

# ALens: An Adaptive Domain-Oriented Abstract Writing Training Tool for Novice Researchers

Chen Cheng\*

Ziang Li\*

chengchen@shanghaitech.edu.cn

liza1@shanghaitech.edu.cn

School of Information Science and Technology, ShanghaiTech University

Shanghai, China

Zhenhui Peng

pengzhh29@mail.sysu.edu.cn

School of Artificial Intelligence, Sun

Yat-Sen University

Zhuhai, China

Quan Li†

liquan@shanghaitech.edu.cn

School of Information Science and

Technology, ShanghaiTech University

Shanghai Engineering Research

Center of Intelligent Vision and

Imaging

Shanghai, China

## ABSTRACT

The significance of novice researchers acquiring proficiency in writing abstracts has been extensively documented in the field of higher education, where they often encounter challenges in this process. Traditionally, students have been advised to enroll in writing training courses as a means to develop their abstract writing skills. Nevertheless, this approach frequently falls short in providing students with personalized and adaptable feedback on their abstract writing. To address this gap, we initially conducted a formative study to ascertain the user requirements for an abstract writing training tool. Subsequently, we proposed a domain-specific abstract writing training tool called *ALens*, which employs rhetorical structure parsing to identify key concepts, evaluates abstract drafts based on linguistic features, and employs visualization techniques to analyze the writing patterns of exemplary abstracts. A comparative user study involving an alternative abstract writing training tool has been conducted to demonstrate the efficacy of our approach.

## CCS CONCEPTS

- Human-centered computing → Human computer interaction (HCI); Visualization; User studies.

## KEYWORDS

Educational Applications, Writing Support Systems, Automated Feedback, Summarizing Learning

### ACM Reference Format:

Chen Cheng, Ziang Li, Zhenhui Peng, and Quan Li. 2018. ALens: An Adaptive Domain-Oriented Abstract Writing Training Tool for Novice Researchers. In *Proceedings of The 19<sup>th</sup> Joint Academic Conference on Harmonious Human-Machine Environment (HHME 2023)*. ACM, New York, NY, USA, 21 pages. <https://doi.org/XXXXXX.XXXXXXX>

\*Both authors contributed equally to this research.

†The corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

HHME 2023, August 25–27, 2023, Harbin, China

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXX.XXXXXXX>

## 1 INTRODUCTION

Paper writing is an essential skill that junior graduate students or researchers should master [111] because of its importance in learning, understanding, applying, and synthesizing new knowledge [35]. Basically, the structure of a typical research paper follows a pattern known as “King Model” [32], which delineates the thematic progression of an article through six sections: *title*, *abstract*, *introduction*, *body*, *discussion*, and *references* [32]. Among the major components of academic papers, abstracts, which usually consist of separate paragraphs outlining the content of the paper [89], have become increasingly important. For example, with the boom in search engines and bibliographic databases, the title and abstract are often the only two parts of a research paper that a potential reader can freely view, while access to the full paper may be subject to charges to the copyright owner [88]. In addition, when researchers conduct systematic investigations of related work, they have to spend time reading the full manuscripts if the corresponding abstracts are obscure, so they may abandon researching them [3, 89]. In addition, during the blind review process, editors use abstracts to invite appropriate reviewers with expertise in the relevant field to evaluate papers [3, 88].

Concerns about the academic abstract writing skills of undergraduate and graduate students in higher education are well documented [25, 38, 86]. From a faculty member’s perspective, writing well is more than just following writing conventions. It also involves creative inspiration, problem-solving, reflection, and editing, culminating in a complete manuscript [25, 57]. From a student’s perspective, writing an abstract can be a daunting task, both in terms of getting ideas on paper and mastering writing rules such as *logic*, *summary*, *argument*, and *grammar* [25, 40]. To help students develop the abstract writing skills typically included in paper writing skills, institutions, such as universities, have conventionally recommended that students attend thematic writing training courses, such as scientific paper writing and biology essay writing, during which it is important for individual students to receive ongoing formative feedback [16]. However, the need to provide optimal formative feedback on individual abstract writing training in traditional large-scale lectures is often hampered by limited financial and pedagogical resources. One possible solution for providing individual feedback is to take advantage of recent advances in Natural Language Processing (NLP) and Machine learning (ML).

