Sci has strong potential as an intelligent assistant for fire science education, research, and building safety applications on fire risk and emergency management.

1. Introduction

Fire science knowledge referring to the basic theories about fire and combustion behavior and firefighting protection, is fundamental in fire science education and advancing the application of technologies in fire safety and engineering (Chen et al., 2026; Quintiere, 2006; Society of Fire Protection Engineers, 2016), including fire modelling (Merci and Beji, 2022), fire protection systems (Zeng et al., 2025), structural design (Buchanan and Abu, 2016), wildfire risk (Chen et al., 2024), firefighting strategies (Huang and Tam, 2024), etc. Research and development of fire science for the researchers and engineers highly rely on the multi-source knowledge database. Existing related simulation

software, e.g., Fire Dynamics Simulator (FDS) (McGrattan et al., 2021) and Pathfinder (Lu et al., 2026a) can assist newcomers in quickly performing fire modeling and calculations. However, a thorough understanding of the underlying principles of the simulation software requires substantial knowledge such as fluid mechanics (Merci and Beji, 2022), numerical methods, and fire dynamics (Quintiere, 2006), which makes the learning process considerably steep. However, the ever-increasing formats and amounts of literature have caused challenges in efficiently learning, retrieving, and synthesizing these domain-specific knowledge.

Although the development of the internet search engines (i.e., Google, Baidu) and scientific databases (i.e., Google Scholar, Web of Science, CNKI) has greatly expanded access to information (Mettrop and

* Corresponding author.

** Corresponding author.

E-mail addresses: yx.zhang@polyu.edu.hk (Y. Zhang), xy.huang@polyu.edu.hk (X. Huang).

<https://doi.org/10.1016/j.jintel.2026.100233>

Received 21 May 2026; Received in revised form 10 August 2026; Accepted 17 August 2026

Available online 18 August 2026

2772-9915/© 2026 The Authors. Published by Elsevier Ltd on behalf of Zhejiang University and Zhejiang University Press Co., Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Nieuwenhuysen, 2001), retrieving relevant fire science knowledge remains challenging. Such difficulties arise because professional information is often fragmented across journal papers, technical standards, numerical simulation manuals, and other user guides of some specific fire services systems (Bromann, 2001; Nolan, 2011). Traditional search engines rely heavily on keyword matching but overlook semantically related content, making the search results unclear and inaccurate. Moreover, the burden of reviewing and understanding large volumes of these search results still falls on students, researchers and engineers, causing huge time costs and learning troubles. These limitations become particularly severe in cross-disciplinary knowledge, for example, studying wildland urban interface (WUI) fire (Chen et al., 2024; Huang et al., 2026; Ramesh et al., 2025) requires mastering wildfire spreading, heat transfer principles, and structural fire response, which requires much higher efficiency and quality of knowledge retrieval and integration. Hence, even in the era of digital information, efficient and accurate retrieval of fire science knowledge remains a significant challenge for related students, researchers, and engineers (Xiong et al., 2024), requiring a higher-tech knowledge search engine to address this issue.

Recent advances in large language models (LLMs) (Zheng et al., 2024) have demonstrated remarkable capabilities in natural language understanding (Tao et al., 2026), semantic search, and scientific reasoning (Jin, 2025). Therefore, LLM-powered chatbots have emerged as a promising solution to the limitations of conventional search engines. Unlike keyword-based retrieval systems, such chatbots can interpret user queries at the semantic level and generate integrated responses by synthesizing information from multiple sources, thereby addressing the issues of fragmented and sometimes inaccurate search results. However, their direct application in fire science poses challenges: generic LLMs often lack domain-specific knowledge, may generate hallucinations, and are not optimized for scientific accuracy. Previous research (Hostetter et al., 2024) has highlighted the potential for current general LLMs (i.e., OpenAI's ChatGPT and Google's Bard) to revolutionize fire engineering practices by providing instant answers to general fire science questions (i.e., "What is fire engineering?" and "Explain the process of fire growth in a compartment"). However, the accuracy and stability of pre-trained general LLMs cannot be satisfied in most specific and professional fire knowledge and fire protection standards (i.e., "Equation of the Grashof number", and "Design value of fire load in residential buildings"). Moreover, the hallucination problem (Farquhar et al., 2024) of LLMs is also serious, as they easily generate answers that appear to be correct and logical but are actually completely unreasonable statements. Too much misinformation may mislead firefighters' actions and decrease users' trust in LLMs-powered chatbots (Ji et al., 2023). In addition, proprietary LLMs such as GPT-4 and GPT-5 can only be accessed through APIs, which do not guarantee data privacy in future applications on sensitive fire accident investigations. However, the open-source LLMs (i.e., Llama and Qwen) may fail to achieve satisfactory results due to a lack of sufficient fire science knowledge.

Although domain-specific LLMs have been successfully developed for fields such as medicine (Thirunavukarasu et al., 2023), law (Ma et al., 2025), building and civil engineering (Cui et al., 2026; Dong et al., 2025; Kampelopoulos et al., 2025), directly transferring these approaches to fire engineering remains challenging. Fire engineering knowledge is highly heterogeneous, spanning fire dynamics, combustion science, structural fire engineering, evacuation, wildfire, and fire protection regulations. In addition, authoritative knowledge is distributed across textbooks, technical standards, engineering guidelines, and research literature with diverse formats and update cycles. Since fire engineering is a safety-critical discipline, inaccurate or hallucinated responses may lead to inappropriate engineering decisions or safety risks. These characteristics impose substantially higher requirements on factual accuracy, knowledge grounding, and domain adaptation than those encountered in many general-purpose question-answering applications. These limitations indicate that a fire-domain LLM is required.

The key question becomes how to build it: training a brand-new LLM from scratch or enabling an existing one to acquire fire knowledge. Training a new LLM from scratch would require billions of tokens and large-scale GPU clusters, making it impractical for most fire research teams. In contrast, fine-tuning an existing open-source LLM enables efficient domain knowledge injection at a small fraction of the cost, while preserving the model's general linguistic and reasoning abilities (Lu et al., 2026a).

In addition, although recent advances in artificial intelligence and physics-based simulation have significantly improved the ability to predict fire development in complex infrastructure environments (Zhang et al., 2021, 2024), most existing fire prediction models and engineering tools (Zeng et al., 2025) primarily output raw physical parameters, such as heat release rate (HRR) (Wang et al., 2022) and temperature field distributions (Cheung et al., 2025; Xie et al., 2025). While these quantitative outputs provide valuable information for fire safety analysis, their engineering interpretation often requires substantial expertise, time, and cognitive effort from fire engineers, emergency responders, or infrastructure operators. In practical fire emergencies, decision-makers are frequently required to assess fire severity, smoke spread, occupant tenability, and evacuation risks within a very limited time window (Xie et al., 2026; Zhang et al., 2023). Under such conditions, directly interpreting multiple evolving physical indicators and translating them into actionable safety decisions can be both challenging and error-prone.

This limitation highlights an important gap between fire forecast and engineering decision-making. Existing AI prediction models can estimate *what is happening in a fire scenario*, but they typically cannot explain *what these predictions mean* for occupant safety, evacuation feasibility, or emergency intervention (Deng et al., 2026). As a result, the burden of transforming numerical predictions into engineering judgments still largely depends on human expertise. Although existing LLM-based engineering decision-support systems (Lu et al., 2026b; Xie et al., 2026) have demonstrated promising capabilities in assisting engineering analysis, most of these systems rely heavily on commercial large language model APIs (e.g., GPT-based services), resulting in relatively high deployment costs and potential concerns regarding data privacy. To this end, the professional fire science LLMs can be used to translate the predicted physical parameters to easy-understanding condition explanation and safety optimization recommendations, but the related products or studies are still limited.

Based on this strategy, we further develop FireSeek-Sci, A fire science knowledge question-answering system based on the light-weight pre-trained foundation LLMs (Qwen2.5 1.5B and 3B versions) that integrates the fine-tuned fire domain knowledge model with retrieval-augmented generation (RAG), enabling semantic-level interaction with fire safety literature and standards. The goal of FireSeek-Sci is to improve accessibility, usability, and learning efficiency of fire science education for both researchers and practitioners, and to lay the foundation for future fire-domain LLM development. In addition to educational applications, FireSeek-Sci is further evaluated in fire and evacuation safety scenarios, where it can assist in interpreting AI-generated simulation results and translating engineering parameters into actionable safety insights for emergency decision-making.

Above all, the main contributions of this study are as follows,

- (1) Fine-tuned a domain-specific LLM for fire science education using authoritative textbook.
- (2) Developed a fire science chatbot for intuitive learning interactions between learners and LLM.
- (3) Integrated RAG from external fire safety knowledge for more reliable and scientific answers.
- (4) Enabled interpreting smart building simulation results to understandable safety assessments.

2. Methodology

2.1. Fire Domain-Knowledge (FDK) dataset

The domain-knowledge dataset plays a critical role as the textbook and learning documents for the LLM training. In this work, the domain-knowledge dataset is original from the famous SFPE Handbook for Fire Protection Engineering (Society of Fire Protection Engineers, 2016), which is regarded as the Bible for fire engineers. This book provides a detailed introduction to the current best practices and knowledge in fire protection engineering and fire safety science as one of the most authoritative and definitive references in this domain. The book contains 90 chapters contributed by over 130 outstanding and experienced engineers and researchers from universities and professional organizations all over the world. The contents of this book cover fundamentals of combustion and fire science (e.g., Introduction of fluid mechanics, Heat transfer), fire dynamics (e.g., Compartment fire modeling, Fire load density), fire evacuation (e.g., Egress concepts and design approaches, Human behavior in Fire), and fire risks (e.g., Building fire risk analysis, Fire Risk in Mass Transportation), etc. While the SFPE handbook serves as the fundamental and general baseline for FireSeek-Sci fine-tuning, its answer generation originates from a single knowledge database, which limits its ability to solve more region-specific problems across different standards and codes. To address this limitation, the system also includes other important and granular documents, e.g., Eurocodes as a supplementary knowledge base in RAG parts (detailed in Section 2.3).

Pre-training and fine-tuning large language models require to format the raw text from the PDF file into machine-interpretable structures. This work aims to develop a fire knowledge Q&A chatbot, so the dataset must comply with a standardized question-answer pair format, using the *Alpaca* dataset (Taori et al., 2023) as the dataset template. The dataset generation process is shown in Fig. 1. In Step 1, the PDF file of the SFPE Handbook was split into 90 sub-files according to the chapters for subsequent lightweight processing of the raw text data. In Step 2, each PDF file was input into the ChatGPT-4 to generate and clean the Q&A pairs. ChatGPT-4 can help to understand the text content, generate related follow-up questions, and extract corresponding answers from the raw text, according to the inputted prompt "Please create 5 long, complete, unique Q&A pairs about according to the PDF file. The answers should be the original paragraph or sentence in the PDF file and cover the whole perspectives". Due to the limitations of the token length of ChatGPT-4, the size of each uploaded PDF file cannot be so large. Therefore, we will split the large PDF file into many small files first. Each PDF can generate up to 25 Q&A pairs, resulting in a total of 2250 pairs.

After that, all generated Q&A pairs have been carefully reviewed by three domain experts, each holding a professorship or senior engineering position in fire safety science and combustion science during the

dataset collection process. All three experts independently reviewed the full set of generated QA pairs, evaluating each item based on their professional expertise and domain knowledge. The review process focused on verifying the accuracy, consistency, and domain relevance of the generated content, and whether the corresponding answer was factually correct and consistent with established fire safety and combustion science principles. No formal written review protocol or scoring rubric was used; assessments relied on the reviewers' individual expert judgment. Any ambiguous or incorrect entries were revised or removed to ensure the overall quality and reliability of the dataset. After data review and cleaning, the total data pairs decrease to 2238. Thirdly, all the questions were rephrased into two alternative forms while keeping the meaning and answers unchanged, so that 2238 pairs were augmented to 6714 pairs. Specifically, only the questions were rephrased to help the model understand different ways of asking. All answers were kept verbatim as original text segments from the handbook to ensure technical accuracy. The data augmentation can increase the model's generalization during training. After that, the dataset was structured as *Alpaca* format by a Python script and were randomly disordered and then split into the training subset (5707 Q&A pairs, 85 %), and the test subset (1007 Q&A pairs, 15 %). The original JSON file containing complete 6714 pairs are attached in the Supplementary Materials.

