陈乔晟

+86-15105380602 qschen@smail.nju.edu.cn cqsss.github.io sdta_cqsss

陈乔晟，博士三年级，来自于**南京大学计算机学院** Websoft 研究组，导师为程龚教授

- 研究方向为 **Web 代码生成、大数据搜索、知识图谱**和**检索增强生成（RAG）**
- 入选**首届中国科协青年人才托举工程博士生专项计划**，托举学会为中国中文信息学会（CIPS）
- 以第一/共一作者身份发表 **CCF-A 类论文 5 篇**，**CCF-B 类论文 3 篇**，其他论文 2 篇
- 获 **ISWC 2023 最佳研究论文提名**，**CCF BigData 2025 最佳应用论文**

教育背景

2021.09 至今 | **南京大学** · 计算机学院 · 计算机科学与技术专业 · 硕博连读
预计 2027.06 毕业 | **博士生国家奖学金** | **优秀研究生标兵**

2017.09 2021.06 | **哈尔滨工业大学（威海）** · 计算机科学与技术学院 · 计算机科学与技术专业 · 工学学士
学分绩：90.11/100 | 专业排名：5/137 | **本科生国家奖学金** | **山东省优秀毕业生**

实习经历

阿里巴巴-通义实验室-千问大模型 2026.03-2026.06

WebDev 预训练

- 参与 Qwen 基础模型代码能力中 WebDev 项目，主要负责预训练数据的清洗、筛选与合成，以及预训练测试和验证。
- 设计 file-level 和 repo-level 代码数据的筛选、分类、打标流程，贡献数据至主线预训练。

腾讯-TEG-混元 AI Data 部（青云计划） 2025.11-2026.02

Deep Research Agent

- 参与 Deep Research Agent 项目，主要负责 agent plan 能力的设计与增强。
- 构建基于 DAG 的 plan 和 summary 策略，SFT+RL 优化模型 plan 能力。

上海人工智能实验室 2025.04-2025.09

可交互科学演示代码生成

- 研究大语言模型在**生成科学演示网站代码**方面的能力，收集并整理种子网站数据，设计评测基准，构建合成数据集。
- 提出针对交互功能的硬测试和基于多截图的软测试方法，突出“真正的交互性验证”与“**snapshot as test**”的特色。
- 工作成果计划以**第一作者**论文形式发表在 **ICML 2026**。

[**ICML 2026**] InteractScience: Programmatic and Visually-Grounded Evaluation of Interactive Scientific Demonstration Code Generation

多模态代码大模型

- 参与训练面向代码绘图、前端网页生成、多模态算法题与冷门可视化语言的语言模型与多模态模型。
- 主要负责 HTML 前端代码的**数据收集**、**合成**与**验证**，推动模型在前端代码生成能力上的提升。
- 工作成果计划以**共同一作**论文形式发表在 **ICLR 2026**；产出数据贡献至 **Intern-S1-Pro** 模型。

[**ICLR 2026**] JanusCoder: Towards a Foundational Visual-Programmatic Interface for Code Intelligence
[**Technical Report**] Intern-S1-Pro: Scientific Multimodal Foundation Model at Trillion Scale

科研经历

[**SIGIR 2025**] Benchmarking Recommendation, Classification, and Tracing Based on Hugging Face Knowledge Graph **第一作者**

- 构建了首个基于 Hugging Face 社区的 AI 资源知识图谱 HuggingKG（包含 260 万节点和 620 万边），系统刻画了**模型**、**数据集**等 AI 资源间的**关系**与**属性**，并进一步设计了 HuggingBench 评测基准，涵盖资源**推荐**、**分类**与**溯源**三类新型测试集合。

[**SIGIR 2024**] Enhancing Dataset Search with Compact Data Snippets **第一作者**

- 针对数据集搜索**对数据内容利用不足**且**结果可解释性差**的挑战，提出了基于子图抽取的紧凑数据片段抽取方法 **CDS**，能够生成**与查询相关的代表性片段**，在提升检索精度的同时，为用户提供数据集中与查询相关的内容，以增强结果的可理解性。

[**SIGIR 2024**] ACORDAR 2.0: A Test Collection for Ad Hoc Dataset Retrieval with Densely Pooled Datasets and Question-Style Queries **第一作者**

