JPhys Photonics

PURPOSE-LED PUBLISHING™

TECHNICAL NOTE • OPEN ACCESS

OptoChat: a large language model with retrieval augmented generation for optics

To cite this article: Xiaoqing Bao *et al* 2026 *J. Phys. Photonics* **8** 026001

Journal of Physics: Photonics

OPEN ACCESS

RECEIVED
7 November 2025

REVISED
27 April 2026

ACCEPTED FOR PUBLICATION
27 May 2026

PUBLISHED
22 June 2026

TECHNICAL NOTE

OptoChat: a large language model with retrieval augmented generation for optics

Xiaoqing Bao[1,2,5], Hairuo Wang[2,5], Silin Chen[3,5], Yali Zhang[3,5], Wenjun Chen[3], Kangjian Di[3], Mengcheng Lv[2], Minghui Zhao[4], Guohao Wang[4], Wenzheng Zhao[4] and Ningmu Zou[3,*]

[1] School of Integrated Circuits, Nanjing University of Information Science and Technology, Nanjing, People's Republic of China
[2] Nanzhi Institute of Advanced Optoelectronic Integration Technology, Nanjing, People's Republic of China
[3] School of Integrated Circuits, Nanjing University, Suzhou, People's Republic of China
[4] ZetaTech Co., Ltd, Shanghai, People's Republic of China
[5] These authors contributed equally to this study.
[*] Author to whom any correspondence should be addressed.

E-mail: nzou@nju.edu.cn

Keywords: large language model, retrieval augmented generation, optical knowledge question answering

Abstract

Large language models (LLMs) show strong performance in general text generation and knowledge-based question answering (QA). However, a substantial performance gap remains in optics, a knowledge-intensive scientific field. General-purpose LLMs, trained and tuned without sufficient optics knowledge, often hallucinate when handling optics-related queries. We present OptoChat, a retrieval-augmented LLM purpose-built for the optical domain. By combining domain-specific retrieval-augmented generation with targeted optimization, OptoChat delivers precise, contextually grounded answers for optics QA. To the best of our knowledge, this is the first work to adapt and optimize LLMs specifically for optics knowledge understanding and QA. Experimental results show that OptoChat achieves state-of-the-art performance on optics-focused QA benchmarks and significantly outperforms recent general-purpose models. The interactive website is available at https://optochatai.com.

1. Introduction

The rapid progress of artificial intelligence (AI) has enabled numerous applications in computational imaging, microscopy automation, and optical design [1–4]. Among AI approaches, large language models (LLMs) are particularly well-suited to optics-related tasks that require knowledge representation and reasoning. LLMs have achieved notable advances in natural language processing, especially in question answering (QA) [5], and have shown broad applicability in specialized domains [6–9]. However, in the field of optics, insufficient incorporation of domain-specific knowledge during pre-training makes LLMs prone to hallucinations and inaccuracies, and the field's knowledge-intensive, scientifically rigorous nature further challenges effective pre-training and fine-tuning.

Recently, retrieval-augmented generation (RAG) has emerged as a promising approach to mitigate the limitations of LLMs by retrieving pertinent information from external knowledge sources and incorporating it into the input in a non-parametric manner [10]. As illustrated in figure 1, external knowledge documents are first processed through a series of preprocessing steps and then encoded into vector representations using an embedding model. These vector representations are stored in a knowledge vector database (VectorDB). When a user submits a query, it is similarly transformed into a vector using the same embedding model and used to retrieve the most relevant knowledge entries from the database. The retrieved documents are then incorporated as auxiliary context, and together with the reformulated user query, form a prompt that is fed into the LLM. The model generates a response based on both the retrieved knowledge and the user query. RAG effectively alleviates the knowledge limitations and hallucination tendencies of general-purpose LLMs, and has demonstrated strong performance across a range of specialized application domains [10–12].

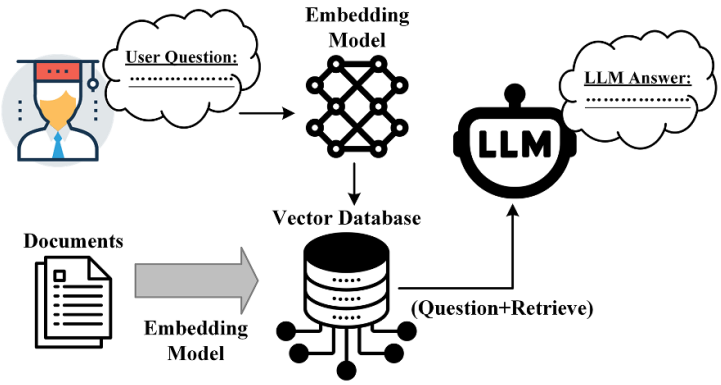

Figure 1. Illustration of the retrieval augmented generation to connect LLMs to the external sources.

However, applying existing RAG methods to the optics domain poses significant challenges due to the intricate physical principles, specialized terminology, and complex scientific reasoning inherent in optics-related documents and queries [13–16].

In this article, we propose a novel retrieval-augmented LLM specifically designed for the optical domain, named OptoChat. To effectively capture user intent in complex, context-rich scientific queries, OptoChat introduces a novel question preprocessing framework that decomposes intricate scientific questions into multiple logically coherent sub-queries and enriches them to enhance semantic completeness. To support robust information retrieval, OptoChat constructs a comprehensive knowledge vector repository from 236 982 articles collected from scientific journals and authoritative texts in optics. This multimodal corpus is systematically parsed, segmented, and indexed to form the foundation of OptoChat's domain-specific knowledge base. To further enhance retrieval performance, OptoChat employs a multi-path retrieval mechanism that enables more diverse and comprehensive evidence retrieval. Additionally, a dedicated Re-rank Agent is introduced to refine the retrieved results and identify the most contextually relevant information.

2. Method

An overview of the OptoChat framework is illustrated in figure 2. In the initialization stage, OptoChat constructs a comprehensive repository of scientific knowledge by encoding a curated corpus of scientific literature and filtered public knowledge bases into high-dimensional vector representations. Upon receiving a user query, OptoChat employs an LLM-driven Query Agent to analyze the original question, extract salient keywords, and decompose it into retrieval-oriented sub-queries. This process is designed to identify the user's intent and convert complex scientific inquiries into a structured representation suitable for retrieval. Subsequently, the generated sub-queries are then embedded into a vector space using a dedicated embedding model. The resulting query vector is then utilized to retrieve relevant documents through a hybrid retrieval approach combining both keyword-based and semantic similarity matching against the VectorDB. The retrieved candidate contexts are scored and re-ordered by a dedicated Re-rank agent, optimizing relevance to the user's intent. Finally, the top-ranked contexts are structured according to a predefined prompt template shown in figure 3 and passed to an LLM to generate a coherent and informative response.