We systematically reviewed the literature on abstract writing in the field of educational technology following the rigorous approach

suggested by Brocke et al. [97]. However, we found that the existing literature is under-researched in terms of **academic abstract writing** training. In contrast, a considerable number of tools have been developed to improve students' **summary writing** skills. It should be noted that abstracts and summaries are different<sup>123</sup>. While there are nuances to various accounts of the difference between an abstract and a summary, the general perception is that a summary of an entire article is a more detailed version of an abstract and that an abstract is usually written in the order of the content of a research paper, while a summary may focus on important aspects of the article<sup>45</sup>. Despite the differences, it is important to acknowledge that both contain important content and require students to have the ability to condense information.

We borrow the research ideas of summary writing training, from which we can conclude that the systems or methods proposed in computer-assisted summary writing training usually involve a three-stage cycle, namely: *reading*, *writing*, and *feedback*. First, reading and understanding the main idea of the source text is critical. Previous work [85] used concept maps to help students identify the main ideas and understand their hierarchy. While it may be suitable for general summary writing training, it is not appropriate in the scenario of academic abstract writing training because the concept maps in [85] need to be generated by consultants and experts, which is too labor-intensive, especially when it comes to academic papers. In order to annotate concept maps for papers in different fields, a large number of experts are needed, as academic terminology and writing styles vary widely. Even in the same discipline, such as human-computer interaction (HCI), abstracts are written differently because of the different types of papers; for example, an application paper and a survey paper typically have different abstract writing styles. In terms of writing, existing summary training tools provide paradigms and summary strategies to instruct writing, which are good guides for abstract writing training. For the last stage, according to the literature review, there are four types of feedback, viz. *providing scores* [24, 28, 46, 98], *peer review* [104], *section content coverage* [24, 28, 98] and *summary writing strategy detection* [1, 28, 51]. However, for the same reasons as concept maps, the coverage of chapter content that requires instructor annotation does not apply to academic writing scenarios. To the best of our knowledge, there are no principles and proofs in the current literature on how to design automatically adaptive computer-assisted academic abstract writing tools to help a researcher learn abstract writing styles and patterns in his/her field.

To clarify the current status and main concerns of the abstract writing and training process for academic papers, we first systematically reviewed the literature in the field of pedagogy and educational technology [97]. Then, we investigated the pain points of L2 (second language) junior researchers when writing abstracts through a formative study (a survey of 164 students and semi-structured interviews with 11 students). We aimed to address the three following research questions: 1) **RQ1: What are current student**

**practices when writing abstracts?** 2) **RQ2: What are the specific challenges students have when writing abstracts?** and 3) **RQ3: What kind of support do students need when writing abstracts?** For **RQ1**, we learned that most L2 junior researchers write abstracts at least after completing the introduction part. For **RQ2**, we found that all the potential assistance we thought based on the literature review and summary writing tools were acknowledged by the participants. For **RQ3**, we extracted four main barriers that learners face when writing abstracts, namely: *lack of skills in rephrasing content*, *organization of ideas*, *identification of main ideas*, and *writing style recognition*. First, rephrasing is considered to be one of the core skills for paraphrasing key content, which is the essence of abstract writing [9]. However, students, especially L2 learners, resort to copying sentences from other parts of the paper rather than rewriting the main ideas in their own words [9, 40]. Second, when it comes to organizing the ideas in each section of the paper, most junior students are not skilled at integrating them in a logical and cohesive manner while making the essay fluent and clear [9]. Third, despite the ability of novice researchers to identify the topic of the essay, secondary and irrelevant information remained easily incorporated, meaning they are deficient in grasping the complete hierarchy of ideas in the text [25]. Fourth, 73% of students mentioned in their interviews that it would have been better to show the writing style or at least give them some hints. It can be quite time-consuming for them to align abstract ideas with lengthy original texts.

