Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

学术论文评审意见生成方法研究

硕士研究生 叶沐阳

2025 年 09 月 21 日

问题回溯

- 相关内容
  - 2023.08.20 杨宗源 《文本生成中的幻觉》

内容提要

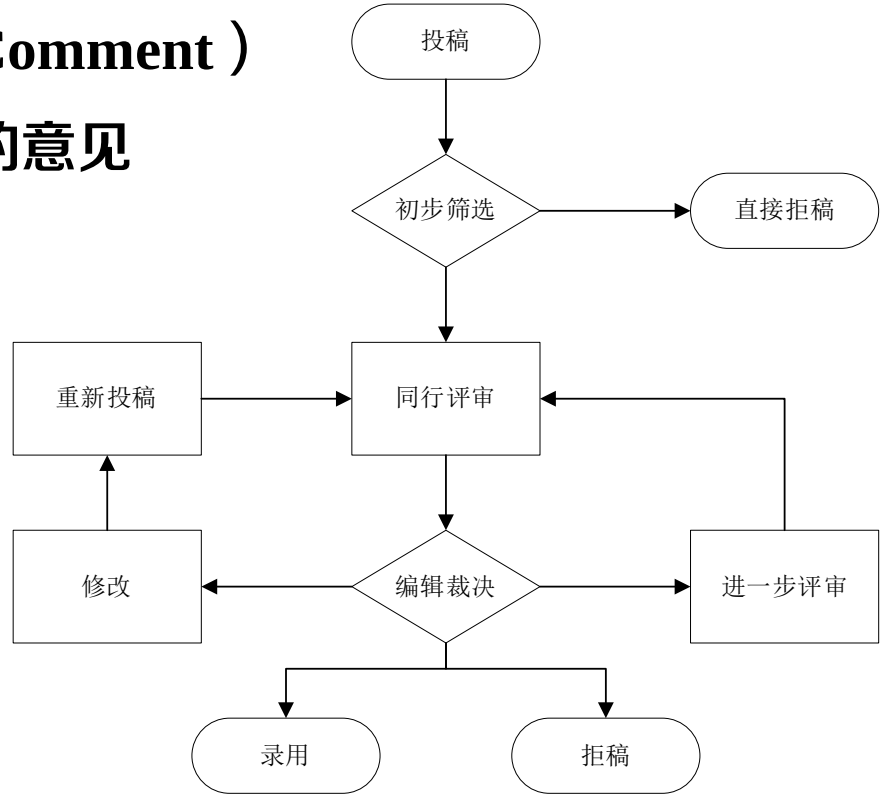
- 预期收获
- 内涵解析与研究目标
- 研究背景与研究意义
- 研究历史
- 知识基础
- 算法原理
  - $SWIF^2T$
  - SEA
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 掌握学术论文评审意见生成的基本概念
  - 了解学术论文评审意见生成的研究背景和研究意义
  - 了解学术论文评审意见生成的前沿方法和未来发展

内涵解析与研究目标

- 内涵解析
  - 同行评审（Peer review）
    - 在期刊发表前，将作者的稿件交给同一领域其他专家进行评审的过程
  - 学术论文评审意见（Academic Paper Review Comment）
    - 同行评审过程中评审专家针对论文稿件提出的意见



- 研究目标
  - 实现自动化学术论文评审
  - 针对论文稿件中的优缺点提出对应意见
  - 提高论文评审效率，促进知识传播

研究背景与意义

- 研究背景
  - 学术论文投稿数量的不断增长带来**巨大审稿压力**
  - 人工智能和大数据为学术论文自动化评审技术提供了发展契机，即利用计算机或其他智能机器对论文稿件进行**初步筛查**，并为合规论文生成**评审报告**
- 研究意义
  - 探索学术论文的自动化评审方法，对**减轻审稿人负担**、**提高评审效率**，具有重要的应用价值

研究历史

Bartoli等从真实评审语料库中提取句子，将其中的领域相关术语替换为目标论文中的对应术语得到句子库，再通过预设的总体评价从句字库中组合句子得到最终的评审意见

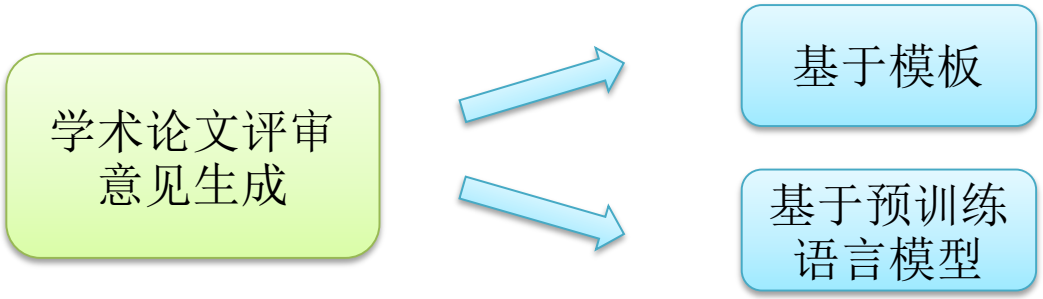
Robertson通过先让GPT-4在论文的不同部分生成注释并将其组织成一致的格式，再让GPT-4通过这些注释和一份评审意见生成的指导指南来生成评审意见