Importantly, the SFPE Handbook was not used for direct reproduction or redistribution of copyrighted content. Instead, it served solely as a reference source for manually extracting and reformulating fire engineering knowledge into a structured question-answer format for research purposes. The released FDK dataset does not contain verbatim reproductions of the original handbook content, nor does it redistribute the original copyrighted material.

2.2. Fine-tune process

General LLMs are typically pre-trained in large-scale general corpora, which may lack sufficient fire domain knowledge. Consequently, this can result in limited understanding and reasoning ability for fire safety tasks. To address this issue, we propose to use a fire domain-knowledge dataset to fine-tune the base LLM, aiming to enhance its answering ability in fire-related problems. For each Q&A pair, the tokenizer of the base model encodes the text into a token sequence $x = (x_0, x_1, \dots)$, and we conduct the fine-tuning on the base model $f(x)$ in an autoregressive manner using the following objective function ($\mathcal{L}$),

$$\mathcal{L}(\Theta, \mathcal{D}_{FDK}) = \mathbb{E}_{x \sim \mathcal{D}_{FDK}} \left[- \sum_{i \in \{\text{output}\}} \log f(x_i | x_0, x_1, \dots, x_{i-1}) \right] \quad (1)$$

where, Θ represents the set of model parameters, initialized from a pre-

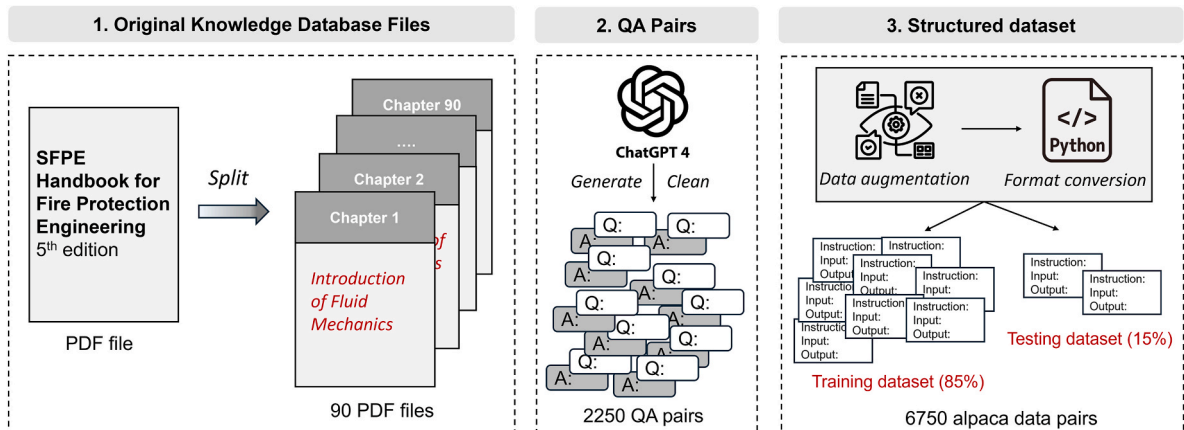


Fig. 1. The generation flowchart of fire domain-knowledge dataset.

FDK Q&A pair with Alpaca format:

```
{
  "instruction": "What is pyrolysis and how is it different from liquid gasification?"
  "input": "",
  "output": "Pyrolysis is the thermal decomposition of solid fuel into smaller gaseous molecules, usually involving complex chemical reactions. Unlike liquids, which can gasify by a simple phase change, solids require chemical breakdown before releasing flammable gases. Pyrolysis is typically endothermic and strongly temperature-dependent."
}
```

Stanford Alpaca-En-Demo:

```
{
  "instruction": "Give an example of a job that a computer can do better than a human being."
  "input": "",
  "output": "An example of a job that a computer can do better than a human being would be analyzing vast amounts of data accurately and quickly. For instance, a computer can analyze and find patterns in large datasets within milliseconds, far surpassing the capabilities of a human being. This makes a computer ideal for jobs such as financial analysis, medical diagnosis or data-driven decision making, where speed and accuracy are crucial."
}
```

Fig. 2. Examples of FDK Q&A pair and Alpaca-En-Demo Q&A pair.

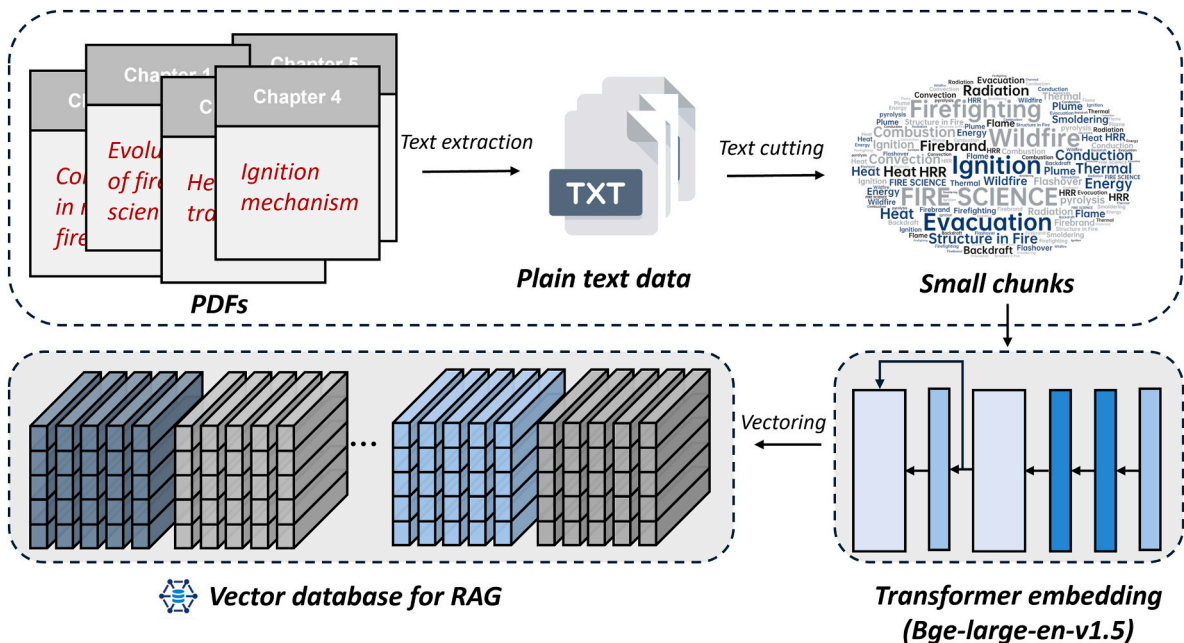
trained checkpoint, $\mathcal{S}_{FDK}$ denotes the fine-tuning fire domain-knowledge (FDK) dataset, $\mathbb{E}_{x \sim \mathcal{S}_{FDK}}$ indicates averaging over all training samples.

In this work, we used the LLaMA-Factory platform to conduct the fine-tuning task. The selected base models are open-access Qwen2.5-1.5B-Instruct and Qwen2.5-72B-Instruct, which are the instruction-tuned versions with 1.54 billion and 3.09 billion parameters. During the training process, we not only use the FDK dataset, but also insert the Stanford Alpaca-En-Demo dataset which is a general corpus with 999

Q&A for generic LLMs to reduce over-fitting risk and retain general linguistic and semantic knowledge while improving its performance on fire-related downstream tasks. Examples of an FDK Q&A pair and a Stanford Alpaca-En-Demo Q&A pair are shown as Fig. 2. The fine-tuning was conducted in Ubuntu 22.04.3 LTS, with the computing source of 8 vCPU, 96 GiB vRAM, 512 GiB OS-vDISK, 1 RTX-6000Ada, and 48 GB GPU Card. The major hyperparameter settings are listed on Table 1.

Table 1
Hyperparameter settings in fine-tuning.

Finetuning type	Learning rate	Max grad norm	Batch size	Gradient accumulation step	Epoch	Enable thinking
LoRA	1e-4	1.0	2	8	3,0	None

**Fig. 3.** Generation process of a vector database for RAG retrieval.

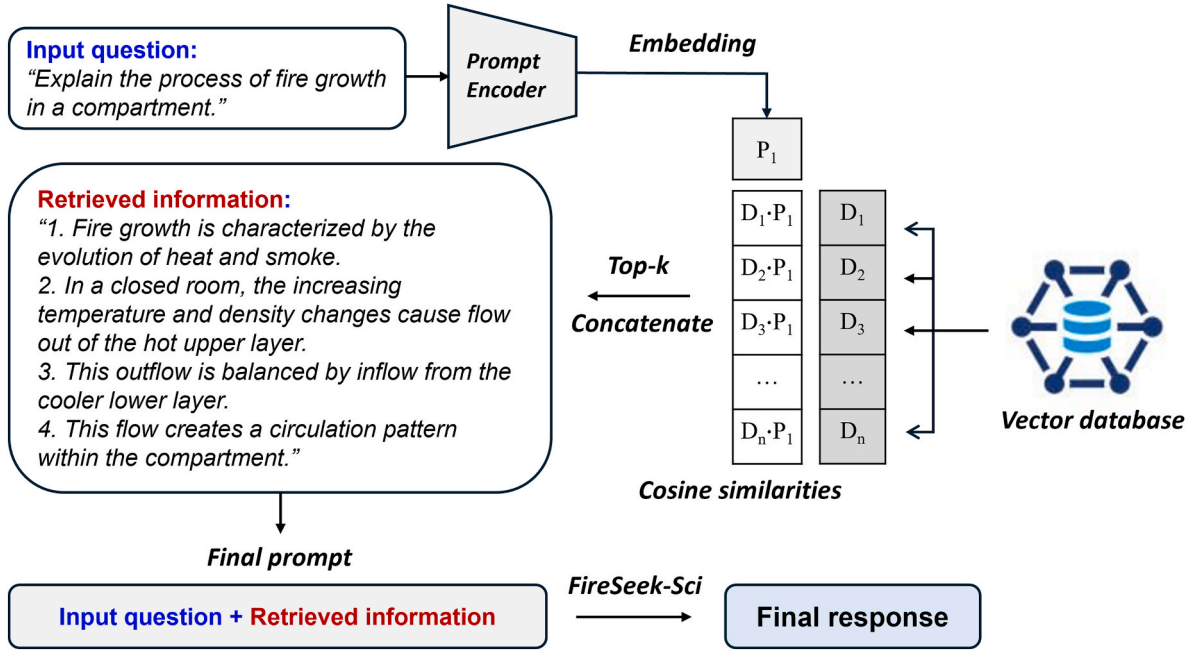


Fig. 4. The flowchart of using RAG to enhance LLM fire-related response.

2.3. Retrieval augmented generation

To improve factual accuracy and domain adaptability, we integrate Retrieval-Augmented Generation (RAG) into our system. RAG combines non-parametric information retrieval with parametric text generation, enabling the LLM to reference external knowledge sources (e.g., fire science textbooks, fire protection codes) during response generation. Unlike pure parametric generation, RAG grounds the model's outputs in retrieved real texts, thus reducing hallucinations and improving reliability on specialized fire knowledge. The flowchart of generating a fire knowledge vector database for RAG is shown as Fig. 3. To support retrieval during inference, external fire engineering references were firstly indexed in a searchable knowledge base. Depending on the source format, textual content was processed into retrievable text segments. For document-based references, textual information was extracted and divided into smaller chunks (e.g., by paragraph, sentence, or fixed character length) to facilitate semantic embedding and retrieval. Emerging machine-readable standards, such as next-generation digital codes, may further simplify this process by providing structured representations without relying on document parsing. Secondly, a transformer-based embedding model (Bge-large-en-v1.5 in this work) is used to convert each chunk into a vector embedding, and the embeddings are saved as a vector database using *LlamaIndex*. External references are used solely to support retrieval during inference. The released FireSeek knowledge base does not redistribute copyrighted documents or contain verbatim reproductions of protected source materials.