To address the above issues and fill the gap in abstract writing training, as well as to take advantage of recent advances in NLP technology, we propose a domain-oriented abstract writing training system *ALens* (abbreviation for Abstract Lens), an adaptive learning tool that uses rhetorical structure parsing to identify main ideas, evaluates their abstracts from different linguistic features and uses visualization to analyze the writing patterns of reference abstracts (i.e., ground truth abstracts). Specifically, to address the first challenge and train users in their paraphrasing, we incorporate linguistic features, such as lexical and syntactic complexity, as assessment metrics. To address the idea organization problem, we run a re-trained sentence classification model that classifies abstract sentences into five genres (i.e. background, objective, method, result, and conclusion) [20, 34, 40, 52] and show the results in different colors. Considering the classification feedback, self-regulation [5] regarding the organization of the abstract will be evoked, which will lead the user to discover which parts of the paper need to be included in specific areas and whether the ideas are expressed in a logical and cohesive order. To address the third challenge, we use discourse parsing with Rhetorical Structure Theory (RST) [47, 69] to construct RST segments from the perspective of identifying logical relationships in the introduction. It separates sentence groups into RST trees with phrases on leaf nodes and logical relations on branches. RST uses rhetorical relations (e.g., elaboration, contrast, etc.) to depict the structure and logic of various parts of the text [70]. By parsing different paragraphs or the whole introduction, users can obtain the hierarchical structure of the text at different granularities and grasp the hierarchy of ideas in the text. Finally, to solve the last issue, following the approach utilized in the works about attention [12, 83, 94, 95], we attempt to find relevant tokens in the generated abstract from the source text, apply semantic similarity

<sup>1</sup><https://smartleadershiphut.com/writing/abstract-vs-summary/>

<sup>2</sup><https://www.scribbr.com/frequently-asked-questions/abstract-vs-summary/>

<sup>3</sup><https://www.mimjournal.com/post/main-differences-between-a-summary-and-an-abstract>

<sup>4</sup><https://smartleadershiphut.com/writing/abstract-vs-summary/>

<sup>5</sup><https://www.scribbr.com/frequently-asked-questions/abstract-vs-summary/>

to align the ideas between the reference abstract and the source text and reconstruct the style used by the authors in writing the reference abstract. In addition, about 27% told us in the interviews that when they have some ideas to write about but do not know where to start, they may get stuck. To facilitate the writing of the first draft, we embed a summary model as an option in the system [42] to generate an initial draft as a prompt to start.

With the proposed research prototype, we further explored the following two research questions: **RQ4: What is the technology acceptance level among junior researchers?** and **RQ5: How effective is ALens in helping users write abstracts compared to the baseline system?** To answer these questions, we demonstrate the impact of *ALens* on users' abstract writing skills by evaluating our system in two writing training scenarios. We quantitatively compare an abstract writing training method with our system. In a user study with 21 students, the results show that with the help of *ALens*, users could organize their content in a more appropriate style when writing abstracts than the alternative tool. In addition, we measure the technology acceptance, user satisfaction, and engagement of both tools using the key constructs [92, 93], and the results are encouraging, suggesting that *ALens* can motivate students to learn abstract writing patterns in their own domain and to write abstracts in an appropriate style. Taken together, the main contributions of this work are:

- We conduct a formative study to understand the problems encountered by L2 junior researchers in the academic abstract writing process.
- We build *ALens*, an automatic feedback learning tool that first incorporates visualization and interactive features into academic abstract writing training.
- We show the effectiveness of *ALens* by comparing it with an alternative abstract writing training tool.