SWIF^{2}T设计四个组件，规划者负责制定问题，调查员负责检索，审查员利用检索到的信息识别需要修订的部分，控制器决定执行或者替换下一步。以此来生成更高质量的评审意见

ReviewRobot使用领域特定的信息抽取技术构建知识图谱，将人工精选的评审语句泛化为模板，自动生成知识丰富且可解释的分数和评审意见

Zhu等人构建自动化的数据预处理管道，通过整合新颖性、可靠性、逻辑完整性等评估模型以提取高质量的监督微调数据集，从训练数据质量角度提升生成意见的有效性

Yu等将一篇论文中包含的多个评审意见进行提炼整合，使用该评审意见对预训练语言模型进行监督微调，避免模型优化过程出现训练目标不一致的现象，从而影响生成评审意见的质量



知识基础 大语言模型

- 大语言模型（Large Language Model，LLM）
  - 基于深度神经网络构建的人工智能模型
  - 在海量的文本数据中训练而成，包含百亿级甚至千亿级参数
  - 可以执行广泛的任务，包括文本总结、翻译、情感分析等
- 监督微调（Supervised Fine-Tuning，SFT）
  - 以预训练模型为基础，通过有标注的目标任务数据集对模型进行二次训练，使其适应特定任务需求
  - 主流方法
    - 全量微调：调整模型的所有参数。下游任务效果好，但算力要求极高
    - 部分微调：仅微调顶层参数。效果相较于全量微调稍差，但计算成本大幅降低
    - 高效微调：仅修改模型中对结果影响大的部分参数。代表方法有LORA、Adapter等

知识基础 BLEU分数

- $N-gram$
  - 将文本拆分成若干个连续的n个词的序列，并统计这些序列在文本中出现的频率
  - 以“我喜欢学习自然语言处理”为例
    - $1-gram$是单个词，如“我”、“喜欢”等
    - $2-gram$是相邻的两个词组成的词组，如“我喜欢”、“喜欢学习”等
    - $3-gram$则是相邻的三个词组成的词组，如“我喜欢学习”等
- $BLEU$分数：主要统计参考文本和生成文本的$N-gram$匹配程度（精确率）

$$BLEU = BP \times \exp(\sum_{n=1}^{N} W_n \times logP_n)$$

其中BP是惩罚因子，用以惩罚生成文本过短带来的精确度提升；P_n是$n-gram$的精确率，且同一个词在生成文本中的计数不超过其在参考文本中的计数； W_n是$n-gram$的权重

知识基础 ROUGE分数

- $ROUGE$分数：主要统计有多少参考文本中的N元词组出现在生成文本中（召回率）
  - $ROUGE-N$：主要统计生成文本的$N-gram$召回率

$$ROUGE-N=\frac{\sum_{gram_N\in S}Count_{match}(gram_N)}{\sum_{gram_N\in S}Count(gram_N)}$$

其中分母是参考文本中$N-gram$数，分子是参考文本与生成文本的共有$N-gram$数

  - $ROUGE-L$：L指最长公共子序列(Longest Common Subsequence, LCS)，主要统计生成文本和参考文本的最长公共子序列召回率

$$ROUGE-L=\frac{LCS(C,S)}{len(S)}$$

Automated Focused Feedback Generation for Scientific Writing Assistance

SWIF²T TIPO

T	目标	针对论文的某一部分提出具体连贯的评审意见
I	输入	学术论文*1、待评审段落*1（2581个学术论文-评审段落对）
P	处理	1.计划者制定任务计划并分解为一系列问题 2.调查者阅读论文或搜索网络回答问题生成上下文信息 3.评审者基于上下文信息，生成具体的评审意见 4.控制者管理执行过程，动态调整计划
O	输出	具体的评审意见*1

P	问题	现有的自动化评审侧重于改进论文的表面形式和风格而非具体内容
C	条件	模型需要访问待评审段落对应的完整学术论文及相关文献
D	难点	如何提高制定计划的质量
L	水平	ACL2024（CCF A）

SWIF²T 算法原理图

- 算法原理图
  - Planner
    - 制定任务计划，分解为一系列问题
  - Investigator
    - 阅读完整论文并搜索网络文献回答问题，生成上下文信息
  - Reviewer
    - 基于上下文信息，识别论文段落中的具体弱点，生成评审意见
  - Controller
    - 管理执行过程，检查计划是否合理并动态调整