As illustrated in figure 4, each document is processed through a dedicated parsing pipeline. Specifically, a YOLOv10-based layout detector [17] is first applied to identify key components within the document, including textual regions, mathematical formulas, tables, and images. This layout detection is followed by secondary segmentation of text regions into structural elements such as the abstract, main body, references, and appendices. For text-containing regions, contextual content is extracted using the PaddleOCR [18] optical character recognition engine. Table contents are independently detected and parsed using a YOLOv8-based table detector [19]. Finally, the extracted elements (including textual segments, tables, and figures) are hierarchically merged into a unified textual representation according to their spatial and logical organization, forming the final input to the embedding pipeline.

The processed text is then segmented using a recursive character splitter, which balances chunk-size constraints with semantic coherence. The splitter recursively divides text using line breaks, spaces, and

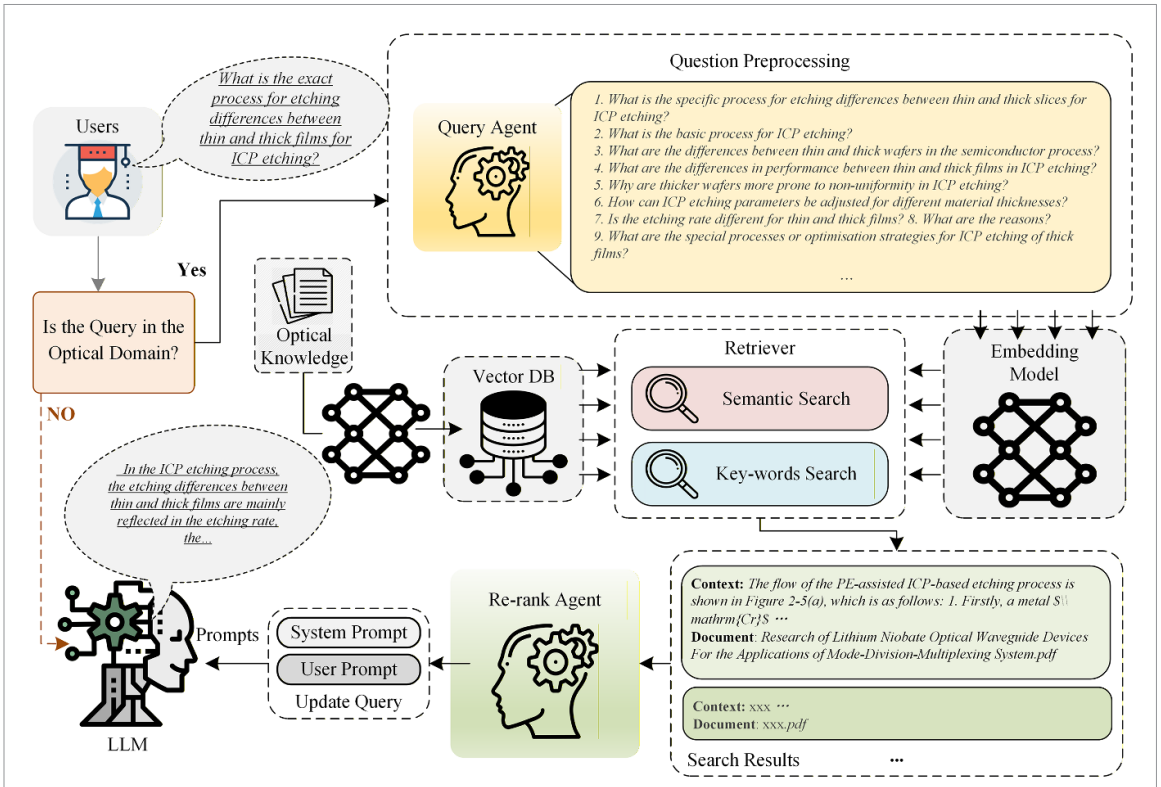

Figure 2. Illustration of OptoChat, a retrieval-augmented LLM specifically designed for the optical domain.

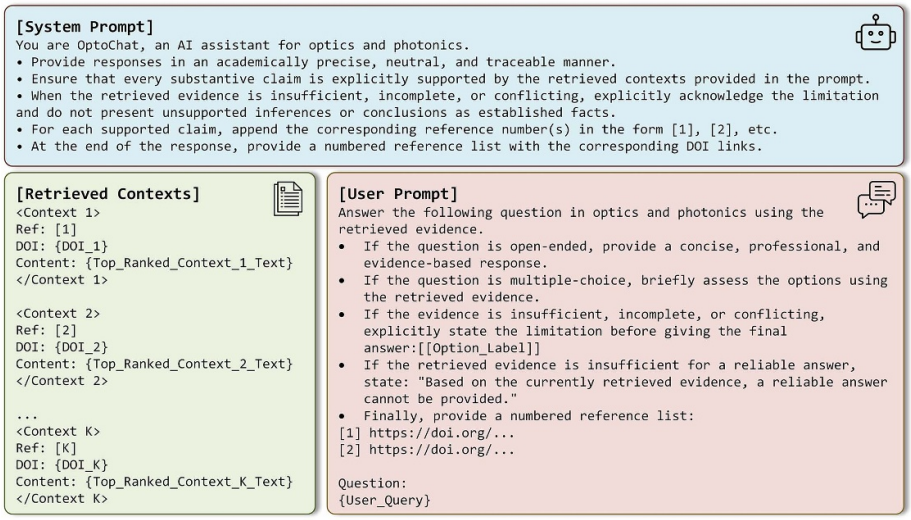

Figure 3. Prompt template used in the generation stage of OptoChat.

special characters, while ensuring contextual overlap between adjacent chunks to preserve semantic continuity. These text chunks are encoded using the bge-m3[6] embedding model, with embeddings indexed and stored in a VectorDB according to their corresponding document identifiers.

2.1. Knowledge base construction and updating

OptoChat maintains a domain-specific knowledge base to support retrieval in the optics domain. This knowledge base is constructed from optics-related scholarly resources. The collected literature is organized into eight optics-related subfields based on the title, abstract, and keywords of each document. This subfield taxonomy is used for corpus organization and subsequent statistical analysis. Specifically,

[6] https://huggingface.co/BAAI/bge-m3.