## 2 RELATED WORK

The literature that overlaps with this work can be grouped into four categories, namely, *technology-mediated summary writing assistance*, *summary evaluation metrics*, *NLP models in the summary task*, and *self-regulated learning*.

### 2.1 Technology-Mediated Summary Writing Assistance

We systematically reviewed the literature on abstract writing in the field of educational technology following the rigorous approach suggested by Brocke et al. [97]. However, although several tools have been developed to improve students' summary writing skills over the past decade, very little literature has focused on the development of learning tools for abstract writing. The main difference between an abstract and a summary of a whole article is the length and purpose<sup>678</sup>. Abstracts usually follow the empirical order of content as specified by the journal or association and cover the main aspects of the research paper. Summaries may not follow specific guidelines, emphasizing certain important aspects of the paper and

providing more details than the abstract. Despite the differences, it has to be acknowledged that both are abbreviated versions of the paper that contain important content and require the ability to understand, express, synthesize and paraphrase [18, 19, 82]. For example, in a study of computer-assisted summary writing training, the **concept map** [24] arranges concepts in the text in layers, with general concepts at a shallower level and specific concepts at a deeper level. It attempts to facilitate students' identification of the main ideas and understanding of the corresponding supporting ideas. Several studies have proposed methods that identify the **summarization strategies**, including deletion, sentence combination, and paraphrasing used by students to help assess teachers' summarization processes and to target them during training. **Worked examples** [24, 46, 98] are exemplars with worked-out steps and predetermined questions and are often used as guides to help students learn to read the original text and summarization strategies. In addition, by comparing multiple worked examples, students gain the ability to identify patterns of relevant and irrelevant information [24]. The **Computer-Supported Collaborative Learning (CSCL) approach** [104] is embedded in the summary writing training system, and students receive peer feedback through online conversations and interactions. As they digest peer feedback, students reflect on their summarization process and make further revisions [49].

However, the above approach cannot be directly applied to academic abstract writing training. Specifically, concept maps and worked examples are carefully prepared by instructors and need to be annotated article by article due to the different topics and progression of the articles. In other words, when it comes to academic papers, the workload of instructors in generating concept maps and worked examples can be very high. To fill the gap in abstract writing training and to take advantage of recent advances in NLP technology, we use rhetorical structure parsing to identify main ideas, evaluate abstracts in terms of different linguistic features, and use visualization to analyze the writing patterns of reference abstracts.

### 2.2 NLP Models in Summary Tasks

The text processing models behind text summarization tools can be broadly classified into two categories, namely *extraction* and *abstraction* [36]. Extractive approaches [54, 66, 72, 73, 81, 110] **copy salient phrases and sentences from the text and merge them to create summaries** [73, 110], thus ensuring that the summaries are factually consistent with the source text [21]. However, the extraction paradigm is often criticized for being logically inconsistent with the input text [80, 81]. Abstraction methods [42, 64, 102, 107] **rearrange the language in the text and add new words or phrases to the abstract as needed** [43]. Since state-of-the-art abstraction methods perform well in generating fluent human-like summaries [107], in our work we embed the abstraction summarization models into our system [42] as an option for generating the initial manuscript to prompt the user to start.

### 2.3 Summary Evaluation Metrics

In the learning process, it is important to provide individual and adaptive feedback [16], and the same is true for abstract writing

<sup>6</sup><https://smartleadershiphut.com/writing/abstract-vs-summary/>

<sup>7</sup><https://www.scribbr.com/frequently-asked-questions/abstract-vs-summary/>