SWIF²T Planner

- 候选计划生成
  - 阅读论文全文提问（如“实验方法是什么？”）
  - 检索外部文献提问（如“其他研究对该方法的评价是什么？”）
- 最佳计划选择
  - 评价标准
    - 遵循正确的格式，确保程序的成功执行
    - 具有连贯性，确保其执行能纳入论文和文献中的相关背景
    - 针对输入段落，探索其中的相关概念
  - Plan re-ranking
    - 训练模型辨别优劣计划
    - 为候选计划打分，选择得分最高的计划

System message You are the ReviewGPT Planner, a world class scientific reviewing assistant. You create plans using the Investigator and Reviewer AI agents to review paragraphs. You will ask the Investigator to gather context from both the web and the paper in the first few steps, then, at the end, the action you ask the Reviewer agent will be exactly “Write a review based on the gathered context.”. DO NOT add a single word to this sentence. Your output MUST be formatted as a numbered list. NEVER write a step that does not involve an action for the Investigator or the Reviewer agents.
User message You will be given a paragraph. Your task is to point out the weaknesses of this paragraph, i.e. ask questions to gather context from the paper and the web before reasoning over it and the paragraph to identify the weaknesses of the passage. Thinking step by step, break the process of scientific reviewing down into small, simple tasks. These should involve gathering context for the paragraph, i.e., gathering information from the paper (such that the paragraph can be understood, verified and criticized without requiring any access to the paper.) and from the literature (such that the Reviewer AI agent can understand cited studies, compare the paper againt other related studies, evaluate its originality and soundness and be aware of criticisms and limitations, all without needing any access to the literature. The questions should be self-contained and formulated to facilitate effective Google searches). The gathered information should allow the Reviewer to comment on the soundness, originality, replicability, meaningfulness of the comparison or the substance of the information discussed in the paragraph. As you make plans for other AI tools, each step should be solvable using one of the following actions:

1. Actor: Investigator | Action: Answer question using the paper | Parameters: question | Description: Answer the provided question from the provided paper. It is important that the query that is searched is a question ending in ‘?’.
2. Actor: Investigator | Action: Answer question using Google | Parameters: question | Description: Use google search to try to answer the provided question. It is important that the query that is searched on Google is a self-contained question ending in ‘?’.
3. Actor: Reviewer | Action: Write review | Parameters: | Description: Write a review that only points out the weaknesses and areas of improvement of a passage based on the plan so far. Can only be called once context has been gathered by another agent.

Your plan should be a numbered list. Steps should be in simple language, and mention which agent should do them.
I will give you now the golden rules by which you NEED to abide. It is of upmost important that none of these rules is broken:
Rule #1: Each step involves requesting the Investigator or the Reviewer to perform an action.
Rule #2: The plan begins with the Investigator answering questions using the paper. Each of these steps should start with “Search the paper to understand”. The next steps should request the Investigator to answer questions by searching the web. These should start with “Search the web to understand” and should be only about one idea. Questions answerable from the web should be self-contained such that they are understandable without referring to another step or the paper. Finally, the last step should be EXACTLY “Reviewer: Write a review based on the gathered context.”
Rule #3: The questions to the Investigator should ONLY ask about one concept at a time. For example, “Search the paper to understand what Attention is ” is valid but “Search the paper to understand what Attention is and how it works” is NOT.
IT IS IMPORTANT TO RESPECT THE THREE GOLDEN RULES I JUST GAVE YOU. Now, the paragraph you will review is:
{*paragraph*}

SWIF²T Investigator

- 相关文档检索
  - 检索论文全文
  - 使用google API检索相关文献，同时过滤已有的评审意见

- QA pipline
  - 将文档拆分为长度相等的段落并创建嵌入
  - 执行相似性搜索，选取最相似的前五个文本段落
  - 根据文本段落回答问题，当无法回答时输出“I don’t know.”