As shown in Fig. 4, given an input question prompt P , the RAG workflow consists of three steps:

3. **Embedding:** The input question P is encoded using an embedding model (Bge-large-en-v1.5 in this work) to obtain the vector representation $e(P)$.
4. **Retrieval:** The embedding $e(P)$ is compared with document embeddings $e(D) = \{e(D_1), e(D_2), \dots\}$ from the external knowledge vector database D . Cosine similarity is calculated, and the Top- k most relevant text segments are retrieved:

$$\{D_1^*, D_2^*, \dots, D_k^*\} = \text{Topk}(\cos(D_1 P_1), \cos(D_2 P_2), \dots, \cos(D_k P_1)) \quad (2)$$

where $\{D_1^*, D_2^*, \dots, D_k^*\}$ denotes the Top- k statements retrieved from the

knowledge base that are most relevant to the input question.

5. **Generation:** The retrieved text segments and the input question are concatenated into a single context block and forwarded to the LLM to generate a grounded, reference-based answer.

By grounding the generation process in authoritative fire safety literature, RAG enhances the model's factual accuracy and significantly reduces hallucination when encountering unfamiliar or specialized queries.

2.4. Website user interface

To visualize the interactive chat of the FireSeek-Sci chatbot, we designed a user-friendly website interface built upon the Gradio framework, shown as Fig. 5. The interface is composed of a conversational interaction component and a dynamic configuration management panel. The conversational module features a primary chat interface incorporating message history display to satisfy the need for continuous chatting, and integrated conversation management tools including retry and clear operations. If the users were not satisfied with the chatbot's answer, they can click the "retry" button to let the chatbot answer it again. The system implements sophisticated state management through event-driven programming functions and chaining mechanisms to ensure proper sequencing of user input processing, conversation history updates, and response generation. The user interface of FireSeek-Sci is designed to interact with users exclusively through natural-language questions. Users may submit conceptual questions related to fire engineering or questions containing contextual information describing a specific fire scenario. In the case studies presented in Section 4, the outputs of the AI-based fire and evacuation prediction models were first summarized into natural-language descriptions, which were then incorporated into user prompts. The current version of FireSeek-Sci does not directly process numerical prediction outputs, images, or other non-textual data.

The configuration management panel enables real-time model switching through a dropdown model selection (1.5B and 3B versions of the model), parameter adjustment, and optional RAG integration. The RAG functionality implements conditional visibility patterns,

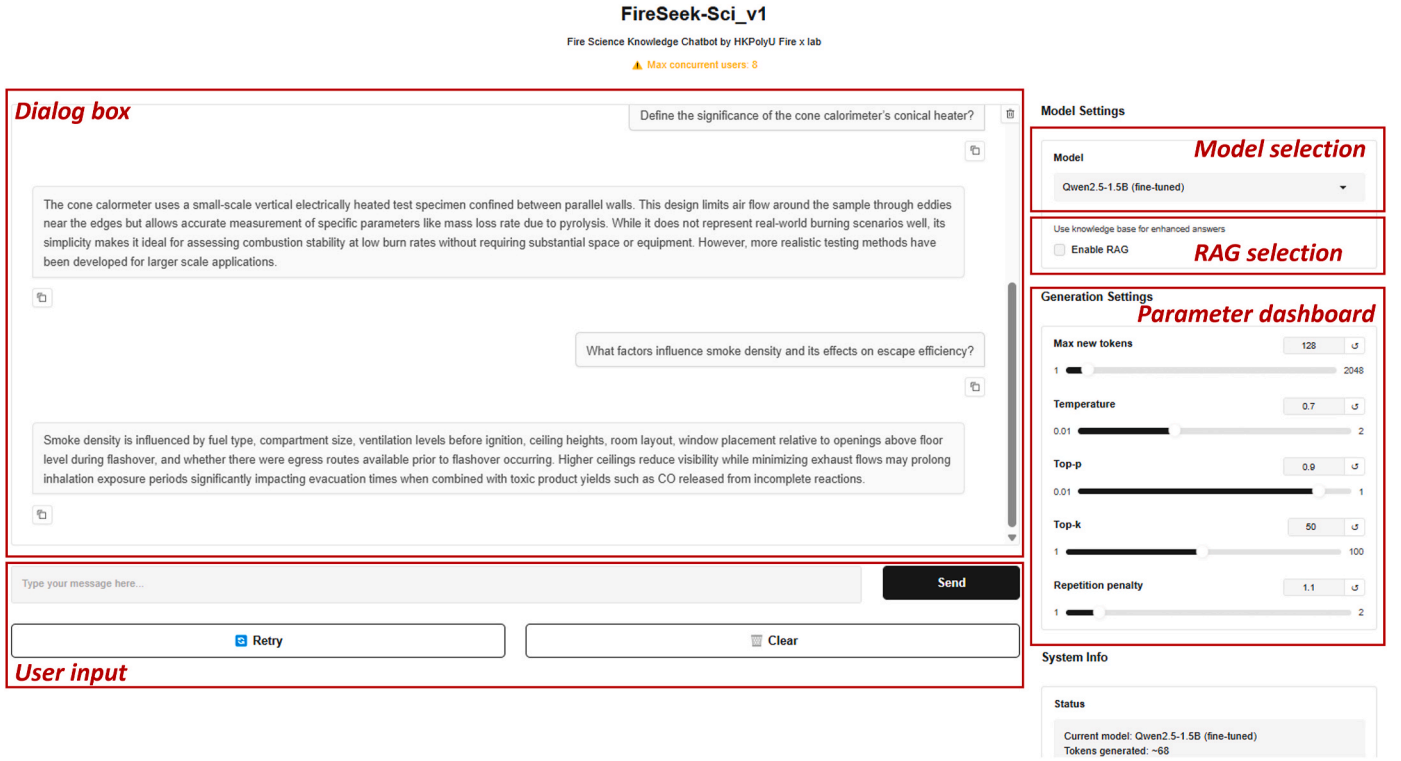


Fig. 5. The user interface of FireSeek-Sci chatbot.

dynamically revealing knowledge base selection options when activated, while maintaining interface simplicity in standard conversational mode. Parameter control panel including token generation limits (1–2048), temperature modulation (0.01–2.0), Top- p (0.01–1.0), Top- k (1–100), and repetition penalty adjustment (1.0–2.0), enabling users to fine-tune model behavior for specific fire science applications.

The practical deployment of FireSeek-Sci and the corresponding website user interface depends on the licensing conditions of the external knowledge sources integrated into the retrieval system. When proprietary fire protection standards or design codes are used, the system should be deployed in compliance with the corresponding licensing agreements and copyright requirements. Consequently, access to certain knowledge resources may be limited to users or organizations that are legally authorized to use these materials.

3. Results and demonstration

3.1. Supervised fine-tuning results

In LLM training, the total number of required training steps depends on the settings of epoch, batch size (B), samples in the training dataset (N), and gradient accumulation steps (GAS), shown as Eq. (3).

$$Total\ step = Epochs \cdot \frac{N}{B \cdot GAS} \quad (3)$$

The loss function is token-level cross-entropy. After a total of 1260 training steps, the loss value decreases to a reasonable convergence range in LLM tasks.

To evaluate the effectiveness of fine-tuning, we compared the performance of the fine-tuned FireSeek-Sci model (1.5B and 3B versions) with several pre-trained base models and fine-tuned models, including Qwen2.5, Llama3.1, and Llama3.2, on the test dataset with 1007 samples. The models were assessed using BLEU-4 and ROUGE metrics, which are commonly used to quantify the quality of generated answers. Specifically, BLEU-4 measures the overlap of 4-g sequences between generated and reference texts, while ROUGE-1, ROUGE-2, and ROUGE-L

quantify the overlap of unigrams, bigrams, and the longest common subsequences, respectively. BLEU-4 measures the precision of n -gram overlaps (up to 4-g) between the generated text and the reference, where n -gram means a sequence of n words, shown as Eq. (4). ROUGE-1 measures recall of the unigram (single word) overlaps between generated and reference text. ROUGE-1 measures the recall of bigram (two-word sequence) overlap. ROUGE-L represents recall and precision of the Longest Common Subsequence (LCS).

$$BLEU - 4 = BP \times \exp \left(\frac{1}{4} \sum_{n=4}^4 \log p_n \right) \quad (4)$$

where, p_n represents n -gram precision (how many n -grams in the generated text also appear in the reference), BP represents a brevity penalty to penalize overly short outputs.

As shown in Table 2, the FireSeek-Sci-3B model achieved substantially higher scores across all metrics, with a BLEU-4 of 41.0 compared to 5.38 for the base model. ROUGE-1, ROUGE-2, and ROUGE-L scores improved from 13.8, 2.89, and 2.89 to 37.9, 13.5, and 30.6, respectively. Similarly, the scores achieved by the FireSeek-Sci-1.5B model and other base models also improved after fine-tuning. These results indicate that fine-tuning on fire domain-specific knowledge (FDK) significantly

Table 2

Comparison of performance benchmark: FireSeek-Sci vs. Baseline Models.

Model	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
FireSeek-Sci-3B	41.0	37.9	13.5	30.6
Qwen2.5-3B-instruct	5.38	13.78	2.89	5.72
FireSeek-Sci-1.5B	39.6	36.8	12.4	29.6
Qwen2.5-1.5B-instruct	5.09	15.9	3.47	5.64
Llama3.1-8B-instruct-fine-tuned	14.65	27.62	12.07	13.39
Llama3.1-8B-instruct	6.49	17.9	4.11	6.82
Llama3.2-3B-instruct-fine-tuned	12.73	22.6	5.28	11.9
Llama3.2-3B-instruct	9.69	19.51	4.48	9.89
Llama3.2-1B-instruct-fine-tuned	12.34	22.36	4.85	11.62
Llama3.2-1B-instruct	11.45	20.7	4.69	11.2

enhances the model's ability to generate accurate and relevant answers in the fire knowledge domain. Notably, FireSeek-Sci with 1.5B parameters is superior to the 8B-parameter version of Llama3.1, demonstrating that fine-tuning with domain knowledge can enable lightweight models to significantly outperform the larger-parameter model.

To check whether the paraphrase-based data augmentation causes reported performance to be inflated by train-test leakage, we additionally trained Qwen2.5-3B-Instruct on the original (non-augmented) dataset and tested it on the corresponding test set. As shown in Table A1 in Appendix, performance differences between the augmented and original training settings were very small across all metrics (less than 0.3 points), suggesting the improvements are not mainly caused by leakage from paraphrase overlap.

To properly evaluate the positive impact of RAG on the overall performance of answer generation, it is insufficient to solely use sentence structure-level metrics (BLEU and ROUGE) matching between generated output and the reference answer in the dataset. Overlap-based metrics may fail to evaluate the semantic depth and knowledge enrichment that RAG aims to provide. Therefore, three complementary metrics, BERTScore-F1, Semantic Similarity (Semantic-Sim), and LLM-Judge Overall Score are used to assess the improvement level of the model enhanced by retrieved information. BERTScore-F1 measures token-level semantic correspondence, while Semantic Similarity evaluates sentence-level embedding relevance, reflecting the accuracy and reasonability of the meanings of generated answers in different phrasings. Additionally, a high-level LLM model with larger parameters (Qwen2.5-14B-Instruct in this work) is used as a referee to judge the performance of the generated answer and provide an overall score (0–10). The overall score follows five preset criteria and the detailed protocol is shown as Algorithm 1 in Appendix,

- Accuracy (0–10): factually correct based on fire science knowledge,
- Completeness (0–10): thoroughly answers the question,
- Relevance (0–10): directly addresses the question,
- Clarity (0–10): clear and well-structured,
- Technical depth (0–10): appropriate level of technical detail.

The comparison results of the comprehensive quality of generated answers by different models with and without RAG across three key semantic performance metrics is shown on Table 3. The results consistently demonstrate that incorporating RAG substantially enhances semantic quality and overall answer reliability across all tested models. Notably, FireSeek-Sci-1.5B improves from 0.83 to 0.87 in BERTScore-F1, and from 0.36 to 0.61 in Semantic similarity, which are both the highest scores among tested models. The LLM-overall score achieved by FireSeek-Sci also improves from 1.79 to 6.08 when RAG is enabled. These results suggest that RAG can help models to generate more accurate, informative, and professional responses based on retrieved specific knowledge.