<sup>8</sup><https://www.mimjournal.com/post/main-differences-between-a-summary-and-an-abstract>

training. We consider assessment methods widely used in summary writing training as a potential approach to abstract writing training. For example, assessment scores are a typical method of providing formative feedback in computer-assisted abstract writing training [24, 85, 98], and there are three types of scores, i.e., content coverage scores [24, 98], scores given by mathematical methods [28, 59, 65, 108] and scores predicted by pre-trained language models [17, 68, 103]. Specifically, content coverage scores are calculated automatically to measure the degree of coverage of each content in the summary. Although calculated automatically, the exact content to be measured is specified by the instructor on an article-by-article basis. However, this is clearly not appropriate for academic abstract writing, as the differences in disciplines, fields, and paper types result in a significant amount of work for instructors to develop content criteria for each type of article. Instead, mathematical methods and deep learning approaches are the most suitable candidates. Although pre-trained deep learning language models can achieve a high degree of agreement with human estimates, their high performance is highly dependent on the availability of relevant datasets. Due to the unavailability of high-quality datasets, we turn to mathematical methods. Specifically, three methods are commonly used: metrics in ML [65, 74, 108], latent semantic analysis (LSA) [63] and linguistic features [29, 60, 61]. Metrics in ML, e.g., ROUGE [65], BERTScore [108] and Bleu [74] and LSA all assess the quality of a summary based on semantic overlap with the reference or source text, thus giving an overall score for the summary. However, this single score is not an appropriate feedback [101], and it does not reveal the gap between what people understand and what they should understand [77]. Therefore, we rate the summaries using different scoring criteria (e.g. lexical complexity and cohesion) based on linguistic features, which is considered more appropriate because it captures different aspects of the summaries and thus provides more informative and instructive feedback [14].

## 2.4 Self-regulated Learning

It has been hypothesized that providing students with feedback about their writing abilities will enhance their learning experience and facilitate the writing of high-quality summaries [112]. In order to achieve self-regulated learning, providing students with formative feedback as well as setting goals is essential [11]. It has been argued that in order for feedback systems to be effective, learners must be provided with goals, their progress tracked, and actions identified to help them achieve those goals [45]. However, individuals are unable to track their own progress in the work [15]. Using targeted assessment and feedback is a good way to enhance the learning process [76]. When students are given feedback on their abilities throughout the intervention, it can increase their chances of achieving better short-term outcomes on specific learning tasks [7, 45, 76]. In this work, we provide students with user-centered adaptive feedback about their abstracts to determine if they can write and improve organized abstracts.

## 3 FORMATIVE STUDY

Based on the similarities between summary writing and abstract writing in terms of the writing process and required competencies,

and in order to fill the gap in abstract writing training tools, we adopt a top-down approach, first distilling possible meta-requirements from the existing literature on summary writing training tools. With this goal in mind, we first selected 27 papers discussing summary writing training for meticulous analysis, from which we distilled the closed loop of summary writing learning. In addition, because abstract writing aids span the fields of education, psychology, and computer science, we focused on literature in these categories. On this basis, additional 32 related papers were selected to further analyze and understand established pedagogical theories in writing [13] and metacognition [67] in the learning process, which is considered a meta-need for adaptive learning tools.

Next, in order to derive the user requirements for the academic abstract writing training system, we first need to understand the problems that students encounter in the academic writing process. Therefore, we design and examine the following three research questions: 1) **RQ1: What are current student practices when writing abstracts?** 2) **RQ2: What are the specific challenges students have when writing abstracts?** and 3) **RQ3: What kind of support do students need when writing abstracts?**

### 3.1 Survey Study

**3.1.1 Survey Protocol.** The survey was administered on the Microsoft Forms online platform. The survey questions included abstract writing practices, difficulties in abstract writing, assistance needed when writing abstracts, and demographic questions. The survey contained textual questions and ranking questions about the academic paper writing process. It also contained questions about students' challenges. 5-point Likert scale questions are used to measure students' attitudes toward several potential types of assistance. The survey also contained open-ended questions about student requests. At the end of the survey, respondents were allowed to leave their contact information if they wished to be interviewed for follow-up.