> **User message** Answer the question based on the context below. IF the question cannot be answered based on the context, return exactly ‘I don’t know’.
> {*context*}
> Question: {*question*}
> Answer:

SWIF²T Reviewer

- 子字符串选取
  - 从待评审段落中选取并引用子字符串
- 评论标签选取
  - 选择合适的评论标签
    - 原创性
    - 实证/理论健全性
    - 可重复性
    - 有意义的比较
    - 实质性
- 评审意见撰写
  - 根据子字符串、评论标签和上下文信息撰写评审意见

System message You are the ReviewGPT Reviewer, a world class AI assistant for scientific reviewers. You write a review that highlights the weaknesses and areas of improvements of a paragraph based on context given to you, and return results as valid JSON. You need to make sure that the review addresses a specific portion of the paragraph, that it is not generic and that it is constructive. If you think that no review is needed, then you can also say this. Also, make sure to use the given context to generate a review: a review that points out the absence of an information in the paragraph should not be made if this information is present in another paragraph of the paper, be careful this is very important!
User message You will be given a paragraph with the following context:
{*context*}
There are five possible review labels: Empirical and Theoretical Soundness, Meaningful Comparison, Substance, Originality, Replicability. Write a review that:

1. Selects and quotes a substring from the given paragraph.
2. Chooses the appropriate review label.
3. Writes a review sentence using the quoted substring, the review label and the context. It is IMPORTANT that you use the provided context to generate a sensible review.
4. Generates a JSON object with the keys "reasoning", "label" and "review". Below are examples that follow all these rules, use them as inspiration + {in-context learning examples}

SWIF²T Controller

- 执行步骤确认
  - Investigator：使用论文全文回答问题
  - Investigator：使用谷歌搜索回答问题
  - Reviewer：撰写具体的评审意见
  - Controller：跳过无法执行或不必要的步骤

System message You are the ReviewGPT Controller, a helpful scientific reviewing assistant. You manage several other AI agents, passing directions to them from the user. You communicate directly with the other AI agents, and as such your answers MUST be ONLY valid json.
User message You are currently following an overall plan to point out the weaknesses in the paragraph:
{*paragraph*}
This is a log of your progress so far:
{*progress*}
The remaining steps are:
{*steps*}
The next step is:
{*next step*}
You will be given a list of actions. Your task is to decide what the best action to take is to accomplish the next step. Each action has several fields, separated by a vertical line (|). These are the actor who takes the action, the name of the action, the parameters that action requires, and a short description of the action. The options are:

- Actor: Investigator | Action: Answer question using the paper | Parameters: question | Description: Answer the provided question from the provided paper. It is important that the query that is searched is a question ending in '?'.
- Actor: Investigator | Action: Answer question using Google | Parameters: question | Description: Use google search to try to answer the provided question. It is important that the query that is searched on Google is a question ending in '?'.
- Actor: Reviewer | Action: Write review | Parameters: | Description: Write a review that only points out the weaknesses and areas of improvement of a passage based on the plan so far. Can only be called once context has been gathered by another agent.
- Actor: Controller | Action: Skip this step | Parameters: | Description: Skip the current step if it is unnecessary or impossible

Provide the best action to take. Your answer must be valid JSON. It should be a JSON object with four entries, "explanation", "actor", "action", "parameters". Actors and actions should be strings, parameters should be another JSON object. Explanations should be a string containing a step-by-step description of why you chose this action. Remember, necessary parameters can be found between curly brackets in the commands. Output JUST the command.

SWIF²T 实验设置

- 数据资源
  - 数据集：由作者自主构建，从NLPeer、PeerRead、ASAP-Review、ARIES等多个公开数据集中筛选出的2581个段落以及与之相对应的人类评审意见
- 对比方法
  - GPT-4：功能强大但单一的大语言模型
  - CoVe（ACL2024）：一个专为减少幻觉设计的系统，流程类似SWIF²T，但缺少动态规划和外部检索
  - 人工撰写的评论：真实的人类评审意见
- 评价指标
  - 人工评价：由11位由论文评审经验的研究人员根据以下三个问题进行盲测对比：是否真正理解论文段落和相关文献？是否点明段落中的具体问题？是否给出建设性的修改方向？
  - 自动评价：ROUGE、BLEU、METEOR

SWIF²T 对比实验

- 实验结果
  - 人工评估：SWIF²T在三个问题维度和总体帮助度上均高于基线

	Winners			
	SWIF²T	GPT-4	CoVe	Human
SWIF²T	-	22.75	28.90	29.50
GPT-4	61.75	-	48.00	45.75
CoVe	50.00	31.50	-	40.25
Human	58.75	43.00	50.25	-
Total	**170.50**	97.25	126.25	115.50

	Winners			
	SWIF²T	GPT-4	CoVe	Human
SWIF²T	-	24.50	23.25	32.25
GPT-4	52.00	-	41.50	44.75
CoVe	46.00	30.00	-	41.75
Human	45.50	34.25	39.25	-
Total	**143.50**	88.75	104.00	118.75

	Winners			
	SWIF²T	GPT-4	CoVe	Human
SWIF²T	-	24.00	28.00	30.75
GPT-4	60.50	-	52.50	48.50
CoVe	53.25	29.00	-	42.00
Human	58.00	39.50	47.00	-
Total	**171.75**	92.50	127.50	121.25