3.2. Answer evaluation of fire science questions

In this analysis, FireSeek-Sci-1.5B was asked some fire science knowledge questions, and its answers with and without RAG were demonstrated and compared. We also tested the Qwen2.5-1.5B-Instruct model using the same questions to evaluate the difference and improvement of the chatbot's response. In the first part, we start with some general and qualitative questions. After that, we present some more professional and quantitative questions to evaluate the mathematical and physical analysis ability. The parameter of maximum new tokens is set as 512. The usage of the FireSeek-Sci interface is demonstrated as Video S1 in the Supplementary Material.

Supplementary video related to this article can be found at <https://doi.org/10.1016/j.jintel.2026.100233>

In part 1, there are three general questions as follows and the response is shown as Fig. 6,

Table 3

Comparison of semantic quality and overall performance of different models with and without RAG.

Model	Use RAG or not	BERTScore-F1	Semantic-sim	LLM-overall
FireSeek-Sci-1.5B	Yes	0.87	0.61	6.08
FireSeek-Sci-1.5B	No	0.83	0.36	1.79
Qwen2.5-1.5B-Instruct	Yes	0.81	0.17	2.54
Qwen2.5-1.5B-Instruct	No	0.80	0.01	0.84
FireSeek-Sci-3B	Yes	0.81	0.54	4.68
FireSeek-Sci-3B	No	0.81	0.52	4.57
Qwen2.5-3B-Instruct	Yes	0.81	0.15	2.30
Qwen2.5-3B-Instruct	No	0.82	0.44	3.58
Llama-3.2-1B-Instruct-finetune	Yes	0.87	0.66	5.20
Llama-3.2-1B-Instruct-finetune	No	0.85	0.51	2.49
Llama-3.2-1B-Instruct	Yes	0.85	0.67	5.78
Llama-3.2-1B-Instruct	No	0.85	0.65	5.21
Llama-3.2-3B-Instruct-finetune	Yes	0.85	0.48	4.55
Llama-3.2-3B-Instruct-finetune	No	0.83	0.20	1.36
Llama-3.2-3B-Instruct	Yes	0.84	0.63	5.03
Llama-3.2-3B-Instruct	No	0.84	0.54	3.63

- Question 1: What is fire engineering?
- Question 2: Explain the process of fire growth in a compartment.
- Question 3: What is fire evacuation?

In part 2, there are three relatively complex questions as follows and the response is shown as Fig. 7,

- Question 1: What is the Grashof number?
- Question 2: How to calculate heat flux induced by heat radiation?
- Question 3: Describe the mechanisms of fire-induced concrete spalling.

Comparisons of responses from different models are listed in Table 4, where FireSeek-Sci can accurately and rapidly answer fire science questions, including the general conception questions and the complex quantitative questions. The comparative analysis also demonstrates that integrating RAG can enhance FireSeek-Sci's answering performance to a certain extent. Without RAG, the model provides generally correct and qualitative answers, sometimes lacking professional nouns, physical explanations, and quantitative reasoning. In contrast, the RAG-enhanced responses are more structured and aligned with professional literature, showing an improved reasoning process. However, some responses using RAG still contain scientific inaccuracies, indicating that retrieval quality and domain-specific validation remain critical. Using RAG also easily produces more hallucinations and redundant nonsense sentences, which requires knowledge graph technology to achieve reasoning freedom of LLM for reducing hallucinations. Using RAG will cause a much longer response time to retrieve references from the vector database. In terms of the basic model, it responds fast, and generation is concise, but it is more suitable for non-specialized tasks due to its limited domain knowledge.

Representative failure cases of FireSeek-Sci are presented in Fig. 8 to illustrate its current limitations in fire science question answering. For the first query (Fig. 8(a)), "Explain the process of fire growth in a compartment", the response generated without RAG provided only a high-level and generic description of fire growth, describing the process simply as ignition, flame spread, and smoke filling. Such an answer lacks the standard compartment fire development stages commonly adopted in fire science, namely the incipient (starting), growth, fully developed, and decay stages, resulting in an incomplete explanation of the underlying fire dynamics. Although the RAG-enhanced model produced a more technically detailed response by incorporating knowledge from

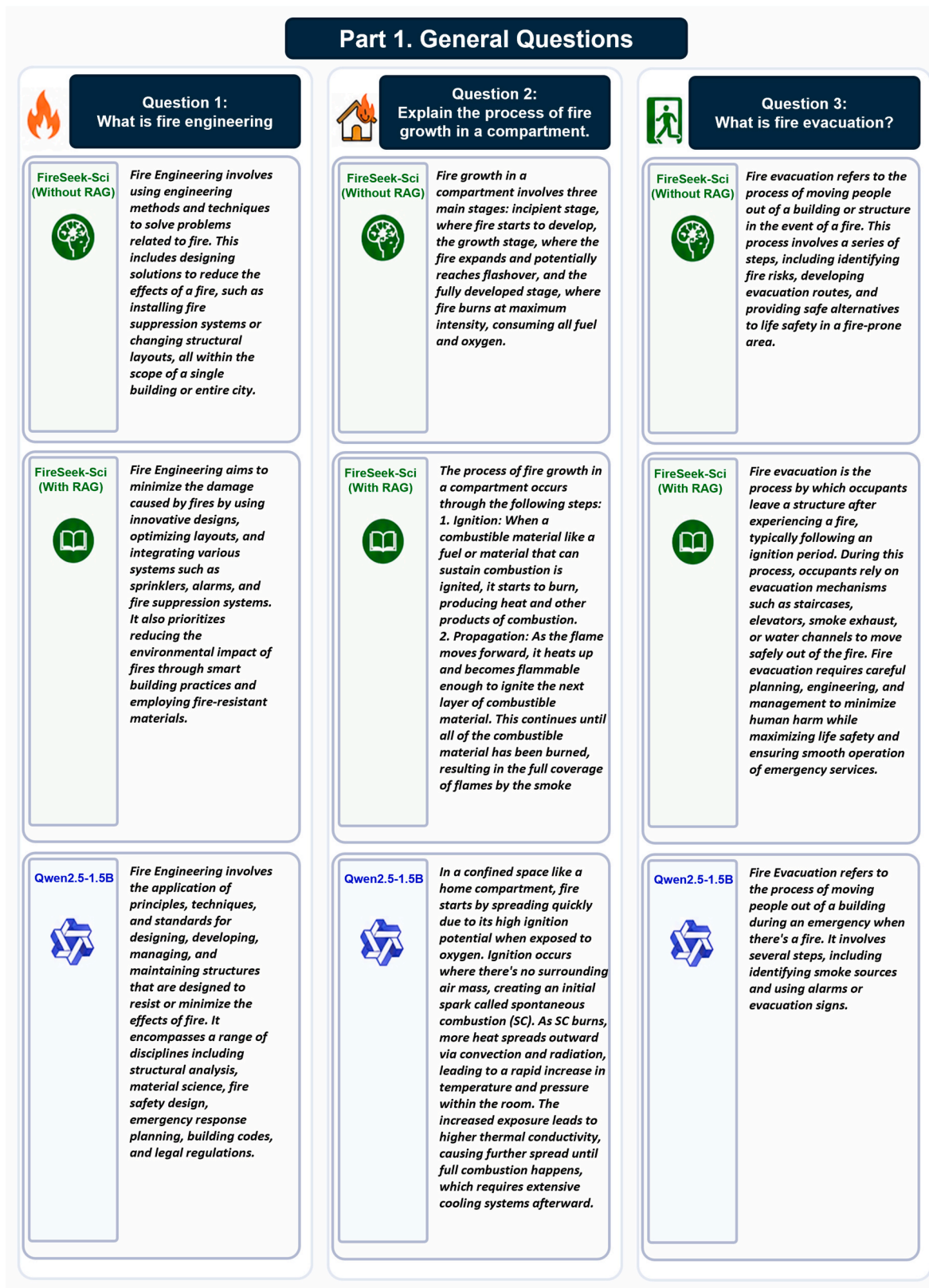


Fig. 6. FireSeek-Sci responses for general fire-related questions.

the retrieved literature, it still generated an irrelevant citation marker, highlighted in Fig. 8(a), which represents a hallucinated artifact introduced during text generation. For the second query (Fig. 8(a)), “What is the Grashof number?”, the without-RAG model produced an incorrect

mathematical expression for the Grashof number, demonstrating that the base model may generate scientifically incorrect equations when relying solely on its parametric knowledge. The RAG-enhanced response successfully corrected the equation but still contained two remaining



Fig. 7. FireSeek-Sci responses for complex fire-related questions.

Table 4
Comparison of answer responses by different methods.

	FireSeek-Sci	FireSeek-Sci (with RAG)	Qwen2.5-1.5B
Accuracy	Good	More accurate, but sometimes hallucinates	General
Response time	1–2 s	Over 10 s	1–2 s
Conciseness	Concise	Little verbose	Concise
Tokens	~110	~180	~160
Application scenario	Professional question	Professional question	General question

issues. First, it incorrectly interpreted the viscosity term as “viscosity’s power relation with temperature” rather than the correct definition of kinematic viscosity, indicating that retrieval alone cannot completely eliminate errors in scientific interpretation. Second, the response again included hallucinated residual retrieval text (shown by the strike-through annotation in Fig. 8b), suggesting imperfect integration between the retrieval and generation modules.

These representative cases demonstrate that, although retrieval augmentation substantially improves the factual correctness and completeness of FireSeek-Sci, the system may still produce overly generic explanations, scientific interpretation errors, and occasional hallucinated artifacts, particularly for technically specialized questions.

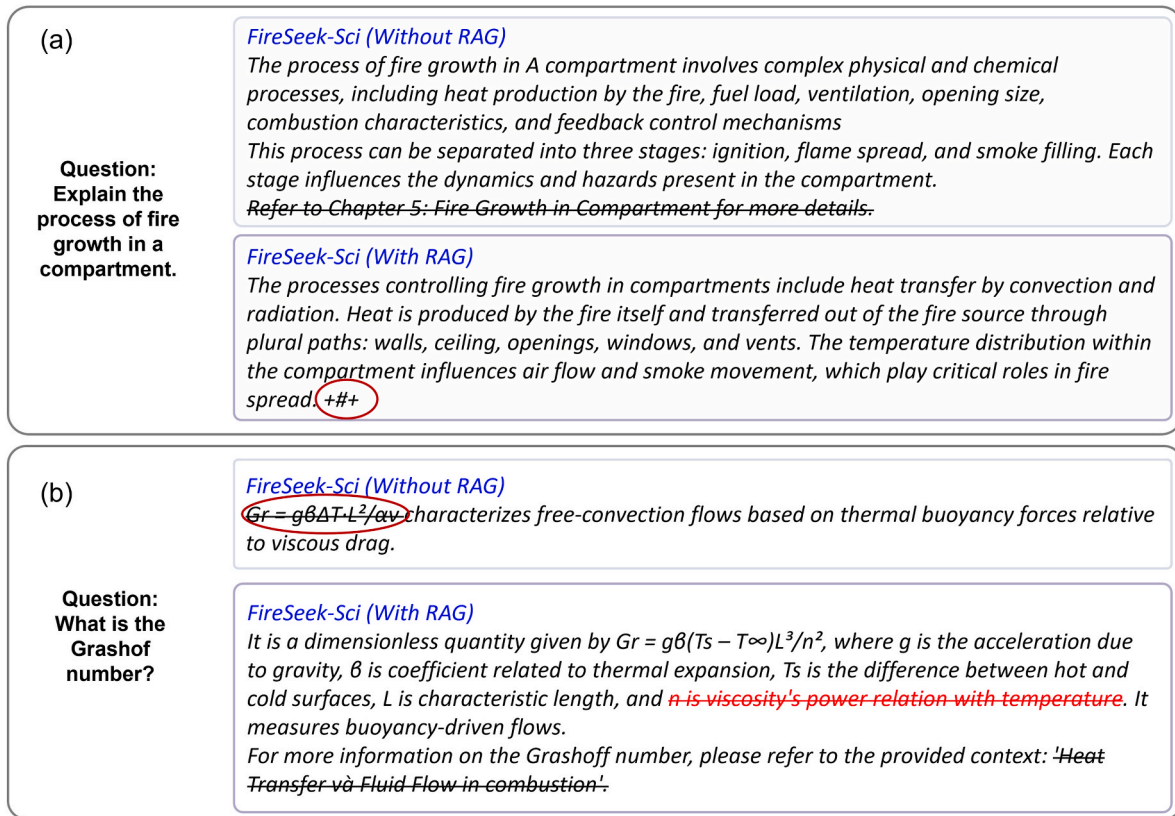


Fig. 8. Inaccurate and unsatisfied cases of FireSeek-Sci responses.