**3.1.2 Respondents and Recruitment.** We recruited 164 respondents (54 female, 106 male, and 4 prefer not to specify) between the ages of 19 and 36 (B.S.: 105, M.S.: 42, Ph.D.: 13, Others: 4) via advertised posts in online university communities. Of all respondents, 125 with academic writing experience answered Q1 – Q3 and the other 39 answered Q2 – Q3.

### 3.2 Interview Study

**3.2.1 Interviewees.** To gain more insight into the challenges and requirements of students when writing their abstracts, we further contacted 11 students (3 female, 8 male; 10 graduate students, 1 PhD. student) who left their email addresses in their questionnaire. Their ages ranged from 20 to 26 (mean age = 23.09, SD = 1.81).

**3.2.2 Interview Protocol and Analysis Method.** We conducted remote semi-structured interviews using an online communication tool and audio-recorded the interviews with consent. The interviews consisted of three main sections: (1) how students typically write abstracts; (2) the challenges in writing abstracts; and (3) what features students need. To analyze the challenges students encounter when writing abstracts for academic papers, we followed

Had academic writing experience				Potential assistance	M	SD
Yes	No			Online revision	4.24	0.88
125/164	39/164		/	Present key information of intro	4.22	0.69
				Word count	4.21	0.35
				Present abstract structure	4.17	0.84
				Abstract writing style recognition	4.10	0.88
				Abstract evaluation	4.09	0.86
				Support intro annotation	4.08	0.83
				Present logic relation of sentences	4.06	0.88
				Present intro structure	4.06	0.88
				Instructional feedback	4.03	0.91

TABLE 1: Results of the survey. On the left is the distribution of the number of distinct respondents; on the right is a tally of 5-point Likert scale questions about potential help (1 – 5: very unhelpful – very helpful).

an iterative coding process [50] for thematic analysis. For each question, one author open-coded the responses to identify the categories that appeared and developed a codebook. We noted that a single response may include multiple categories. Therefore, we treated each category as binary. For each response, we labeled whether each category was present or absent. Two coders coded all responses independently. They then discussed inconsistencies, refined the code definitions, and independently re-coded the responses based on the new definitions. They iteratively coded the responses until they reached a Cohen's kappa above 0.7 for all categories. Finally, we came up with five subcodes for the students' challenges.

### 3.3 Findings and Design Requirements

For RQ1, We recapitulated the following findings from the survey and interview results. Literature [56, 99] and reports<sup>9</sup> indicate that "...the Abstract must be written after completing the entire manuscript. Ensure that important points made in the main manuscript are included in the abstract...". We also found that most learners (survey: 139/164, interview: 10/11) wrote the abstract after writing the introduction or body of the paper. They usually first determined the structure of the abstract, i.e., what sections need to be included and which are more important. Then they wrote down each part purposefully. And most learners (survey: 119/164, interview: 8/11) would refer to the introduction or body of the paper to ensure consistency. For example, after reading the introduction, some of them would extract important sentences or paragraphs by highlighting these texts and reorganizing these texts into an abstract. More than half of them would use writing aids such as Grammarly<sup>10</sup>. For RQ2, The detailed results of the survey are shown in Table 1. The result shows that all the potential assistance we refined from the summary writing training tool and pedagogical theories are verified by our respondents. The main findings for RQ3 are summarized in subsubsection 3.3.1.

**3.3.1 Challenges of Academic Abstract Writing.** We combined our survey research and interview study to present the following five challenges.

**C1: Lack of skills in rephrasing content (N=7/11).** Sometimes students tend to rewrite key sentences in the introduction,

especially when writing background and conclusion sentences (P1, P3 – P7). However, rephrasing content is sometimes tricky because students cannot directly use sentences from the introduction. Current summarization techniques produce wording that is still too close to the original text, which does not help solve this problem. "How to express the same meaning precisely in a new way is sometimes an annoying problem (P2, male, age=24)."