  - 自动评估
    - SWIF²T与人工撰写的评论有最高的相似度

Model	METEOR	BLEU@4	ROUGE-L
GPT-4	18.13	28.50	18.37
CoVe	18.76	29.07	19.62
SWIF²T $_{RR}$	19.17	29.77	19.39
SWIF²T	**20.04**	**30.06**	**20.44**

SWIF²T 算法总结

- 算法贡献
  - 设计了一个多代理系统，可以生成具体且特异的评审意见
  - 帮助科学写作辅助以及学术论文评审

- 算法不足
  - 依赖于GPT-4和Google查询，金钱和时间成本过高
  - 使用Google搜索，可能遗漏知名度不高或非英语的研究结果
  - 大语言模型本身的幻觉问题

Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis

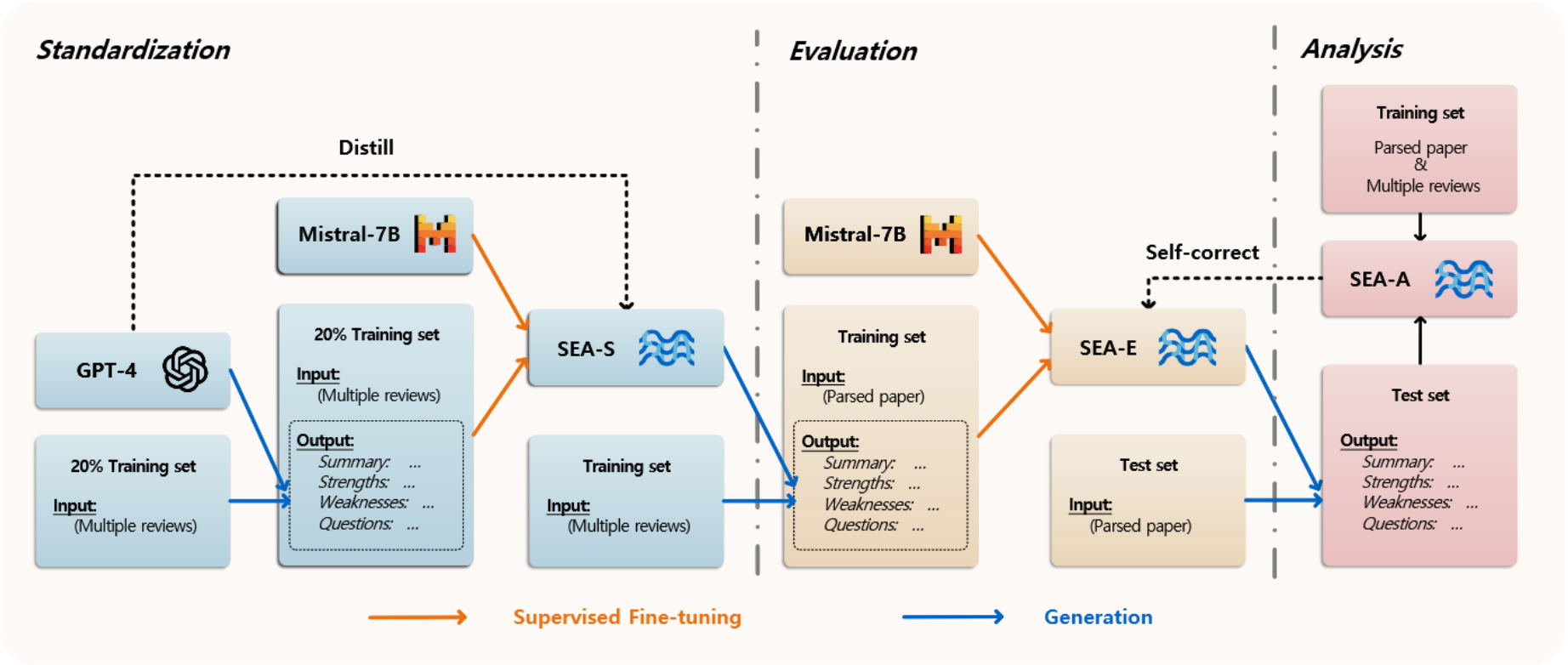
SEA TIPO

T	目标	针对学术论文生成高质量、高一致性的评审意见
I	输入	待评审的学术论文*1（12296篇学术论文）
P	处理	1.标准化同一篇论文的多条评审意见 2.使用整合后的数据监督微调大语言模型 3.引入不匹配分数衡量生成评审意见与论文内容的一致性，一致性差则重新生成评审意见
O	输出	具体的评审意见*1

P	问题	现有方法生成的评审意见通常是通用的或部分的
C	条件	模型需要有能力处理长上下文
D	难点	如何提高生成评审意见与学术论文的一致性
L	水平	EMNLP 2024（CCF B）