- 针对现有评测集中**词汇偏置严重**和**查询形式单**一的问题，构建了基于内容的数据集检索评测集 **ACORDAR 2.0**，利用稠密检索模型扩展候选数据集以缓解偏置，并基于大语言模型将关键词查询改写为高质量的问题式查询提升了评测多样性。

[**ISWC 2024**] DUNKS: Chunking and Summarizing Large and Heterogeneous Data for Dataset Search **第一作者**

- 针对数据集内容**大规模**和**异构性**的挑战，提出基于图的**统一表示模型**支持多种数据格式的统一处理，并基于该模型提出基于贪心模式覆盖的**多片段排序**方法 **DUNKS**，在 NDCG 和 MAP 等指标上显著提高数据集排序效果。

[**ISWC 2023**] Dense Re-Ranking with Weak Supervision for RDF Dataset Search **第一作者** 🏆 **最佳研究论文提名**

- 针对数据集搜索**标记数据不足**的挑战，提出方法 **DR2**，通过外部训练数据**远监督**和生成模型**自训练**方法生成伪标记数据，并采用**由粗到细**的训练策略微调稠密检索模型，提升了 DPR 和 ColBERT 两个稠密检索模型的排序效果。

[**SIGIR 2025**] µDS: Multi-Objective Data Snippet Extraction for Dataset Search 第二作者

- 提出将**紧凑性**、**相关性**、**代表性**和**内聚性**四个目标联合优化的数据片段抽取方法 **µDS**，将其建模为一个新的组合优化问题，并设计了具有最坏情况近似比保证的高效算法，实验验证了该方法在提升数据片段质量和数据集检索效果方面的有效性。

[**SIGIR 2022**] ACORDAR: A Test Collection for Ad Hoc Content-Based (RDF) Dataset Retrieval 第二作者

- 针对数据集搜索评测集**依赖元数据**进行标注的问题，构建了**首个基于数据内容**的 RDF 数据集检索评测集 **ACORDAR**，验证了数据集内容在检索中的价值，还设计了专门用于浏览 RDF 数据集的 dashboard，以便全面、高效地浏览 RDF 数据集。

[**ISWC 2025**] mmRAG: A Modular Benchmark for Retrieval-Augmented Generation over Text, Tables, and Knowledge Graphs 第二作者

- 设计了用于评估多源数据 RAG 系统的模块化评测基准 **mmRAG**，整合了来自六个不同问答数据集（**文本**、**表格**与**知识图谱**）的查询和数据，统一转换为可检索文档，使用大模型进行相关性标注，使得检索精度、查询路由等组件能够被细粒度评估。

[**EMNLP 2023**] An Empirical Investigation of Implicit and Explicit Knowledge-Enhanced Methods for Ad Hoc Dataset Retrieval 第二作者

- 复现并分析了多种检索模型（如 coCondenser、ANCE、monoT5 等）在显式和隐式知识设置下的数据集搜索上的表现，揭示了**知识增强**方法在数据集搜索场景的有效性。

💻 系统研发

全国公共数据搜索系统 2024 年

- 主导设计并研发全国公共数据搜索系统，**收集**、**整合**、**索引**来自全国 25 个省级行政区的 148 个公共数据开放网站的数据集，实现**关键词搜索**、**分面搜索**、**结果展示**等功能。创新性地利用大语言模型（LLM）实现**自动化元数据整合**、**高精度数据集排序**和**搜索结果相关性解释**。
- 相关论文分别发表在 CCF BigData 2023 和 CCF BigData 2025 会议，获 🏆 **CCF BigData 2025 最佳应用论文**，并推荐至 **CCF-B 类期刊 DSE** 发表。

🏅 荣誉奖励

首届中国科协青年人才托举工程博士生专项计划（全国各专业共 3226 人） 2024 年
国家奖学金（博士生）（计算机学院共 14 人） 2024 年
南京大学优秀研究生标兵（计算机学院共 15 人） 2024 年
山东省优秀毕业生 2021 年
CCPC 中国大学生程序设计竞赛总决赛铜奖 2019 年
ACM-ICPC 国际大学生程序设计竞赛亚洲区域赛（上海站）银奖 2019 年
CCPC 中国大学生程序设计竞赛厦门站银奖 2019 年
ACM-ICPC 国际大学生程序设计竞赛亚洲区域赛（徐州站）银奖 2018 年
国家奖学金（本科生）（计算机学院共 5 人） 2018 年