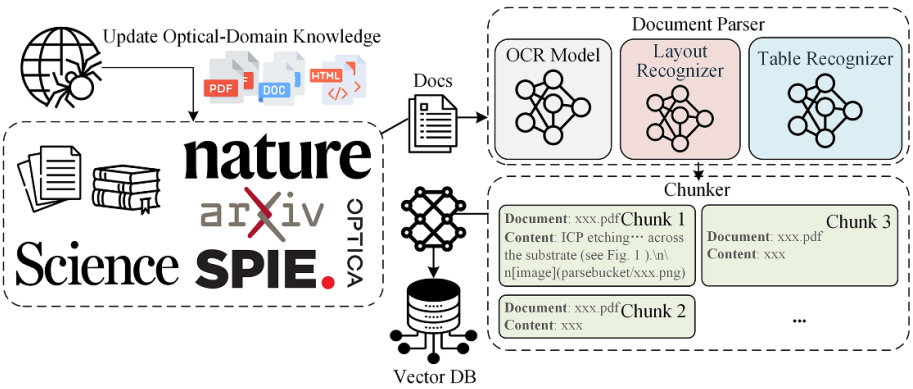

Figure 4. Illustration of the construction of the VectorDB.

the eight subfields are topological photonics (Topo. photonics); computational optics, modeling, and inverse design (Comp. optics); quantum optics and quantum photonics (Quantum optics); optical devices, integrated photonics, and measurement (Opt. devices); wave optics and electromagnetic theory (Wave optics/EM); metasurfaces, nanophotonics, and phase engineering (Metasurfaces/nano.); general and interdisciplinary optics literature (General/interdisc.); and others (Others).

As illustrated in figure 4, the collected resources are processed through an indexing pipeline that includes text normalization, content segmentation, chunk construction, and vectorization. The resulting representations are stored in the VectorDB for subsequent retrieval. To keep the knowledge base up to date, OptoChat adopts an incremental updating strategy rather than rebuilding the entire database from scratch. During each update cycle, newly added and modified records are synchronized with the database, and duplicate entries are removed. Through this procedure, the knowledge base remains current and suitable for retrieval in the optics domain.

2.2. Question preprocessing

Conventional RAG systems typically perform retrieval directly from raw user queries. Due to the knowledge-intensive and complex nature of the optics domain, raw-query matching may not fully capture the physical meaning and reasoning structure contained in user questions. To address this issue, OptoChat adopts an agent-based approach for the preprocessing of user queries involving specialized domain knowledge. Specifically, the Query Agent is built on Qwen2.5-32B-Instruct [20] and is configured through a task-specific system prompt. The preprocessing stage begins with domain discrimination. If the query is assessed to fall outside the scope of optical knowledge, it is forwarded directly to a general-purpose LLM for response generation. However, if the query is identified as pertaining to the optical domain, it undergoes further processing by the Query Agent. In this case, the Query Agent is configured to function as a complex logical query decomposer. It decomposes the original query into up to ten logically related sub-queries while preserving contextual and logical dependencies. It then generates a retrieval-oriented representation composed of keywords and sub-queries. The keywords are used for keyword-based retrieval, whereas the sub-queries are used for semantic retrieval, as illustrated in figure 2. This structured decomposition improves retrieval coverage and evidence alignment for downstream generation by aligning query components with relevant knowledge segments in the optical domain. Unlike the Re-rank Agent, which operates after retrieval to score candidate contexts, the Query Agent operates before retrieval and is responsible for domain discrimination and query decomposition.

2.3. Retriever and re-ranker

The OptoChat retriever employs a hybrid retrieval strategy, integrating both keyword-based and semantic retrieval mechanisms to enhance the accuracy of optical knowledge QA. For keyword-based retrieval, the Query Agent initially extracts salient keywords from user queries. These keywords are then used to search the knowledge base. Specifically, a term frequency-inverse document frequency (TF-IDF) [21] method is applied for initial keyword matching, complemented by the BM25 algorithm [22] for robust sparse retrieval. The keyword-based retrieval process can be formulated as follows:

$$\text{Result}_{\text{keyword}} = \underset{d \in D}{\text{search}} \left(d, \text{Agent}\left([\text{user\_prompt}] \right) \right), \tag{1}$$

where $\text{Result}_{\text{keyword}}$ represents the retrieval results based on keyword-based retrieval, D is the knowledge base, d denotes the documents in the knowledge base, Agent is the Query Agent, and search(˙) is the TF-IDF and BM25 algorithm. The combined application of TF-IDF and BM25 identifies the most relevant knowledge chunks based on the lexical similarity between query keywords and the knowledge base content. While keyword-based retrieval offers rapid and computationally efficient access to relevant information, it has a limited ability to capture deeper semantic relationships. To address these semantic limitations, OptoChat further incorporates semantic retrieval. This process begins by encoding the sub-queries (generated by the Query Agent) into high-dimensional vectors using the bge-m3 embedding model. These query vectors are projected into the same semantic space as the VectorDB. Within this semantic space, similarity metrics are used to measure the relevance between the query vectors and the stored knowledge chunk vectors. The top-K most semantically relevant chunks are then selected as retrieval results. This hybrid retrieval approach, synergizing sparse (keyword-based) and dense (semantic vector-based) methodologies, effectively balances retrieval efficiency with accuracy, thereby significantly augmenting the overall knowledge retrieval capabilities of the OptoChat system.

The initial candidate retrieval results, generated by the Query Agent and the hybrid retrieval methods, often contain documents that are weakly relevant or entirely irrelevant to the user query. This can arise from imperfections in VectorDB construction or inaccuracies in Query Agent processing. As demonstrated in [23], even a small number of weakly relevant documents can lead RAG-based LLMs to produce erroneous responses. To mitigate this issue, OptoChat incorporates a dedicated Re-rank Agent designed to filter out weakly relevant and irrelevant information. Specifically, OptoChat leverages bge-reranker-v2-m3[7] as its Re-rank Agent. This agent takes the user query and each candidate document as an input pair. The Re-rank Agent then directly computes a relevance score for each query-document pair using a cross-attention mechanism. The candidate documents are then ranked in descending order according to their relevance scores. The top-K contexts are selected as the final retrieval results for generation and organized using the predefined prompt template before being passed to the LLM. This re-ranking step improves the quality and reliability of the evidence provided to the LLM. Here, K denotes the number of top-ranked contexts used in the generation stage.

3. Results and discussions

To evaluate OptoChat's performance, we established a specialized evaluation benchmark within the optics domain. Specifically, we randomly sampled more than 1000 source documents according to the subfield taxonomy defined in the OptoChat database, thereby reducing the risk that benchmark construction would be concentrated on only a limited range of topics. Subsequently, an LLM was employed to generate an initial set of question-answer pairs, which were then validated by domain experts. This review by more than 10 researchers helped ensure the accuracy and relevance of the generated content, resulting in the final evaluation benchmark. The benchmark comprises two distinct components: OptoQA, consisting of 580 question-answer pairs on optics, and OptoMCQA, comprising 252 multiple-choice questions.