SEA 算法原理图

- 算法原理图
  - SEA-S：将一篇论文的多个原始审稿意见整合成统一格式的标准化评论
  - SEA-E：监督微调大语言模型，生成评审意见
  - SEA-A：引入不匹配分数，采用自我纠错策略重新生成评论



- 评审意见标准化
  - 评审数据集局限性
    - 每篇论文与多条评审意见相关联
    - 每条评审意见都根据审稿人的领域和专业知识提供有限的视角
    - 不同学术场所的评审格式和标准有所不同
  - 模型微调
    - 从训练集中随机选取20%的数据，使用GPT-4整合多条评审意见
    - 使用上述数据微调Mistral-7B模型，得到SEA-S模型
  - 使用SEA-S对完整数据集进行标准化

INSTRUCTION:

As an experienced academic paper reviewer, you are presented with different review contents for the same paper. Please analyze these contents carefully and consolidate them into a single review. The review should be organized into nine sections: Summary, Strengths, Weaknesses, Questions, Soundness, Presentation, Contribution, Rating and Paper Decision. Below is a description of each section:

1. Summary: Combine the 'Summary' sections from all reviews into a cohesive summary, aiming for a length of about 100-150 words.
2. Strengths/Weaknesses/Questions: Combine the Strengths/Weaknesses/Questions sections from all reviews into a unified, cohesive bullet-point list that avoids redundancy while preserving the specific details and depth of each point.
3. Soundness/Presentation/Contribution: Aggregate the Contribution/Soundness/Presentation score from each review to determine a suitable overall score (the score must be an **integer**), then, match this integer score to the corresponding criterion from the list below and provide the result. For example, if the score is 3, the result should be '3 good'. The possible scores and their criteria are:

   1 poor \n 2 fair \n 3 good \n 4 excellent

4. Rating: Aggregate the 'Rating' from each review to determine a suitable overall Rating (the Rating must be an **integer**), then, match this integer Rating to the corresponding criterion from the list below and provide the result. For example, if the Rating is 1, the result should be '1 strong reject'. The possible Ratings and their criteria are:

   1 strong reject
   2 reject, significant issues present
   3 reject, not good enough
   4 possibly reject, but has redeeming facets
   5 marginally below the acceptance threshold
   6 marginally above the acceptance threshold
   7 accept, but needs minor improvements
   8 accept, good paper
   9 strong accept, excellent work
   10 strong accept, should be highlighted at the conference

5. Paper Decision: It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on Meta-review, the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation, etc. Please ensure your Decision (Accept/Reject) matches the value of the 'Decision' key in the JSON, if present.

Here is the template for a review format, you must follow this format to output your review result:
Summary: \n <Summary content> \n

Strengths: \n <Strengths result> \n
Weaknesses: \n <Weaknesses result> \n
Questions: \n <Questions result> \n

Soundness: \n <Soundness result> \n
Presentation: \n <Presentation result> \n
Contribution: \n <Contribution result> \n
Rating: \n <Rating result> \n

Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content

- 不匹配分数
  - 给定某篇论文p，它有m条真实人工评审意见，对第j条评审意见有评分spr_i和对应的置信度cpr_j
  - 将置信度加权平均评分作为“群体参考值”，那么第i条评审意见的不匹配分数为该条评审意见的评分减去群体参考值：

$$y_{pr_i}^{true} = spr_i - \frac{\sum_{j=1}^{m} cpr_j \cdot spr_j}{\sum_{j=1}^{m} cpr_j}$$

$y_{pr_i}^{true}$的绝对值越大，代表该条评审意见的偏离程度越大，评审质量越低

- 回归模型训练
  - 给定某篇论文p和与之相对应的评审意见r，使用专为长上下文设计的预训练句子表示模型SFR-Embedding-Mistral将其转化为向量表示h_p和h_r 然后，分别计算二者的查询向量和键向量：

$$q_p = W^q h_p, \qquad q_r = W^q h_r$$
$$k_p = W^k h_p, \qquad k_r = W^k h_r$$

其中W^q和W^k是可学习的权重矩阵。

  - 我们计算的不匹配分数为：

$$y_{pred}^{pr} = \omega\left(q_p {k_r}^T + q_r {k_p}^T\right) + b$$

  - 使用均方误差为目标训练回归模型，得到SEA-A模型

SEA SEA-E

- 评审意见生成
  - 使用经SEA-S模型标准化后的数据集对Mistral-7B模型监督微调，得到SEA-E模型
  - 使用SEA-E模型生成评审意见

- 自我纠错策略
  - SEA-E模型完成评审意见生成后，使用SEA-A模型来评估生成评审意见和原论文的一致性
  - 当不匹配分数大于预设的阈值时，将当前的不匹配分数作为额外提示词重新生成评审意见