Therefore, FireSeek-Sci should currently be regarded as an assistive tool for fire science knowledge retrieval, while engineering-critical decisions should continue to rely on expert verification. The potential improvement strategies should be considered in future versions, for example, adopting the hierarchical navigable small world (HNSW) algorithm (Malkov and Yashunin, 2018) to increase search efficiency from the vector knowledge database, and transitioning to distilled embedding models to improve embedding speed. To mitigate scientific inaccuracies and hallucinations, future work should integrate a knowledge graph technology that maps verified physical laws and fire dynamics knowledge as a ground-truth constraint layer. The system will perform a dual-check where retrieved text chunks must align with the structured relationships and numerical constants stored within the knowledge graph before generation. This strategy ensures that the model's outputs remain strictly bounded by established principles and knowledge structures, effectively filtering out "redundant nonsense".

3.3. Results of the user satisfaction survey

The objective of this work is to develop a fire science chatbot to help students, researchers, and engineers to search for fire science knowledge more efficiently and accurately. The objective of the user study was to evaluate the usability, user experience, and perceived usefulness of FireSeek-Sci in practical fire engineering applications. Scientific correctness was evaluated separately using automatic metrics and LLM-overall assessment. Therefore, the users' satisfaction feedback is critical for us to revise and update the further versions of FireSeek-Sci. In this work, we invited over 20 fire engineering-related participants involved in the user satisfaction survey including experiencing the website user interface of FireSeek-Sci and filled in the satisfaction questionnaire. The questions in our satisfaction survey are listed in Table S1 in the Appendix.

A total of 24 participants were invited to take part in the internal testing and satisfaction survey of FireSeek-Sci. Each participant

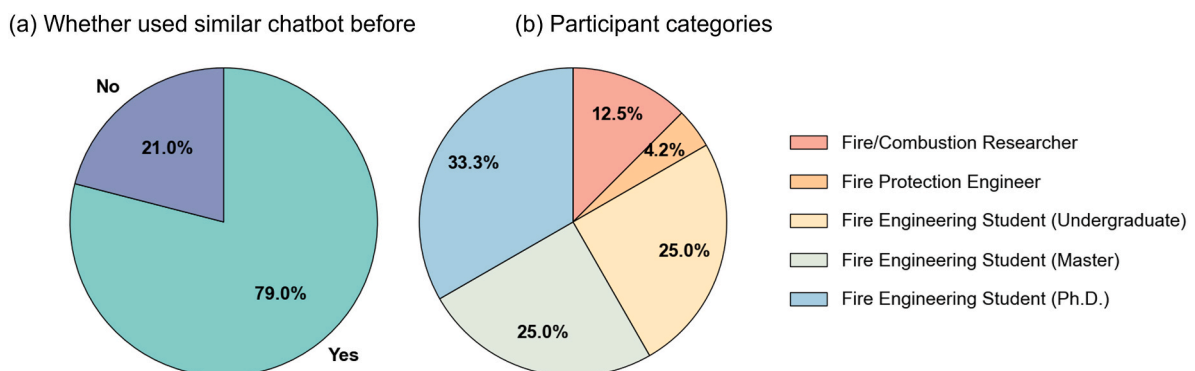


Fig. 9. Backgrounds of the survey participants.

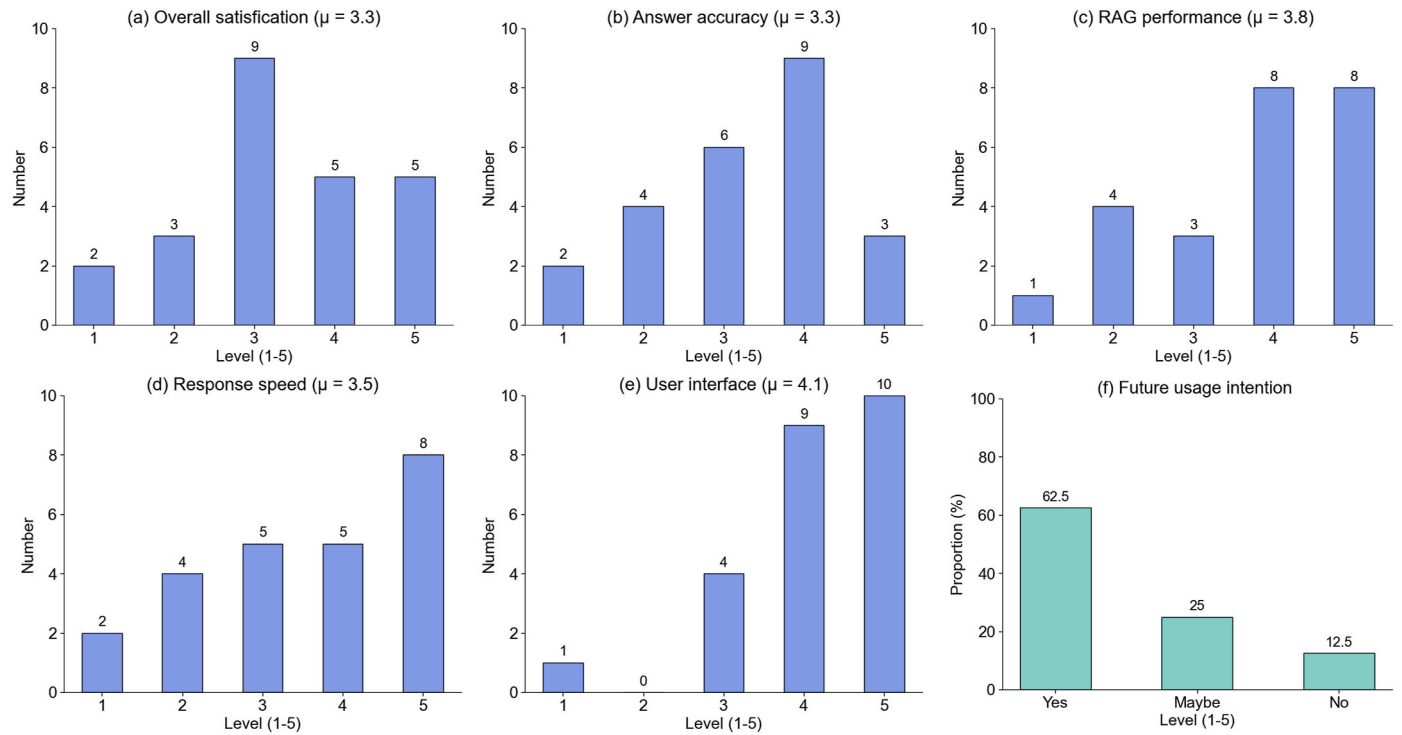


Fig. 10. Results of user satisfaction, (a) Overall satisfaction, (b) Answer accuracy, (c) RAG performance, (d) Response speed, (e) User interface, (f) Future usage intention.

interacted with the FireSeek-Sci chatbot by asking fire science-related questions, both with and without the RAG function enabled, and then evaluated the accuracy and response speed of the answer generation. After using the chatbot, they were required to fill in a satisfaction questionnaire to assess the performance of the chatbot. The whole evaluation process was unsupervised and anonymous. Most participants (83.3 %) were fire engineering students at different academic levels, including undergraduate (25 %), master's (25 %), and Ph.D. (33.3 %) students, as shown in Fig. 9a. The remaining participants consisted of three scientific researchers specializing in fire or combustion fields and one fire protection engineer. In addition, participants' prior experience with similar chatbots was investigated. As illustrated in Figs. 9(b), 79 % had previously used similar chatbots, while 21 % were unfamiliar with this type of system.

The results of the satisfaction questionnaire for the internal testing of FireSeek-Sci are shown in Fig. 10, covering six aspects: (a) overall satisfaction, (b) answer accuracy, (c) RAG performance, (d) response speed, (e) user interface, and (f) future usage intention. As shown in Fig. 10(a), the overall satisfaction with FireSeek-Sci reached a mean score of 3.3 (on a 1–5 scale), indicating an acceptable but moderate user experience. The answer accuracy evaluation in Fig. 10(b) reached a similar mean score of 3.3, suggesting that FireSeek-Sci was able to provide partially useful responses for fire science-related queries. However, this score also indicates that the generated answers do not yet consistently achieve a high level of perceived accuracy, particularly for complex questions requiring specialized domain reasoning. In Fig. 10(c), the RAG performance achieved a relatively higher mean score ($\mu = 3.8$), demonstrating that participants believed that the RAG function is beneficial in enhancing the relevance and depth of generated answers. In Fig. 10(d), the response speed was rated with a mean of 3.5, reflecting generally satisfactory performance with acceptable latency for most users. The user interface (UI) received the highest evaluation ($\mu = 4.1$), shown in Fig. 10(e), indicating that participants approved the interface as intuitive and user-friendly. Finally, Fig. 10(f) shows that 62.5 % of participants expressed a clear intention to use FireSeek-Sci and its further versions in the future, while 25 % indicated possible future use

and 12.5 % reported no intention of reuse. Overall, the results suggest that FireSeek-Sci shows promising potential as an intelligent assistant for fire science knowledge support, with acceptable usability and user experience. However, the evaluation results should be interpreted cautiously, as the moderate satisfaction and answer accuracy scores indicate that the system still has limitations in achieving highly reliable scientific responses. The user feedback suggests that the overall experience may be affected by multiple factors, including response latency during real-time interactions and residual inaccuracies or hallucinations in complex quantitative queries. These limitations highlight the need for further improvements in retrieval efficiency, knowledge coverage, and scientific reasoning capability. Future work will focus on optimizing retrieval strategies, incorporating knowledge graphs and expert-verified knowledge sources, and developing more robust mechanisms to improve factual consistency and reliability in fire science applications.

4. Case studies of smart building fire safety applications

To further illustrate the potential applicability of FireSeek-Sci beyond educational question answering, this section presents two representative case studies in which FireSeek-Sci is introduced as a domain-specialized large language model capable of converting fire dynamic and pedestrian evacuation outputs and to generate natural-language safety-oriented summaries. The case studies are intended as qualitative demonstrations of the system's ability to incorporate quantitative simulation outputs into its interpretation and to produce structured, hazard-oriented natural-language descriptions.

4.1. Case 1: corridor fire prediction

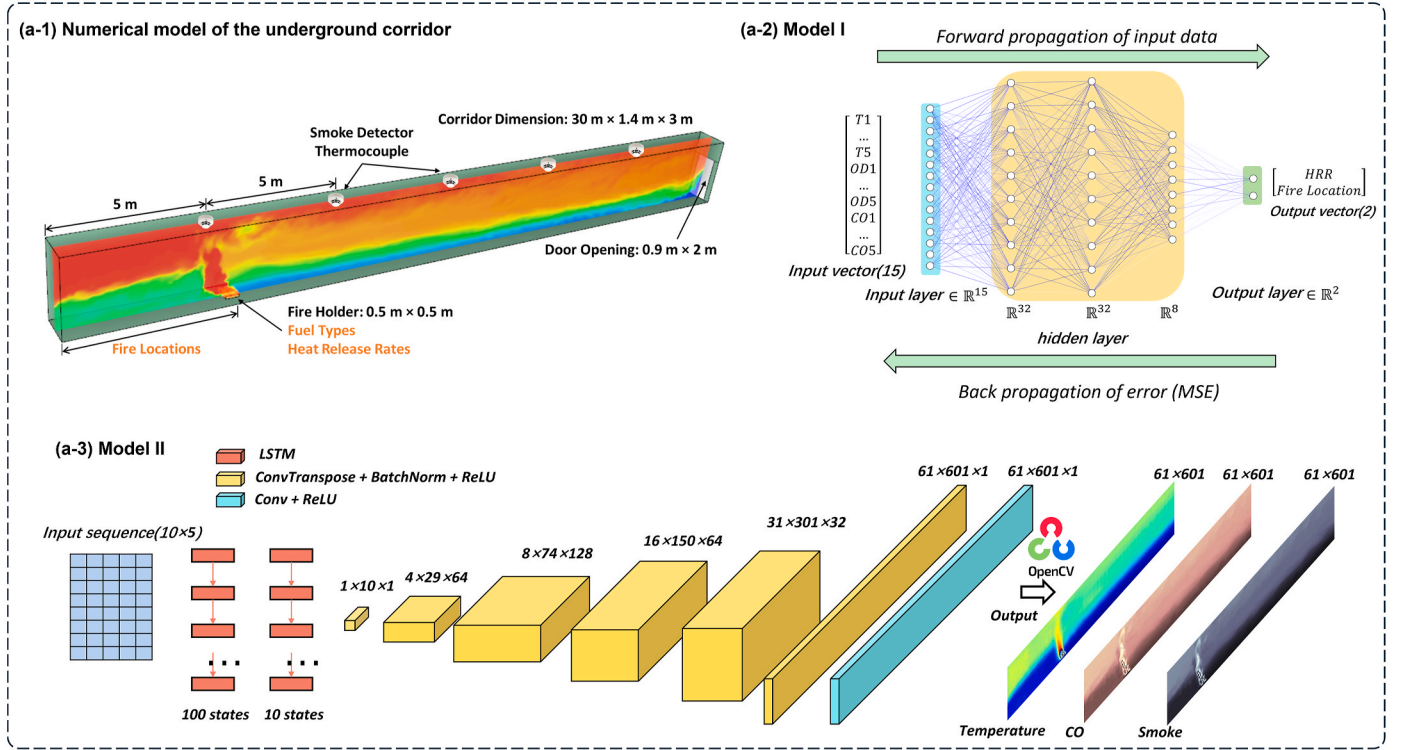
The first case study demonstrates the ability of FireSeek-Sci to interpret AI-predicted fire dynamics in a ventilation-limited corridor (Figs. 11) and transform numerical outputs into actionable fire safety intelligence based on our previous work (Cheung et al., 2025). As illustrated in Fig. 11, the input to FireSeek-Sci was derived from the outputs of an AI-based fire prediction framework (Fig. 11), which