To further characterize both the knowledge corpus and the source documents used to construct the evaluation benchmark, we analyzed a stratified statistical sample of 5000 documents and a benchmark source sample of 1000 documents under the subfield taxonomy defined in the OptoChat database. As shown in figure 5(a), both sampled subsets are broadly consistent with the full database in terms of subfield composition. In particular, the benchmark source sample preserves the main disciplinary coverage of the full corpus, suggesting that the evaluation benchmark was constructed from documents spanning the major optics-related subfields rather than from a narrowly concentrated subset. The source distribution of the statistical sample is shown in figure 5(b), indicating that the knowledge corpus includes both authoritative optics journals and arXiv resources. The t-SNE visualization of the partial chunk embeddings is presented in figure 5(c), where related subfields exhibit meaningful local clustering in the embedding space. In addition, figure 5(d) shows the citation-count distribution of the statistical sample, indicating that the corpus covers multiple citation levels. A relatively high proportion of low-citation or uncited records is observed. This is partly attributable to the inclusion of arXiv documents and recently indexed records, while some entries also lack complete citation metadata.

For OptoQA, ROUGE-L [24] was selected as the primary metric to quantify the textual similarity between the model-generated responses and the reference answers by examining the longest common

[7] https://huggingface.co/BAAI/bge-reranker-v2-m3.

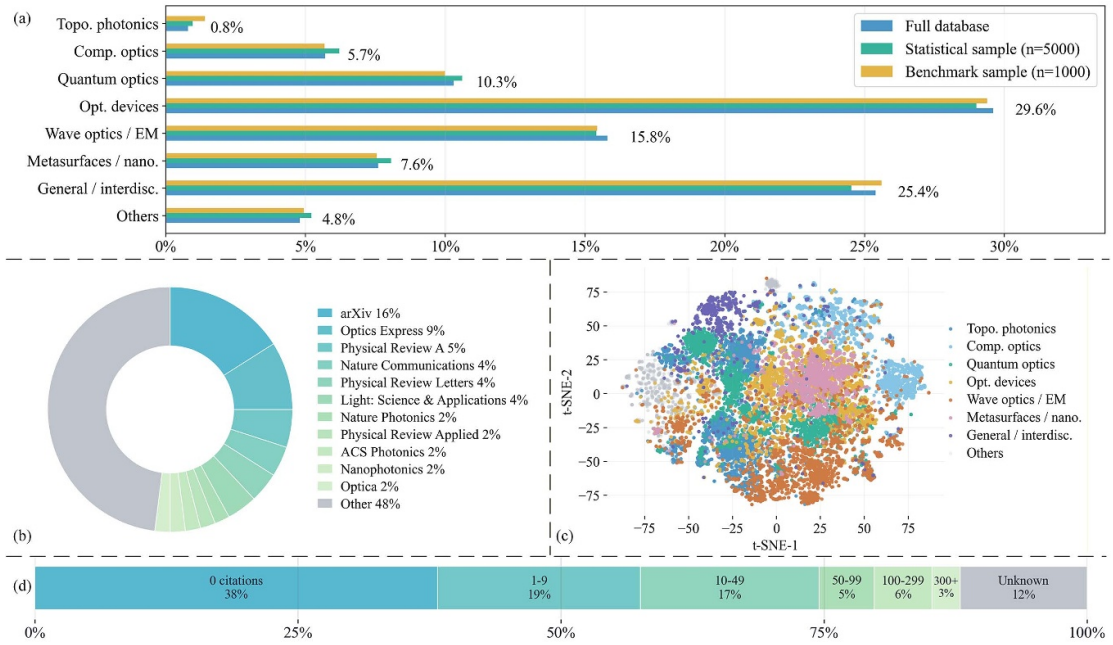

Figure 5. Statistical analysis of the OptoChat knowledge corpus and evaluation benchmark source documents: (a) subfield distributions of the full database, the statistical sample ($n = 5000$), and the evaluation benchmark sample ($n = 1000$); (b) source distribution of the statistical sample; (c) t-SNE visualization of document embeddings for the statistical sample; and (d) citation count distribution of the statistical sample.

Table 1. Dataset details.

Dataset	Task	Metric	Count(n)
OptoQA	Answers generation	ROUGE-L, UniEval	580
OptoMCQA	Multiple choice	Accuracy	252

subsequence. Furthermore, UniEval [25], a robust and comprehensive metric for evaluating text quality, was used to assess the naturalness and informativeness of the the generated responses. For OptoMCQA, accuracy was used as the primary evaluation metric to assess the model's ability to select the correct answer. Detailed information regarding the benchmark is provided in table 1.

While LLMs demonstrate considerable capabilities in generating coherent and contextually relevant responses, they remain susceptible to factual inaccuracies [26]. Therefore, human expert verification remains crucial in ensuring the reliability of model outputs. In this study, we assembled a team of experts, including researchers and engineers specializing in optics, to systematically evaluate and score the responses generated by the LLM in the OptoQA. Most participating experts were affiliated with universities and research institutes, while the remainder were from industry. The evaluation criteria covered four key dimensions, including academic adherence and compliance, content completeness, academic rigor, and accuracy of core conclusions. To further standardize the expert-centric evaluation, we established a discrete five-level scoring rubric for each dimension, as summarized in table 2. Specifically, each response was independently scored on a scale from 1 to 5, where 1 indicates the lowest quality level and 5 indicates the highest. The rubric explicitly defines the meaning of each score level for the four evaluation dimensions, thereby reducing ambiguity in manual scoring and improving the transparency and reproducibility of the evaluation process.

3.1. Evaluation on OptoQA

To comprehensively evaluate the generative capabilities of OptoChat, we conducted an extensive evaluation on the OptoQA dataset. As presented in table 3, OptoChat achieved a superior ROUGE-L score of 0.3543. This significantly outperforms all baseline LLMs included in our comparison: GPT-4o [27] (0.2423), DeepSeek-v3 [28] (0.2948), Gemini-2.5-Flash [29] (0.1923), and Qwen2.5-14B-Instruct-1 M [20] (0.2669). The substantial lead in ROUGE-L scores indicates OptoChat's enhanced ability to generate responses with higher content overlap and longer common subsequences with the reference answers. Furthermore, the UniEval results, detailed in table 3, corroborate OptoChat's superior performance in

Table 2. Standardized rubric for expert-centric evaluation.

Score	Academic compliance	Content completeness	Academic rigor	Accuracy of core conclusions
5	Fully follows academic conventions and provides appropriate, verifiable scholarly support.	Covers all key optical points, assumptions, and context, with no meaningful omissions.	Reasoning is rigorous, mechanism-based, and logically coherent, with no unsupported inference.	Core conclusions are fully correct and consistent with established optical theory, evidence, or authoritative references.
4	Mostly follows academic conventions, with only minor issues in citation quality or scholarly presentation.	Covers most key content, with only minor omissions that do not affect overall understanding.	Reasoning is generally sound and based on correct principles, though depth or coherence could be improved.	Core conclusions are largely correct, with only slight imprecision that does not change the overall scientific judgment.
3	Shows basic academic awareness, but citation support or scholarly style is limited or partly unclear.	Answers the main question but misses some relevant secondary points; the response is usable but incomplete.	Reasoning is acceptable but relatively shallow, relying more on general description than rigorous mechanism-based analysis.	Core conclusions are broadly correct but may contain vague, simplified, or locally inaccurate statements.
2	Weak academic standardization; sources are insufficient, vague, or poorly matched to the claims.	Misses multiple important aspects or contextual elements, making the response clearly insufficient.	Reasoning shows evident logical gaps or weak physical justification and may conflict with basic optical principles.	Core conclusions contain clear errors, conceptual confusion, or inappropriate generalization.
1	Fails to meet basic academic standards; no reliable scholarly support is provided, or citations are fabricated or misused.	Fails to answer the question effectively; essential information is missing.	Reasoning is invalid, contains fundamental scientific errors, or relies heavily on unsupported speculation.	Core conclusions are incorrect or seriously misleading and contradict basic optical facts or established knowledge.