**C2: Identification and organization of ideas (N=9/11).** On the one hand, according to the survey results, almost half of the respondents (47.6% of the 164 participants) answered that it was quite difficult to summarize all the key points in a limited space, or to write them concisely enough. On the other hand, having a good insight into the logic of the introduction is crucial for students to write excellent abstracts. According to the survey results, most students (82.96% of 164 participants) wrote their abstracts after writing the main body of the paper. However, students may forget the logical flow of the introduction after writing the main body of the paper, especially sections with complex logical relationships. For example, "in my field, the prior experiments section in the introduction includes too many experimental methods. When writing the abstract, I always need to figure out again how they relate to each other (P9, male, age=21)."

**C3: Recognition of writing pattern and style (N=8/11).** Abstract writing is often field-oriented, as different disciplines and different types of papers, and different journals and conferences generally differ in style and writing patterns. The survey found that 53.0% of 164 respondents found it time-consuming to master the style abstracts and find their regularity by perusing articles in the field, and 65.9% thought it would be better if they were shown the regularity.

**C4: Requirement for the first draft.** 3 out of 11 students responded in the formative interview that they did not know where to start writing. "I'm used to revising from other people's drafts and I can't start from a completely blank space (P7, female, age=22)."

**3.3.2 Design Requirements.** Based on the identified challenges in academic paper abstract writing and users' expectations for satisfactory assistance results and comprehensive functionality, we derive the following design requirements of an adaptive abstract writing training tool.

**R1: Provide assistance on rephrasing.** Through the formative study, we found that how to rewrite key sentences and other

<sup>9</sup><https://writingcenter.gmu.edu/writing-resources/different-genres/writing-an-abstract>

<sup>10</sup><https://app.grammarly.com>

information extracted from the introduction is a great challenge for learners (C1). To address this issue, we should provide guidance on representing information in another way, such as sentence transformation and phrase substitution based on self-regulated learning theory [11].

**R2: Help learners better understand the introduction and organize the main ideas.** From the previous formative study, we found that many learners encountered difficulties in selecting core information in a limited space, i.e., the problem of main idea identification (C2). To effectively address this problem, learners' mastery of the structure and content of the introduction is exceptionally demanding. On the one hand, the introduction is a distillation of the main text, and the relationship between some sentences is difficult to grasp. Therefore, we should provide guidance on identifying the logical relationships of sentences in the introduction. On the other hand, the introduction is long and requires extra time to reread because the content is forgotten. We should also help learners quickly review the structure and content of the introduction. After understanding the introduction, some learners still have difficulty organizing these key elements fluently (C2). We should help them to have a better understanding of the information from a new perspective and help them to organize the main ideas in a rational way.

**R3: Assist learners in understanding domain-specific abstract styles.** As shown in Table 1, most participants indicated that knowledge about how abstracts are written was very useful to them (Mean = 4.10, SD = 0.88) (C4). The underlying style can guide learners to write abstracts that are more accurate in content and organization. Therefore, in addition to directly presenting the reference abstract, we should also demonstrate its style in a clear and intuitive manner (C3).

**R4: Prepare for the first draft.** In the semi-structured interviews, 3 students mentioned the difficulty of writing abstracts from scratch. Despite the relatively low rate, we observed the rise of text summarization platforms such as *TLDR this*<sup>11</sup>, *Resoomer*<sup>12</sup>, and *Wordtune Read*<sup>13</sup>. Therefore, we believe that there is a trend to harness the power of NLP techniques to facilitate abstract writing, for example, to generate the first draft version (C4).

**R5: Easy to access and use.** From the survey, some potential users expressed concerns about the complexity and difficulty of using the features of the general academic abstract writing assistance system. Therefore, we had to ensure that the tool would not become burdensome and responsive to users while providing practical features to address the above challenges. For example, users did not need to install additional software or hardware, using the typical writing assistance platform interface design familiar to learners. Despite the above requirements, basic functionality is also highly valued, as shown in Table 1. Therefore, the tool also needs to have the basic features of a writing aid to ensure effectiveness, such as online revision and word count functions.