estimates the fire source location, heat release rate (HRR), and spatio-temporal distributions of temperature, smoke density, and CO concentration from multi-sensor measurements. These prediction results were subsequently converted into structured natural-language prompts and provided to FireSeek-Sci for interpretation and engineering decision support. Detailed descriptions of the prediction model are available in

Ref. (Cheung et al., 2025).

In the demonstration (Fig. 11(b)), the input prompts for FireSeek-Sci are extracted from the AI estimated key fire parameters at a representative fire stage ($t = 120$ s), including a HRR value of approximately 520 kW, a fire source location at 12 m, a ceiling temperature of 118°C , a smoke density of 0.38 g/m^3 , and a CO concentration of 460 ppm. While

(a) Underground corridor fire predictions



(b) FireSeek-Sci interpretation

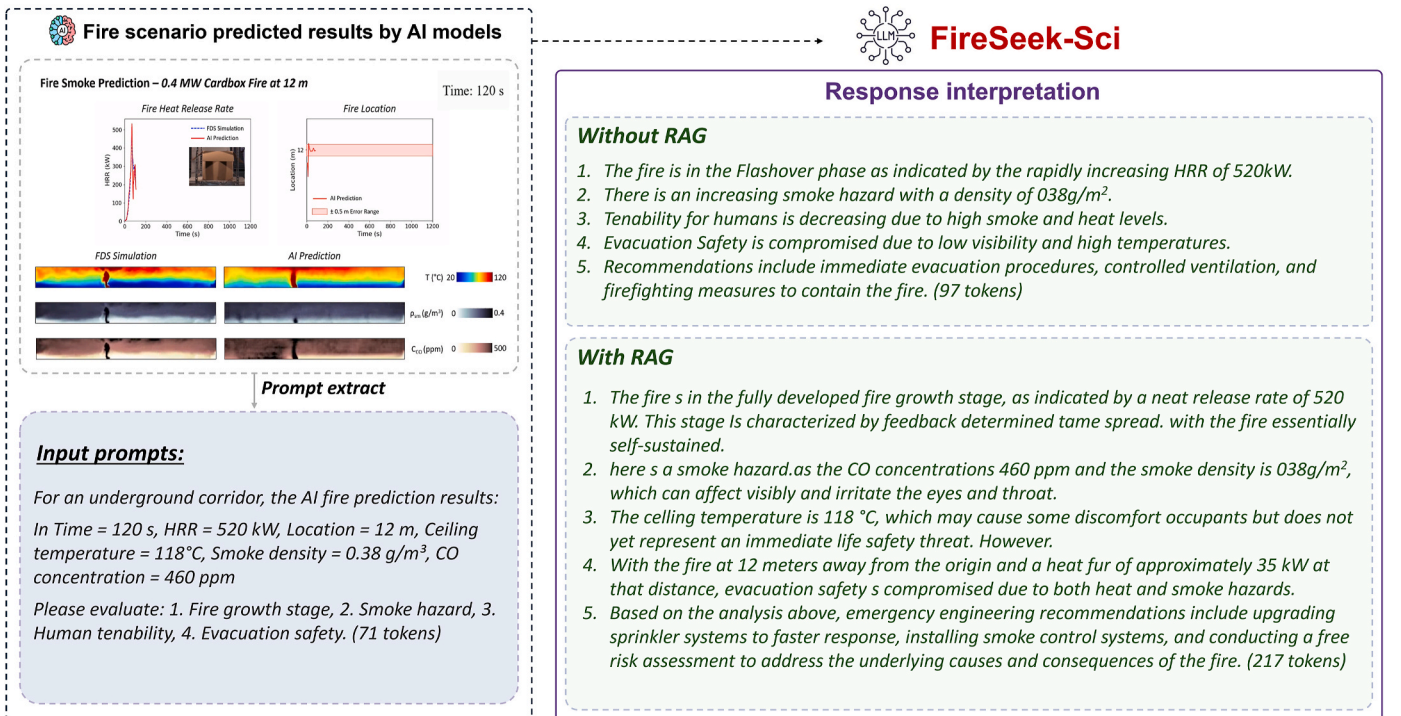


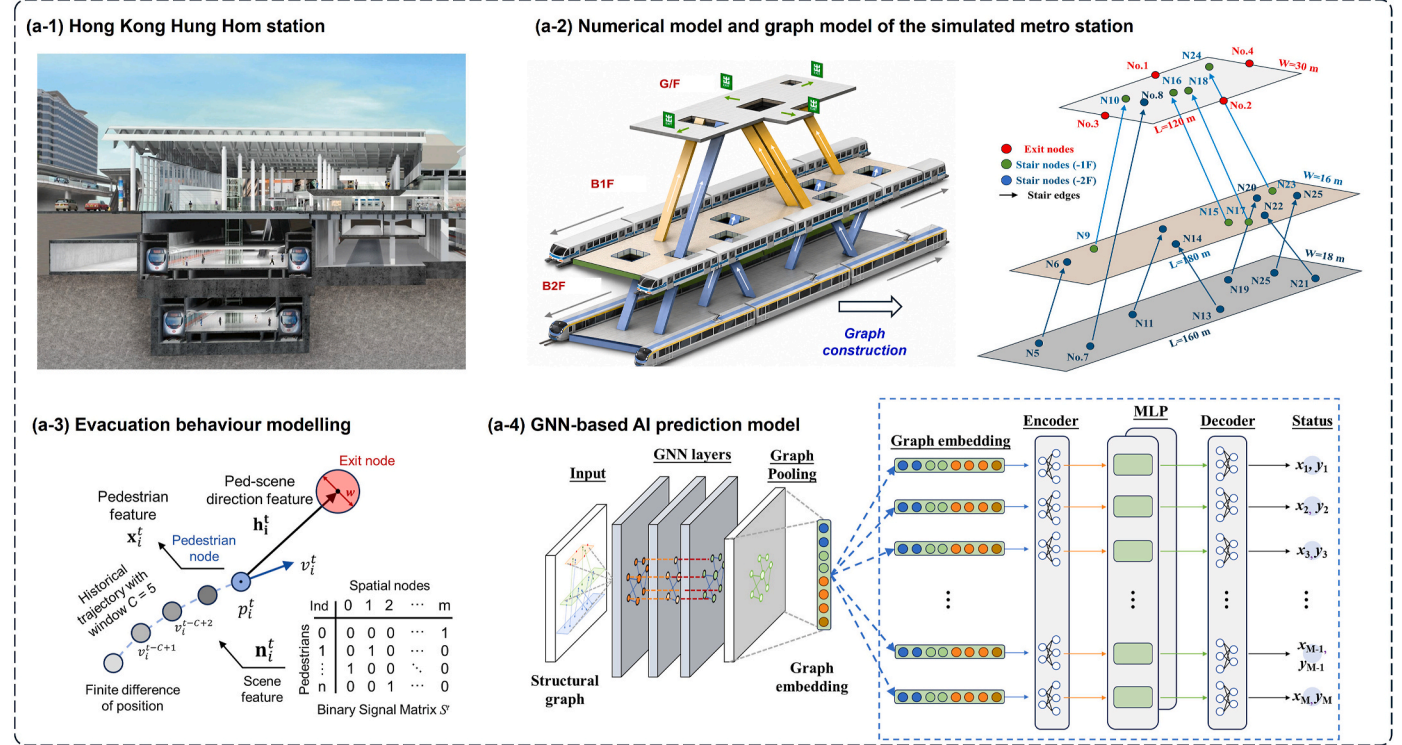
Fig. 11. Workflow of the corridor fire case, (a) numerical corridor scenario generation and AI-based fire prediction, and (b) prompt construction, and engineering interpretation by FireSeek-Sci with and without retrieval augmentation.

these parameters provide a quantitative description of fire development, they do not directly communicate the associated safety implications to firefighters or emergency decision-makers. The purpose of FireSeek-Sci in this scenario was therefore to interpret these physical predictions and convert them into engineering-oriented natural-language assessments. Therefore, the input prompts contain the parameter information and

propose to provide evaluative interpretations of the fire development and safety considerations, i.e., 1. *Fire growth stage*, 2. *Smoke hazard*, 3. *Human tenability*, and 4. *Evacuation safety*.

Without RAG, FireSeek-Sci was able to identify several major hazards based on its internal fire science knowledge. The model successfully recognized that the fire was in an advanced growth stage and identified

(a) Underground corridor fire predictions



(b) FireSeek-Sci interpretation

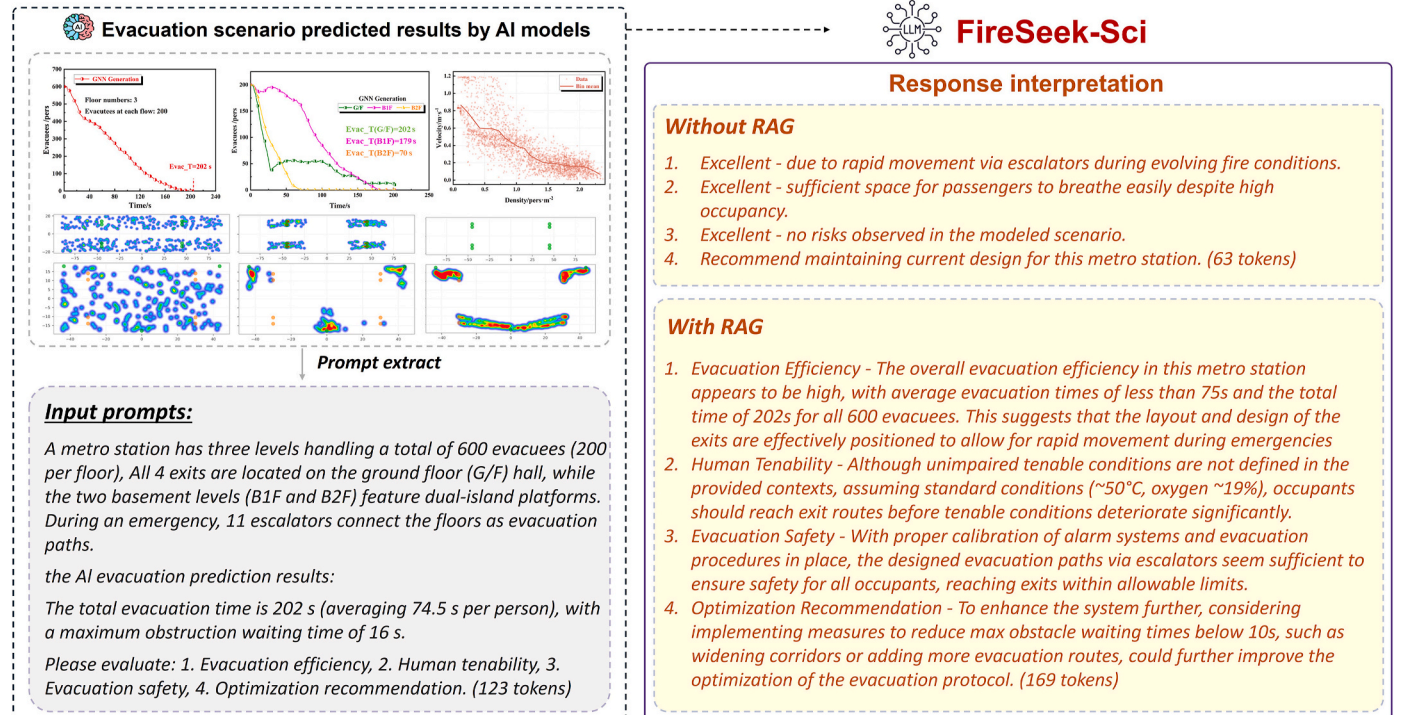


Fig. 12. Workflow of the metro station evacuation prediction case, (a) numerical model of simulated metro station and AI-based evacuation prediction model, and (b) prompt construction, and engineering interpretation by FireSeek-Sci with and without retrieval augmentation.