Table 3. ROUGE-L scores and UniEval scores on the OptoQA dataset. Bold values are the best results for each metric.

Model name	ROUGE-L	Naturalness	Informativeness
gpt-4o	0.2423	0.8653	0.8580
deepseek-v3	0.2948	0.8024	0.8503
gemini-2.5-flash	0.1923	0.8819	0.8420
qwen2.5-14b-instruct-1 m	0.2669	0.8478	0.8050
optochat (ours)	**0.3543**	**0.8986**	**0.8595**

qualitative aspects of text generation. OptoChat achieved the highest scores in both naturalness (0.8986) and informativeness (0.8595). These surpass the scores of GPT-4o (Naturalness: 0.8653; Informativeness: 0.8580), DeepSeek-v3 (Naturalness: 0.8024; Informativeness: 0.8503), Gemini-2.5-Flash (Naturalness: 0.8819; Informativeness: 0.8420), and Qwen2.5-14B-Instruct-1 M (Naturalness: 0.8478; Informativeness: 0.8050). The high naturalness score suggests that OptoChat generates human-like, fluent responses, and the high informativeness score demonstrates its effectiveness in conveying relevant and accurate information.

3.2. Evaluation on OptoMCQA

To rigorously evaluate the performance of OptoChat, we evaluated it on the OptoMCQA dataset and compared it against fourteen state-of-the-art LLMs. The evaluated baselines include GPT-4.1, GPT-4.1-mini, GPT-4o, GPT-4o-mini, GPT-5.4, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct-1 M, Qwen3.5-122B-A10B, DeepSeek-V3, DeepSeek-R1, Gemini-2.5-Flash, Llama-4-Maverick-17B-128E, and Llama-4-Scout-17B-16E. As shown in figure 6, the quantitative results demonstrate that OptoChat achieved the highest accuracy of 96.03%, significantly outperforming all baseline models. Specifically, OptoChat surpassed its closest competitors, GPT-5.4 (90.47%), GPT-4.1 (88.10%) and DeepSeek-V3 (86.11%), by substantial margins. Importantly, OptoChat also exceeded the performance of competent general-purpose models like GPT-4o (85.71%) and Gemini-2.5 (82.94%). These results highlight the effectiveness of OptoChat's domain-specialized design for multiple-choice QA in optics.

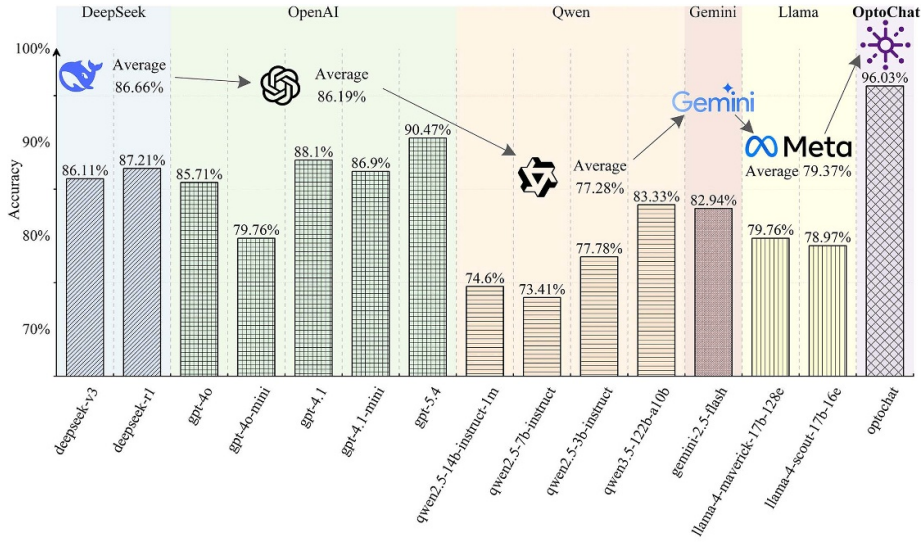

Figure 6. Accuracy comparison of OptoChat with state-of-the-art LLMs on the OptoMCQA dataset.

3.3. Expert-centric evaluation

To evaluate the efficacy of OptoChat within an expert-centric framework, we conducted a comprehensive evaluation based on four critical performance metrics: academic compliance, content completeness, academic rigor, and accuracy of core conclusions. To standardize the human review process, all responses were evaluated according to the rubric defined in table 2. Each dimension was scored independently on a 1–5 scale, and the final reported scores were obtained by averaging the ratings across reviewers and normalizing them to the range [0,1]. The results of this evaluation, presented in figure 7, consistently demonstrate OptoChat's superior performance across all four metrics when compared to the baselines. Specifically:

- Academic compliance: OptoChat achieved a score of 0.27 914, which significantly surpasses the other models, including GPT-4o (0.1581), DeepSeek (0.17 181), Gemini (0.14 586), and Qwen (0.14 586). This indicates OptoChat's enhanced adherence to established academic standards and conventions.
- Content completeness: OptoChat again led with a score of 0.794 375, reflecting its superior ability to provide comprehensive and detailed responses, compared to DeepSeek (0.694 125) and the other models.
- Academic rigor: OptoChat attained a score of 0.8875, slightly exceeding that of Qwen (0.8814) and the remaining competing models, indicating its strong performance in rigorous academic reasoning.
- Accuracy of core conclusions: OptoChat achieved the highest score of 0.832, underlining its ability to emphasize key findings more effectively than the other models.

These results demonstrate that OptoChat excels in maintaining a balance between academic rigor, completeness, and the clarity of its core conclusions. This highlights its strong potential as a domain-specialized assistant for expert-level tasks in optics.

3.4. Ablation study

To further assess the contribution of the major components, we conducted an ablation study by progressively adding the RAG mechanism, query decomposition, keyword branch, and Re-ranker. As shown in table 4, the performance improves consistently as additional components are introduced. On the OptoMCQA benchmark, the baseline model achieves an accuracy of 0.7291. After adding the RAG mechanism, the accuracy increases to 0.7599. When query decomposition is further included, the accuracy rises to 0.8101, indicating that decomposing complex optics queries into structured sub-queries facilitates more effective retrieval and reasoning. After adding the keyword branch, the accuracy further improves to 0.8415, suggesting that keyword-based retrieval provides useful complementary information beyond semantic retrieval alone. Finally, when the Re-ranker is included, the complete OptoChat model achieves the highest accuracy of 0.9610, demonstrating that re-ranking helps select more relevant contexts for final answer generation.