INSTRUCTION:

You are a highly experienced, conscientious, and fair academic reviewer. Please help me review this paper. The review should be organized into nine sections:

1. Summary: A summary of the paper in 100-150 words.
2. Strengths/Weaknesses/Questions: The Strengths/Weaknesses/Questions of paper, which should be listed in bullet points, with each point supported by specific examples from the article where possible.
3. Soundness/Contribution/Presentation: Rate the paper's Soundness/Contribution/Presentation, and match this score to the corresponding criterion from the list below and provide the result. The possible scores and their criteria are:
   1 poor
   2 fair
   3 good
   4 excellent
4. Rating: Give this paper an appropriate rating, match this rating to the corresponding criterion from the list below and provide the result. The possible Ratings and their criteria are:
   1 strong reject
   2 reject, significant issues present
   3 reject, not good enough
   4 possibly reject, but has redeeming facets
   5 marginally below the acceptance threshold
   6 marginally above the acceptance threshold
   7 accept, but needs minor improvements
   8 accept, good paper
   9 strong accept, excellent work
   10 strong accept, should be highlighted at the conference
5. Paper Decision: It must include the Decision itself (Accept or Reject) and the reasons for this decision which is based on the criteria of originality, methodological soundness, significance of results, and clarity and logic of presentation.

Here is the template for a review format, you must follow this format to output your review result:
Summary: \n <Summary content> \n

Strengths: \n <Strengths result> \n
Weaknesses: \n <Weaknesses result> \n
Questions: \n <Questions result> \n

Soundness: \n <Soundness result> \n
Presentation: \n <Presentation result> \n
Contribution: \n <Contribution result> \n
Rating: \n <Rating result> \n

Paper Decision:
- Decision: Accept/Reject
- Reasons: reasons content

Please ensure your feedback is objective and constructive. The paper is as follows: <paper content>

SEA 实验设置

- 数据集

	CONLL-16	ACL-17	COLING-20	ARR-22	NeurIPS-16-22	ICLR-17-23	NeurIPS-23	ICLR-24	Total
# papers	22	136	88	364	1,048	1,617	3,368	5,653	12,296
# tokens per paper	8,163	8,400	7,571	8,229	10,499	9,586	11,205	9,815	10,142
# reviews	39	272	112	684	3,847	5,779	15,027	21,839	47,602
# tokens per review	532	558	539	539	527	602	642	594	603
% accepted	50%	67%	93%	100%	97%	30%	95%	37%	60%
domain	NLP/CL	NLP/CL	NLP/CL	NLP/CL	ML	ML	ML	ML	multi

- 对比方法
  - Mistral-7B：直接使用大语言模型
  - 监督微调：分别采用随机选取、GPT-3.5、SEA-S三种方法对单篇论文的多条评审意见进行标准化，使用监督微调后的模型进行评审意见生成
- 评价指标
  - BLEU、ROUGE、BERTScore

SEA 对比实验

- 实验结果
  - SEA在所有测试场景中都优于其他基线模型

Method	BLEU	ROUGE (Recall)			ROUGE (F1-score)			BERTScore	Tokens
		R-1	R-2	R-L	R-1	R-2	R-L		
CONLL-16									
M-7B	18.92	20.81	4.81	10.30	28.66	6.81	14.18	82.49	554
M-7B-R	18.16	21.96	5.17	10.62	29.56	7.18	14.31	82.57	357
M-7B-3.5	19.70	26.51	5.58	13.96	30.19	6.45	15.37	82.01	627
SEA-E	29.07	34.91	7.79	15.29	38.64	8.67	16.73	82.85	793
SEA-EA	**31.01**	**36.96**	**8.91**	**16.34**	**40.49**	**9.68**	**17.57**	**82.94**	798
ACL-17									
M-7B	18.92	21.53	5.23	10.50	27.99	6.93	13.54	82.75	569
M-7B-R	18.15	21.84	5.19	10.76	27.71	6.87	13.55	82.56	357
M-7B-3.5	16.73	27.27	6.26	14.47	26.09	6.19	13.19	82.37	636
SEA-E	25.67	33.13	7.71	14.94	35.52	8.45	15.62	83.08	772
SEA-EA	**27.90**	**35.83**	**8.84**	**15.83**	**38.03**	**9.48**	**16.36**	**83.19**	806
COLING-20									
M-7B	21.97	29.11	6.42	14.80	31.91	7.01	15.83	82.76	579
M-7B-R	19.49	29.21	6.69	15.20	30.23	6.80	15.25	82.27	361
M-7B-3.5	18.13	34.03	7.56	18.43	28.49	6.10	14.77	82.12	617
SEA-E	22.93	40.62	9.23	20.05	34.37	7.65	16.15	82.85	774
SEA-EA	**24.85**	**42.97**	**10.57**	**20.89**	**36.67**	**8.76**	**16.96**	**83.09**	782
ARR-22									
M-7B	22.07	25.28	6.96	12.46	32.60	9.16	15.99	83.25	575
M-7B-R	20.27	24.89	6.70	12.60	31.22	8.66	15.71	82.70	357
M-7B-3.5	20.18	31.70	7.90	16.38	30.82	7.86	15.33	82.65	650
SEA-E	27.92	37.64	9.37	17.18	38.94	9.84	17.35	83.38	787
SEA-EA	**30.05**	**40.34**	**10.82**	**18.17**	**41.37**	**11.19**	**18.20**	**83.59**	818