the increasing risks associated with smoke accumulation, elevated temperature, and reduced visibility. It also correctly associated these hazards with deteriorating occupant tenability and compromised evacuation safety. These results indicate that the domain-specific fine-tuning of FireSeek-Sci provides a strong baseline capability for fire scenario interpretation. However, the interpretation occasionally included generalized terminology, such as classifying the fire as approaching flashover conditions, without explicitly relating the conclusion to the predicted numerical values. Some recommendations were qualitatively reasonable but lacked direct quantitative justification. This suggests that although the model possesses fire science reasoning capability, its responses may still be influenced by general language priors or incomplete recall of engineering thresholds.

When RAG was introduced, the quality and engineering reliability of the response improved significantly. FireSeek-Sci with RAG directly incorporated the predicted numerical parameters into its interpretation, explicitly referencing the elevated CO concentration, smoke density, and ceiling temperature as key indicators of occupant hazard. The model provided a more structured hazard analysis, including fire development stage, smoke toxicity, thermal exposure, human tenability, and evacuation feasibility. More importantly, its engineering recommendations became substantially more practical, including immediate evacuation activation, smoke control enhancement, and suppression system intervention. The comparison also confirms that retrieval augmentation substantially improves the grounding, consistency, and engineering applicability of large language model responses in complex infrastructure fire scenarios.

4.2. Case 2: metro station evacuation prediction

The second case study evaluates the capability of FireSeek-Sci in interpreting AI-predicted evacuation performance for a complex metro station environment (Fig. 12). The evacuation predictions were generated using a previously developed graph neural network (GNN)-based evacuation prediction framework, shown as Fig. 12. The model provides key evacuation performance indicators, including evacuation trajectories, congestion information, and evacuation efficiency. In the presented scenario, the predicted results included a total evacuation time of 202 s, an average evacuation time of 74.5 s per occupant, and a maximum congestion waiting time of 16 s. These prediction outputs were transformed into structured prompts and supplied to FireSeek-Sci for engineering interpretation and decision support. These numerical outputs were subsequently transformed into structured natural-language prompts and provided to FireSeek-Sci for engineering interpretation, i.e.,

“The AI evacuation prediction results: The total evacuation time is 202 s (averaging 74.5 s per person), with a maximum obstruction waiting time of 16 s. Please evaluate: 1. Evacuation efficiency, 2. Human tenability, 3. Evacuation safety, 4. Optimization recommendation.”

Shown as Fig. 12(b), FireSeek-Sci was able to produce a general qualitative assessment of evacuation performance without RAG. The model recognized the positive impact of escalator-based vertical circulation and concluded that the overall evacuation performance was “excellent”. It also suggested that the station layout provided sufficient safe space and no severe fire evacuation hazards. While these responses were overly optimistic and intuitively reasonable, they lacked direct references to the predicted evacuation metrics. More importantly, the model failed to identify potential congestion bottlenecks or discuss the implications of the observed waiting time.

After introducing RAG, the response quality improved significantly. FireSeek-Sci explicitly incorporated the evacuation prediction metric into its reasoning process, directly interpreting the total evacuation time of 202 s and the average evacuation time of 74.5 s as indicators of relatively high evacuation efficiency. In addition, the model identified the maximum waiting time of 16 s as a potential congestion indicator and proposed optimization strategies, including corridor widening,

route redistribution, and evacuation path enhancement to reduce bottleneck delays. Compared with the non-RAG response, the retrieval-enhanced model demonstrated stronger quantitative interpretations, better engineering consistency, and more actionable recommendations.

4.3. Discussion and perspective

Above all, the case studies of corridor fire and metro station evacuation interpretations illustrate that local-deployed FireSeek-Sci is capable of translating AI-based prediction results outputs into structured, natural-language safety-oriented summaries, which qualitatively resemble the type of interpretation an engineer might produce. This suggests a potential role for such a system in helping engineers, infrastructure operators and firefighters more quickly organize and communicate multiple evolving physical variables and evacuation indicators under time-critical conditions. However, these case studies are illustrative demonstrations rather than a validation study. The correctness, completeness, and engineering reliability of the specific hazard interpretations and recommendations generated in these scenarios were not independently verified by domain experts, and the two cases are not sufficient to establish general reliability across the broader range of fire and evacuation scenarios that may arise in practice. The rigorous validation of answer correctness (e.g., via blinded expert evaluation against standardized case questions) remains an important direction for future work.

Moreover, another significant advantage of FireSeek-Sci is money-saving. Table 5 summarizes the token consumptions of each response modes and equivalent commercial API cost. For the two representative safety engineering tasks, the average input prompt length was 97 tokens. In the non-RAG mode, FireSeek-Sci generated relatively concise responses, with an average output of 80 tokens. After RAG was enabled, the response length increased to an average of 219 tokens, reflecting the model's tendency to produce more detailed and evidence-supported engineering analysis. However, the main source of token growth did not come from the generated response itself, but from the retrieved contextual knowledge injected into the prompt. In the two case studies, the retrieved fire engineering references contributed approximately 800–1000 additional tokens per query. To provide a more intuitive engineering deployment comparison, the token usage was further converted into equivalent cloud-based API costs based on a typical commercial pricing scheme and normalized to 100 user queries.

As a result, the corresponding estimated API cost of non-RAG mode is \$0.18 per 100 queries. RAG mode increased the total token consumption to an average of 1215 tokens per interaction, approximately 6.9 times higher than the non-RAG mode, with an equivalent API cost of about \$1.22 per 100 queries. Because FireSeek-Sci is deployed locally using lightweight foundation models, these additional tokens do not result in recurring API charges. Instead, the increased token consumption mainly affects local computational resources such as GPU inference time and memory usage. This deployment strategy offers a practical advantage for high-frequency engineering applications, where continuous cloud-based API could become a significant long-term operational cost.

Table 5
Comparison of token consumptions and equivalent commercial API cost.

	Case 1	Case 2	Average	Equivalent API cost/100 queries (USD)
Input	71	123	97	0.09
Non-RAG response	97	63	80	0.08
RAG response	217	220	219	0.89
Retrieved context injection	~800	~1000	~1000	0.22
Total (Non-RAG)	168	186	177	0.18
Total (RAG)	1088	1343	1215	1.22

5. Conclusions

This work developed an LLM-powered fire science knowledge rapid retrieval and response chatbot named FireSeek-Sci. The backbone LLM is fine-tuned from the Qwen2.5 model using a fire domain-knowledge database. FireSeek-Sci improves by over 30 points in multiple accuracy metrics in the generated corpus compared with the basic Qwen model. The response accuracy and richness of chatbot using RAG technology achieved significant improvements in semantic similarity (increased 69 %) and advanced LLM scoring (increased over 4 points on a 0–10 scale), respectively. FireSeek-Sci without RAG provides fast and concise responses with good accuracy, making it well-suited for professional and technical questions compared to the basic Qwen model. However, it may lack the deeper contextual understanding and knowledge coverage that RAG can offer.

To further evaluate its potential applicability beyond educational question answering, two fire safety engineering case studies were conducted in this work. In the first case, FireSeek-Sci was used to interpret AI-predicted fire dynamics in a ventilation-limited corridor. In the second case, FireSeek-Sci was applied to interpret AI-based evacuation predictions in a multi-level metro station. In both cases, FireSeek-Sci was able to transform raw simulation outputs into understandable engineering assessments, covering fire development, smoke hazards, human tenability, evacuation safety, and optimization recommendations. A noticeable difference was observed between the two inference modes that FireSeek-Sci without RAG tended to generate broad qualitative judgments, while retrieval augmentation enabled more evidence-based reasoning closer to how fire engineers typically interpret numerical predictions in practice.

Overall, FireSeek-Sci validates the feasibility of developing a fire-domain large language model from a general-purpose foundation model, through domain-specific fine-tuning, and further toward an engineering-oriented retrieval-enhanced reasoning system. The proposed framework not only demonstrates strong potential for fire education and rapid knowledge learning but also shows promising

capability in fire safety assessment and emergency decision support. Future work will emphasize expanding the fire domain database with more fire codes, standards, and real case studies, incorporating knowledge graph-based retrieval constraints to reduce hallucinations, and enabling contextual multi-turn reasoning for complex fire scenario analysis. Future efforts will gradually transform FireSeek-Sci from a simple engineering educational prototype into a reliable and powerful engine for next-generation intelligent fire research and engineering support.

CRedit authorship contribution statement

Yifei Ding: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Hongtao Zheng:** Writing – review & editing, Software, Investigation, Formal analysis. **Tong Lu:** Methodology, Data curation, Conceptualization. **Yanfu Zeng:** Writing – review & editing, Data curation, Conceptualization. **Rong Deng:** Writing – review & editing, Formal analysis, Conceptualization. **Yuxin Zhang:** Writing – original draft, Supervision, Methodology, Data curation, Conceptualization. **Xinyan Huang:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by National Key R&D Program of China (Grant No. 2024YFE0216700), Hong Kong Innovation and Technology Commission (Grant Nos. MHP/018/24, ITP/013/25LP), and PolyU Teaching and Learning Grant (Grant No. 9BNH).

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.iintel.2026.100233>.

Appendix

Table A1
Performance differences between the augmented and original dataset.

Model	Dataset	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
Qwen2.5-3B-instrcut	Augmented	41.03	37.88	13.47	30.61
Qwen2.5-3B-instrcut	Original	40.75	37.79	13.36	30.60

Table A2
Questions listed in our satisfaction survey questionnaire.