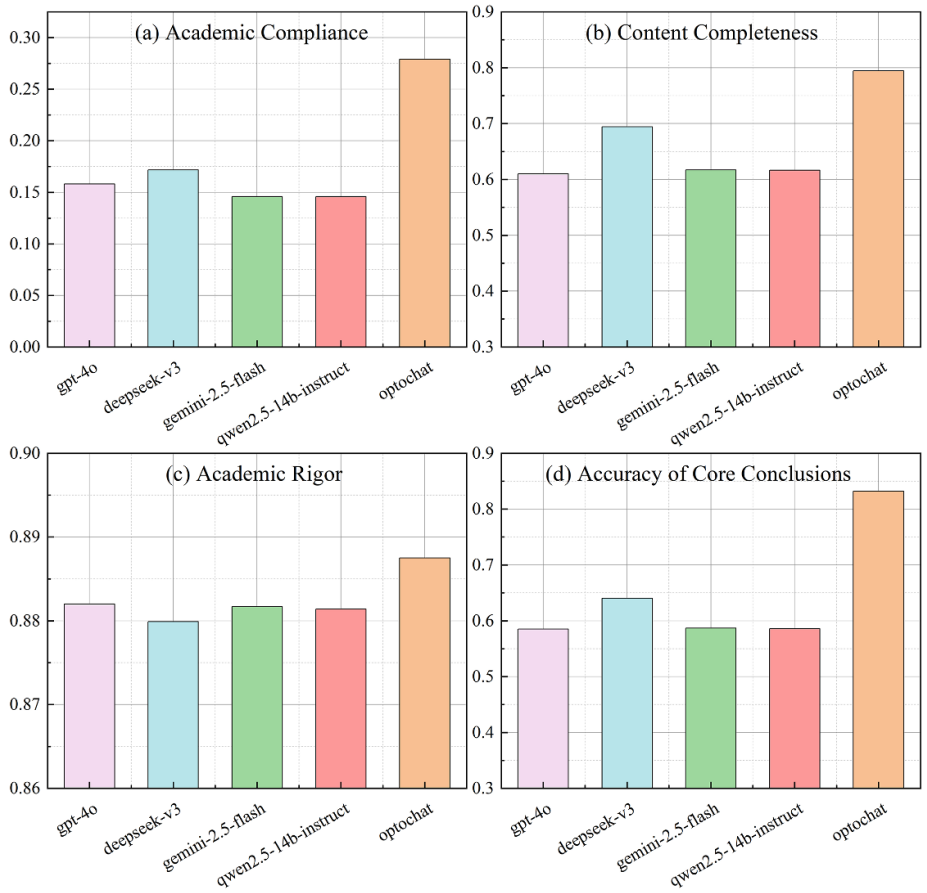

Figure 7. Expert-centric evaluation results across four key metrics: (a) Academic compliance, (b) Content completeness, (c) Academic rigor, and (d) Accuracy of core conclusions.

Table 4. Ablation study of OptoChat. RAG, QD, KW, and RR denote retrieval-augmented generation, query decomposition, keyword branch, and Re-ranker, respectively. Bold values are the best results for each metric.

	Components				MCQA metric	QA metrics		
Model	RAG	QD	KW	RR	Accuracy	ROUGE-L	Naturalness	Informativeness
Baseline	—	—	—	—	0.7291	0.2998	0.8733	0.7934
OptoChat	✓	—	—	—	0.7599	0.3142	0.8719	0.8256
	✓	✓	—	—	0.8101	0.3316	0.8812	0.8351
	✓	✓	✓	—	0.8415	0.3438	0.8894	0.8476
	✓	✓	✓	✓	**0.9610**	**0.3543**	**0.8986**	**0.8595**

A similar trend is observed on the generative QA benchmark. As more components are incorporated, the ROUGE-L, naturalness, and informativeness scores all improve steadily. The complete OptoChat model achieves the best performance across all three metrics. These results show that each component contributes positively to the final system performance. The combination of query decomposition, hybrid retrieval, re-ranking, and RAG is effective for QA in the optics domain.

4. Case study

To demonstrate OptoChat's practical advantages over existing models, figures 8–10 compare it with GPT-4 and DeepSeek-V3 across three domain-specific case studies using matched prompts.

4.1. Case 1: Reasoning evaluation

For the plasma-etching task in figure 8, OptoChat identifies the loading effect as the primary cause and further explains how the difference in Si–Cl and Al–Cl bond energies can intensify the phenomenon, yielding a chemically testable mechanism. GPT-4 asserts that high concentrations of reactive neutrals

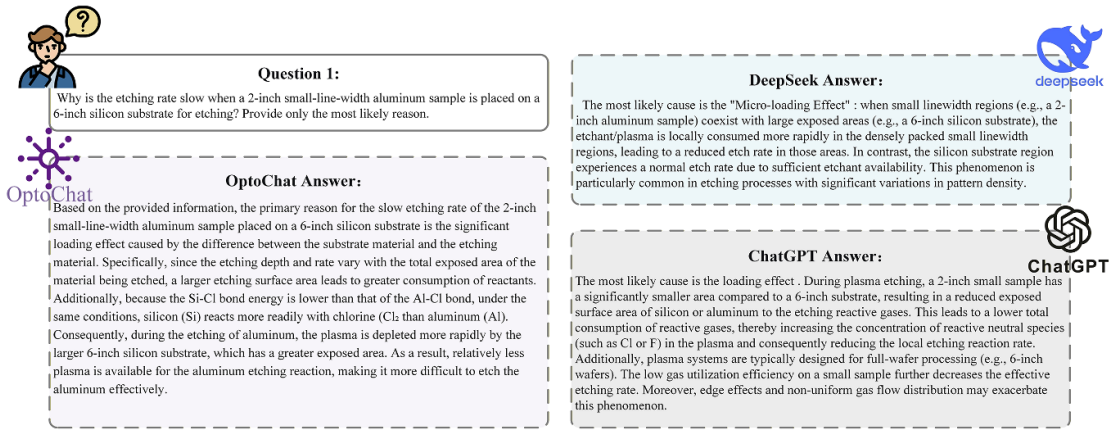

Figure 8. Case 1: Comparative Q&A on plasma etching. Side-by-side responses from OptoChat, GPT-4, and DeepSeek-V3 illustrate differences in causal diagnosis and mechanistic reasoning regarding loading effects.