Method	BLEU	ROUGE (Recall)			ROUGE (F1-score)			BERTScore	Tokens
		R-1	R-2	R-L	R-1	R-2	R-L		
NeurIPS-16-22									
M-7B	14.91	14.47	4.89	7.15	23.31	7.94	11.56	83.10	612
M-7B-R	13.94	14.47	4.79	7.29	22.70	7.67	11.44	82.73	362
M-7B-3.5	16.95	20.41	6.02	10.72	26.45	8.13	13.45	82.56	629
SEA-E	24.83	24.12	7.31	10.66	34.06	10.44	15.11	83.35	782
SEA-EA	**27.08**	**26.76**	**8.38**	**11.55**	**36.91**	**11.69**	**15.99**	**83.52**	838
ICLR-17-23									
M-7B	13.75	13.10	4.42	6.51	21.65	7.36	10.80	83.26	607
M-7B-R	12.98	13.38	4.45	6.85	21.36	7.26	10.91	82.80	359
M-7B-3.5	17.85	18.26	5.70	9.27	27.37	8.69	13.94	82.87	637
SEA-E	23.34	22.38	6.84	9.93	32.50	10.07	14.49	83.58	783
SEA-EA	**25.47**	**24.80**	**7.87**	**10.81**	**35.23**	**11.32**	**15.43**	**83.73**	841
NeurIPS-23									
M-7B	12.42	11.96	4.96	6.13	20.55	8.55	10.55	83.86	617
M-7B-R	11.92	11.88	4.87	6.16	20.14	8.31	10.49	83.44	366
M-7B-3.5	16.71	16.80	6.12	8.53	26.51	9.74	13.50	83.20	650
SEA-E	21.34	20.32	7.27	9.14	31.34	11.26	14.14	84.02	794
SEA-EA	**23.32**	**22.49**	**8.38**	**9.91**	**34.03**	**12.73**	**15.03**	**84.20**	844
ICLR-24									
M-7B	13.93	13.48	5.29	6.73	22.55	8.89	11.28	83.79	614
M-7B-R	13.91	14.17	5.41	7.21	22.94	8.85	11.69	83.81	380
M-7B-3.5	18.72	19.40	6.52	9.64	29.26	9.93	14.58	83.29	649
SEA-E	23.88	23.28	7.90	10.13	34.29	11.71	14.98	84.04	793
SEA-EA	**25.96**	**25.62**	**8.97**	**10.97**	**36.97**	**13.02**	**15.88**	**84.15**	852

SEA 算法总结

- 算法贡献
  - 提出了一个新的自动论文评审框架
  - 提出了一个构建标准化审稿数据集的新范式
  - 提出了一个新的评估指标来衡量论文和生成评审意见之间的一致性

- 算法不足
  - 仅停留在机器学习领域，尚未探索物理学或数学等其他学科领域
  - 没有在模型训练阶段使用基于分数的自然语言指导来增强评审意见生成模型
  - 缺乏作者对评审意见的反驳探索以纠正审稿人可能的误解

特点总结与未来展望

应用与展望

- 特点总结

算法	SWIF^{2}T	SEA
优势	1.聚焦于段落，可以生成具体连贯的评审意见 2.生成评审意见时附带推理步骤，可解释性强	1.标准化输入数据，数据集质量较高 2.采用自我纠错策略，生成评审意见与原始论文一致性更高
劣势	1.计算成本和时间成本过高 2.直接使用大语言模型，可能引发幻觉问题	领域泛化能力待验证

- 未来展望
  - 模型权重层面的进一步训练
  - 更宽广的学科领域
  - 时间和金钱成本的降低

参考文献

[1] Chamoun E, Schlichtkrull M, Vlachos A. Automated Focused Feedback Generation for Scientific Writing Assistance[C]. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 9742-9763

[2] Yu J, Ding Z, Tan J, Luo K, Weng Z, Gong C, Zeng L, Cui R, Han C, Sun Q, Wu Z, Lan Y, Li X. Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis[C]. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA: Association for Computational Linguistics, 2024: 10164-10184

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！