Question	Fill
Part 1. Basic Information	
1. What is your profession or background?	Researcher/Student/Engineer
2. Have you used similar Q&A chatbot before?	Yes/No
Part 2. Function satisfaction	
1. How satisfied are you with FireSeek-Sci overall?	Very dissatisfied (1) – Very satisfied (5)
2. Would you like to continue using this software in the future?	Definitely/No/Maybe
3. How accurate are the basic Q&A answers?	Very not accurate (1) – Very accurate (5)
4. Do you agree that external knowledge retrieval (RAG) improves answer quality?	Very disagree (1) – Very agree (5)
5. How satisfied are you with the system's speed/performance?	Very dissatisfied (1) – Very satisfied (5)
6. How user-friendly is the interface design and interaction?	Very dissatisfied (1) – Very satisfied (5)

Algorithm 1. LLM-overall evaluation for answer quality

Algorithm 1 LLM-as-a-Judge Evaluation for Answer Quality

Require: Test set $D = \{(q_i, r_i)\}_{i=1}^N$; generation model M ; judge model J ; retrieval index I (optional)

Ensure: Per-instance scores $S = \{s_1, \dots, s_N\}$; aggregate score $\bar{S}$

```

1: for  $i = 1$  to  $N$  do
2:    $q_i, r_i \leftarrow D[i]$ 
3:   if RAG enabled then
4:      $c_i \leftarrow \text{Retrieve}(I, q_i, \text{top\_k} = 3)$ 
5:      $a_i \leftarrow \text{Generate}(M, q_i, c_i)$ 
6:   else
7:      $a_i \leftarrow \text{Generate}(M, q_i)$ 
8:   end if
9:    $p_i \leftarrow \text{BuildJudgePrompt}(q_i, r_i, a_i, c_i)$ 
10:   $\text{resp}_i \leftarrow J(p_i)$ 
11:  if  $\text{resp}_i$  is valid JSON then
12:    (Acc, Comp, Rel, Clar, Tech)  $\leftarrow \text{ParseScores}(\text{resp}_i)$ 
13:  else
14:    (Acc, Comp, Rel, Clar, Tech)  $\leftarrow (5.0, 5.0, 5.0, 5.0, 5.0)$   $\triangleright$  fallback
15:  end if
16:   $s_i \leftarrow \frac{\text{Acc} + \text{Comp} + \text{Rel} + \text{Clar} + \text{Tech}}{5}$ 
17:   $S \leftarrow S \cup \{s_i\}$ 
18: end for
19:  $\bar{S} \leftarrow \frac{1}{N} \sum_{i=1}^N s_i$ 
20: return  $S, \bar{S}$ 

```

To complement the automatic lexical and semantic metrics, we employed an LLM-as-a-Judge protocol to assess the qualitative correctness of generated answers. For each test instance, the judge model (GPT-4o-mini or a locally deployed Qwen2.5-14B-Instruct model) was presented with the original question, the reference (gold-standard) answer, and the candidate answer produced by the evaluated system, along with the top retrieved contexts when retrieval-augmented generation (RAG) was enabled. The judge was instructed, through a structured prompt, to rate the candidate answer along five dimensions on a 0–10 scale (shown as Eq. (A1)): (1) Accuracy, the factual correctness of the response relative to established fire-science knowledge; (2) Completeness, the extent to which the answer addressed all relevant aspects of the question; (3) Relevance, the degree to which the response directly answered the question asked; (4) Clarity, the readability and coherence of the explanation; and (5) Technical Depth, the appropriateness of the level of domain-specific detail. The judge returned its assessment as a structured JSON object containing the five sub-scores together with a brief textual justification. An overall score was then computed as the unweighted arithmetic mean of the five sub-scores.

$$\text{Overall} = \frac{1}{5} (\text{Accuracy} + \text{Completeness} + \text{Relevance} + \text{Clarity} + \text{Technical Depth}) \quad (\text{A1})$$

References

- Bromann, M., 2001. The Design and Layout of Fire Sprinkler Systems, 2nd ed. CRC Press, Boca Raton. <https://doi.org/10.1201/9781420031911>.
- Buchanan, A.H., Abu, A.K., 2016. Structural Design for Fire Safety, 2nd ed. John Wiley & Sons Ltd, Hoboken. <https://doi.org/10.1002/9781118700402>.
- Chen, B., Wu, S., Jin, Y., Song, Y., Wu, C., Venevsky, S., Xu, B., Webster, C., Gong, P., 2024. Wildfire risk for global wildland–urban interface areas. *Nat. Sustain.* 7 (4), 474–484. <https://doi.org/10.1038/s41893-024-01291-0>.
- Chen, D., Chen, L., Yang, Y., Lu, Q., Hancock, C., Lock, R., Sölvsten, S., 2026. A knowledge-driven approach for automated fire safety compliance checking in operational buildings. *Knowl. Based Syst.* 336, 115156. <https://doi.org/10.1016/j.knsys.2025.115156>.
- Cheung, W.K., Zeng, Y., Ding, Y., Węgrzyński, W., Huang, X., 2025. Predicting smoke hazards and burning fuel via smart building fire sensor network and dual-agent deep learning. *Indoor Built Environ.* 34 (6), 1107–1125. <https://doi.org/10.1177/1420326X251331180>.
- Cui, J., Zhai, C., Wang, Y., Li, Y., 2026. Large language models and AI agents in disaster-resilient infrastructure: concepts, applications, pathways, and challenges. *Reliab. Eng. Syst. Saf.* 272, 112497. <https://doi.org/10.1016/j.res.2026.112497>.
- Deng, R., Cheung, W.K., Ding, Y., Huang, X., Usmani, A., 2026. Agent-human multimodal interactive framework for inspection and maintenance of infrastructure fire safety. *J. Ind. Inf. Integr.* 53, 101184. <https://doi.org/10.1016/j.jii.2026.101184>.
- Dong, Y., Zhan, Z., Hu, Y., Doe, D.M., Han, Z., 2025. AI BIM coordinator for non-expert interaction in building design using LLM-driven multi-agent systems. *Autom. Constr.* 180, 106563. <https://doi.org/10.1016/j.autcon.2025.106563>.
- Farquhar, S., Kossen, J., Kuhn, L., Gal, Y., 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630 (8017), 625–630. <https://doi.org/10.1038/s41586-024-07421-0>.
- Hostetter, H., Naser, M.Z., Huang, X., Gales, J., 2024. The role of large language models (AI chatbots) in fire engineering: an examination of technical questions against domain knowledge. *Nat. Hazards Res.* 4 (4), 669–688. <https://doi.org/10.1016/j.nhres.2024.06.003>.
- Huang, X., Lin, S., Chen, Y., Qin, Y., Zhang, Y., Ding, Y., 2026. A review of interactive smoldering-flaming fires at the wildland-urban interface. *Proc. Combust. Inst.* 42, 106298. <https://doi.org/10.1016/j.proci.2026.106298>.
- Huang, X., Tam, W.C., 2024. Intelligent Building Fire Safety and Smart Firefighting. Springer, Cham. <https://doi.org/10.1007/978-3-031-48161-1>.
- Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., Fung, P., 2023. Towards mitigating LLM hallucination via self reflection. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, pp. 1827–1843. <https://doi.org/10.18653/v1/2023.findings-emnlp.123>.
- Jin, Y., 2025. Bridging the knowledge-skill gap: the role of large language model and critical thinking in education. *Comput. Educ.* 235, 105357. <https://doi.org/10.1016/j.compedu.2025.105357>.
- Kampelopoulou, D., Tsanousa, A., Vrochidis, S., Kompatsiaris, I., 2025. A review of LLMs and their applications in the architecture, engineering and construction industry. *Artif. Intell. Rev.* 58 (8), 250. <https://doi.org/10.1007/s10462-025-11241-7>.
- Lu, T., Ding, Y., Zhang, Y., Xie, W., Huang, X., 2026a. A review of artificial intelligence in pedestrian evacuation safety: applications, challenges, and future LLM-XAI ecosystems. *Adv. Eng. Inform.* 76, 105080. <https://doi.org/10.1016/j.aei.2026.105080>.
- Lu, T., Zhang, Y., Xie, W., Huang, X., 2026b. Human-AI interactive framework for smart evacuation safety analysis in large infrastructures. *Reliab. Eng. Syst. Saf.* 266, 111695. <https://doi.org/10.1016/j.res.2025.111695>.

- Ma, H., Lu, Y., Xiao, Z., Feng, J., Zhang, H., Yu, J., 2025. SDD-LawLLM: advancing intelligent legal systems through synthetic data-driven fine-tuning of large language models. *Electronics* 14 (4), 742. <https://doi.org/10.3390/electronics14040742>.
- Malkov, Y.A., Yashunin, D.A., 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42 (4), 824–836. <https://doi.org/10.1109/tpami.2018.2889473>.
- McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., Weinschenk, C., Overhold, K., 2021. *Fire dynamics simulator user's guide*. In: NIST Special Publication 1019, 6th ed. National Institute of Standards and Technology (NIST), Gaithersburg.
- Merci, B., Beji, T., 2022. *Fluid Mechanics Aspects of Fire and Smoke Dynamics in Enclosures*, 2nd ed. CRC press, London <https://doi.org/10.1201/9781003204374>.
- Mettrop, W., Nieuwenhuysen, P., 2001. Internet search engines—fluctuations in document accessibility. *J. Doc.* 57 (5), 623–651. <https://doi.org/10.1108/eum000000007096>.
- Nolan, D.P., 2011. *Fire Fighting Pumping Systems at Industrial Facilities*, 2nd ed. William Andrew, Norwich <https://doi.org/10.1016/C2010-0-66226-2>.
- Quintiere, J.G., 2006. *Fundamentals of Fire Phenomena*. John Wiley & Sons, Ltd, Hoboken. <https://doi.org/10.1002/0470091150>.
- Ramesh, M., Sun, Z., Li, Y., Zhang, L., Annam, S.K., Fang, H., Tong, D., 2025. Assessing WildfireGPT: a comparative analysis of AI models for quantitative wildfire spread prediction. *Nat. Hazards* 121 (11), 13117–13130. <https://doi.org/10.1007/s11069-025-07344-7>.
- Society of Fire Protection Engineers, 2016. *SFPE Handbook of Fire Protection Engineering*, 6th ed. Springer, Cham <https://doi.org/10.1007/978-3-031-59212-6>.
- Tao, J., Wang, P., Jiang, H., Han, Y., 2026. A knowledge-formalization and vision-language-model system for fire emergency decision support. *Adv. Eng. Inform.* 71, 104409. <https://doi.org/10.1016/j.aei.2026.104409>.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T. B., 2023. Alpaca: a strong, replicable instruction-following model [dataset]. Stanford Center for Research on Foundation Models. <https://crfm.stanford.edu/2023/03/13/alpaca.html#3>.
- Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. *Nat. Med.* 29, 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>.
- Wang, Z., Zhang, T., Wu, X., Huang, X., 2022. Predicting transient building fire based on external smoke images and deep learning. *J. Build. Eng.* 47, 103823. <https://doi.org/10.1016/j.jobe.2021.103823>.
- Xie, W., Zeng, Y., Zhang, X., Wong, H.Y., Zhang, T., Wang, Z., Wu, X., Shi, J., Huang, X., Xiao, F., Usmani, A., 2025. AIoT-powered building digital twin for smart firefighting and super real-time fire forecast. *Adv. Eng. Inform.* 65, 103117. <https://doi.org/10.1016/j.aei.2025.103117>.
- Xie, W., Zhang, Y., Lu, T., Huang, X., Shi, J., Huang, X., Xiao, F., Usmani, A., 2026. Integrating smart fire forecast with LLM-powered emergency response. *Engineering* 63, 252–267. <https://doi.org/10.1016/j.eng.2026.02.023>.
- Xiong, H., Bian, J., Li, Y., Li, X., Du, M., Wang, S., Yin, D., Helal, S., 2024. When search engine services meet large language models: visions and challenges. *IEEE Trans. Serv. Comput.* 17 (6), 4558–4577. <https://doi.org/10.1109/tsc.2024.3451185>.
- Zeng, Y., Liu, X., Ding, Y., Zheng, Z., Zhang, T., Huang, X., Lu, X., 2025. AI-powered automatic design of fire sprinkler layout for random building floorplans. *J. Infrastruct. Intell. Resil.* 4 (4), 100167. <https://doi.org/10.1016/j.iintel.2025.100167>.
- Zhang, X., Jiang, Y., Wu, X., Nan, Z., Jiang, Y., Shi, J., Zhang, Y., Huang, X., Huang, G.G. Q., 2024. AIoT-enabled digital twin system for smart tunnel fire safety management. *Dev. Built Environ.* 18, 100381. <https://doi.org/10.1016/j.dibe.2024.100381>.
- Zhang, X., Wu, X., Park, Y., Zhang, T., Huang, X., Xiao, F., Usmani, A., 2021. Perspectives of big experimental database and artificial intelligence in tunnel fire research. *Tunn. Undergr. Space Technol.* 108, 103691. <https://doi.org/10.1016/j.tust.2020.103691>.
- Zhang, Y., Zhang, X., Huang, X., 2023. Design a safe firefighting time (SFT) for major fire disaster emergency response. *Int. J. Disaster Risk Reduct.* 88, 103606. <https://doi.org/10.1016/j.ijdrr.2023.103606>.
- Zheng, H., Wang, M., Wang, Z., Huang, X., 2024. FireDM: a weakly-supervised approach for massive generation of multi-scale and multi-scene fire segmentation datasets. *Knowl. Based Syst.* 290, 111547. <https://doi.org/10.1016/j.knosys.2024.111547>.