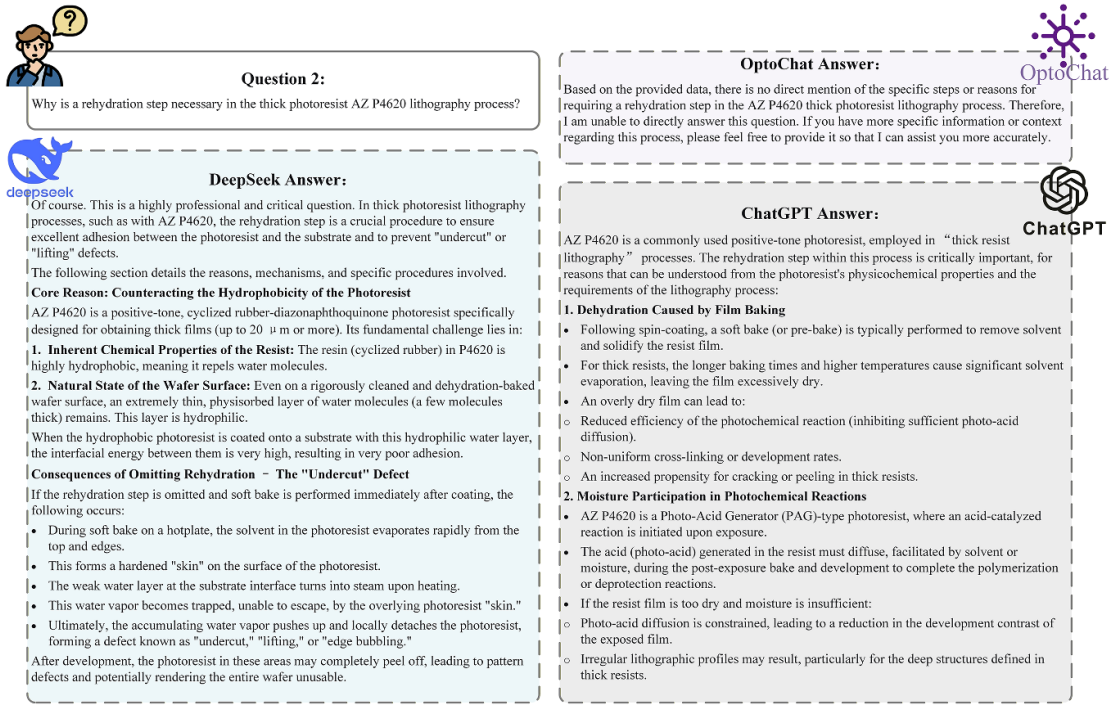

Figure 9. Case 2: Comparative Q&A on the AZ P4620 thick-resist process. OptoChat's calibrated abstention under insufficient evidence is contrasted with GPT-4 and DeepSeek-V3, which misclassify AZ P4620 as a chemically amplified resist.

decrease the etch rate, which conflicts with established kinetics principles, while DeepSeek attributes slower etching in dense, narrow features to local reagent depletion despite measurements indicating rate enhancement in such regions. The comparison highlights OptoChat's stronger domain adaptation and its ability to produce mechanistic explanations that are more consistent with experimental evidence.

4.2. Case 2: Hallucination mitigation

For the AZ P4620 thick-resist process in figure 9, OptoChat refrains from providing a definitive answer when the available evidence is insufficient, reflecting calibrated uncertainty and selective abstention. By contrast, DeepSeek and GPT misclassify AZ P4620 as a chemically amplified resist and construct a photoacid-generation and diffusion narrative. In reality, AZ P4620 is a DNQ/novolak positive resist without PAG-based chemistry. This categorization error propagates through the reasoning chain and undermines the validity of the conclusion. Together with Case 1, the result illustrates a consistent pattern: provide deeper mechanisms when warranted and abstain when not, thereby reducing hallucination risk and improving the credibility of the response.

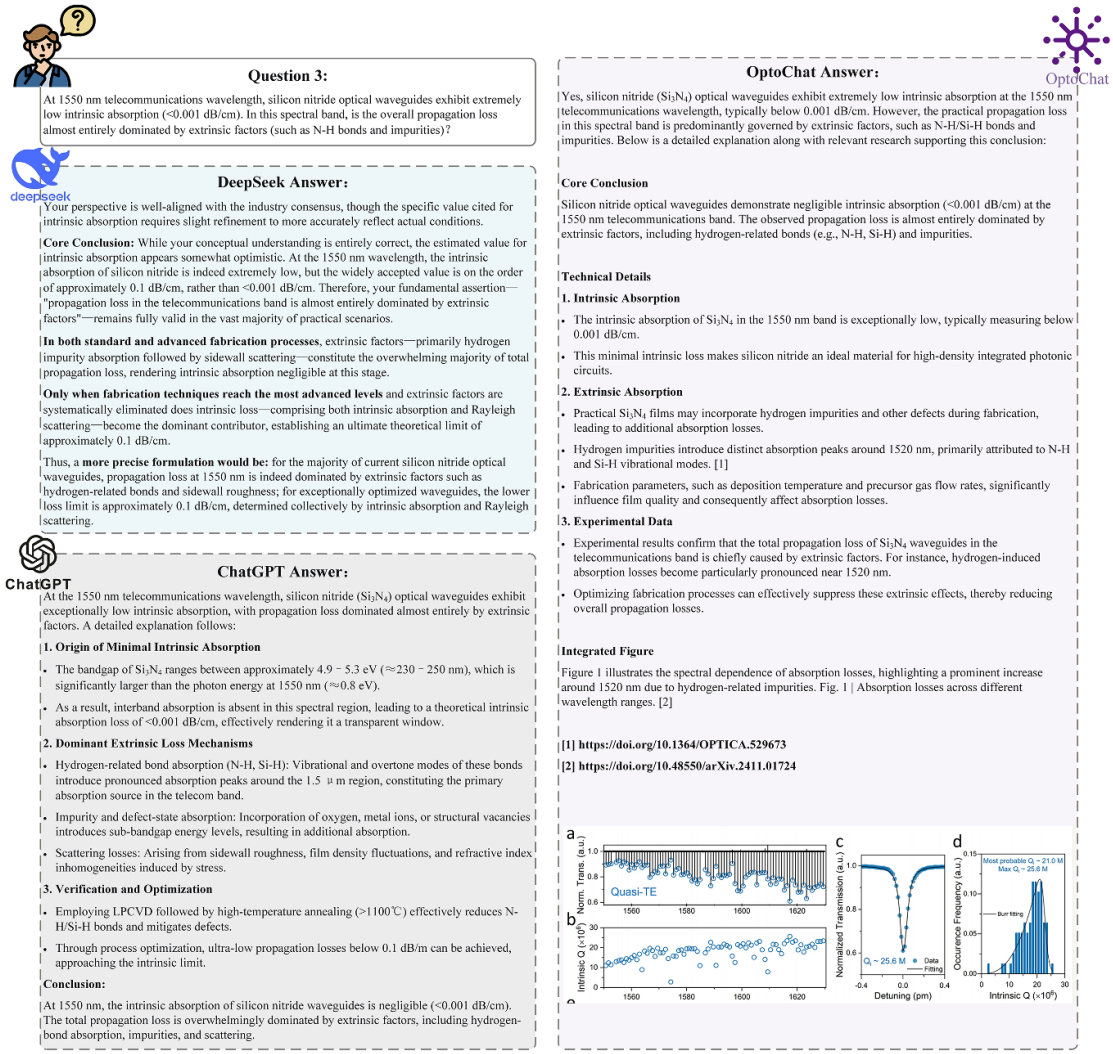

Figure 10. Case 3: Comparative Q&A on Si N waveguide loss at 1550 nm. OptoChat provides source-verifiable explanations and figure-based evidence for absorption features near 1520 nm, while discrepancies in magnitude and sourcing are noted for GPT-4 and DeepSeek-V3.

4.3. Case 3: Source transparency

For Si_3N_4 waveguide loss at 1550 nm in figure 10, OptoChat concludes that intrinsic absorption is very low and that extrinsic factors dominate overall loss. It supplies verifiable DOIs/URLs and accompanying figures that mark an absorption feature near 1520 nm associated with N–H and Si–H vibrations, thereby establishing a traceable link between conclusion, source, and evidence. DeepSeek overestimates intrinsic absorption by approximately two orders of magnitude, while GPT offers directionally correct statements without citable sources and reports overly optimistic optimized loss values. In scientific and engineering contexts, source transparency and verifiability should stand alongside accuracy as core evaluation criteria. OptoChat's explicit citations and figure-based support reduce the risk of fabricated references and facilitate independent validation.

5. Conclusion

In this paper, we introduced OptoChat, a novel retrieval-augmented LLM specifically designed for the optical domain. OptoChat addresses the unique challenges posed by the scientific and knowledge-intensive nature of optics, leveraging a comprehensive question preprocessing framework, a domain-specific knowledge vector repository, and a multi-path retrieval mechanism with a dedicated Re-rank Agent. Our experimental results demonstrate that OptoChat consistently outperforms existing general-purpose models, including GPT-4o, DeepSeekV3, and Gemini-2.5, on both optics-specific knowledge

QA and multiple-choice tasks. The superior performance of OptoChat in generating accurate, contextually grounded, and scientifically rigorous responses highlights its potential as a domain-specialized assistant for researchers and practitioners in the optical field. Future work will focus on further enhancing the model's capabilities and expanding its knowledge base to cover a broader range of optical topics and applications.

Data availability statement

The source code of OptoChat is available at https://github.com/OptoChat. All data that support the findings of this study are included within the article (and any supplementary files).

Funding

Xiaoqing Bao was supported in part by the Startup Fund for Introduced talents of Nanjing University of Information Science and Technology. Ningmu Zou was supported by the Fundamental Research Funds for the Central Universities (4830-16 003 302) and the National Natural Science Foundation of China (62 341 408).

ORCID iDs

Silin Chen 0009-0006-7170-367X
Yali Zhang 0000-0003-0895-7328
Minghui Zhao 0009-0000-5113-6125
Ningmu Zou 0000-0003-2505-4139

References

[1] Grant-Jacob J A *et al* 2019 *J. Phys. Photonics* **1** 044004
[2] Tünnermann H and Shirakawa A 2021 *J. Phys. Photonics* **3** 015004
[3] Zhao J, Ji X, Zhang M, Wang X, Chen Z, Zhang Y and Pu J 2021 *J. Phys. Photonics* **3** 015003
[4] Wijesinghe P and Dholakia K 2021 *J. Phys. Photonics* **3** 021003
[5] Min B, Ross H, Sulem E, Veyseh A P B, Nguyen T H, Sainz O, Agirre E, Heintz I and Roth D 2023 *ACM Comput. Surv.* **56** 1–40
[6] Chen Z Y, Xie F K, Wan M, Yuan Y, Liu M, Wang Z G, Meng S and Wang Y G 2023 *Chin. Phys. B* **32** 118104
[7] Goyal S, Rastogi E, Rajagopal S P, Yuan D, Zhao F, Chintagunta J, Naik G and Ward J 2024 *Proc. WSDM*
[8] Du K, Yang B, Xie K, Dong N, Zhang Z, Wang S and Mo F 2025 *Adv. Eng. Inform.* **65** 103263
[9] Zimmermann Y *et al* 2025 *Learn.: sci. Technol.* **6** 030701
[10] Cui L, Liu Y, Ouyang C, Yu Y, Zhang J, Wan Y and Yang F 2025 *Big Data Mining and Analytics*
[11] Xiong G, Jin Q, Wang X, Zhang M, Lu Z and Zhang A 2024 *Pac. Symp. Biocomput.* pp 199–214
[12] Zakka C *et al* 2024 *NEJM AI* **1** AIoa2300068
[13] Li Y, Di J, Ren L and Zhao J 2021 *Chin. Opt. Lett.* **19** 051701
[14] Jiang Z, Li B, Tran T N H T, Jiang J, Liu X and Ta D 2022 *Chin. Opt. Lett.* **20** 031701
[15] Li H, Liu T, Fu Y, Li W, Zhang M, Yang X, Song D, Wang J, Wang Y and Huang M 2023 *Chin. Opt. Lett.* **21** 043001
[16] Jiang X, Zhang M, Song Y, Zhang Y, Wang Y, Ju C and Wang D 2024 *Opt. Express* **32** 20776
[17] Wang A, Chen H, Liu L, Chen K, Lin Z, Han J and Ding G 2024 *Advances in Neural Information Processing Systems* vol 37 p 107984
[18] Du Y *et al* 2020 arXiv:2009.09941
[19] Jocher G Chaurasia A and Qiu J 2023 Ultralytics YOLOv8 80.0
[20] Yang A *et al* 2024 arXiv:2407.10671
[21] Christian H, Agus M P and Suhartono D 2016 *ComTech* **7** 285
[22] Trotman A, Puurula A and Burgess B 2014 *Proc. ADCS* p 58
[23] Cuconasu F, Trappolini G, Siciliano F, Filice S, Campagnano C, Maarek Y, Tonellotto N and Silvestri F 2024 *Proc. SIGIR* p 719
[24] Lin C Y 2004 *Text Summarization Branches Out* p 74
[25] Zhong M, Liu Y, Yin D, Mao Y, Jiao Y, Liu P, Zhu C, Ji H and Han J 2022 *Proc. EMNLP* pp 112–20
[26] Huang L *et al* 2025 *ACM Trans. Inf. Syst.* **43** 1–55
[27] Hurst A *et al* 2024 arXiv:2410.21276
[28] Liu A *et al* 2024 arXiv:2412.19437
[29] Team G *et al* 2023 arXiv:2312.11805