## 4 DESIGN OF THE ABSTRACT WRITING TRAINING SYSTEM

### 4.1 Approach Overview

Based on the requirements derived from the formative study, we design an abstract writing training process and incorporate it into a web-based writing assistance platform named *ALens*. It facilitates users to quickly grasp the main ideas of an essay, optionally write using a summarization model from NLP, recognize their deficiencies in abstract writing, and gain knowledge about style in specific scenarios. To support the writing process in a convenient and user-friendly manner (R5), visualization and interaction are integrated into *ALens* to cater to the mental habits of users with different granularity requirements. *ALens* consists of a *Rhetorical Structure View*, a *Writing Area*, an *Evaluation Dashboard*, and a *Reference Abstract with a Flow Map*. Figure 1 describes the general stages of the designed abstract training pipeline. First, the user can select an article to be learned for abstract writing and upload it. Subsequently, the rhetorical structure of the original article is analyzed to help the user quickly identify the main ideas in terms of logical structure (R2). Then, the user can choose to write the first draft from scratch or with the help of a summarization model (R4). Given the lack of content organization, the sentences in the abstract are divided into several types (e.g. background and conclusion) [4, 27, 71], and the completeness of the abstract is checked against the domain of the paper, i.e., whether the first draft properly covers and arranges the domain typical of the abstract's required information and guide users to reflect on them (R2). Meanwhile, *ALens* can automatically analyze the linguistic features of the abstract to check whether it is comprehensible, concise, fluent, and consistent with the source text (R1). In addition, paraphrase detection is applied to the feedback to guide the user to rephrase sentences instead of copying them. Finally, users can check the writing style of the reference abstract and analyze its linguistic features and organization. Specifically, a flow map is used to align ideas in the reference abstract with the source text, that is, to find the most relevant content from the source text. By comparing these features of different articles with similar academic domain "styles", users are expected to discover writing patterns and learn writing styles (R3).

### 4.2 NLP Pipeline

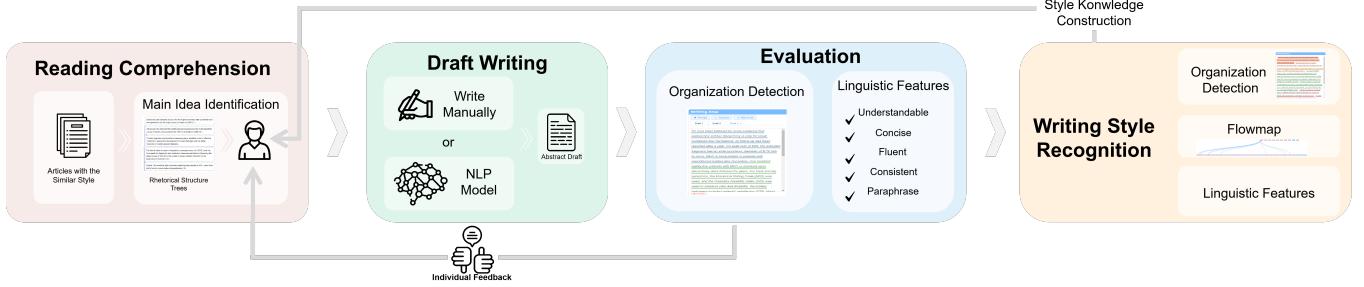
The back-end engine of *ALens* first helps users to identify ideas by parsing an article into a rhetorical tree, and then supports the production of the initial draft for later revision. The sentences in the abstract are then divided into several genres to detect the organization of the abstract. At the same time, the abstract is evaluated by different linguistic features, providing personal feedback to stimulate revision according to a self-regulated learning theory [11].

**4.2.1 Main Idea Identification.** To enable learners to quickly grasp the hierarchical structure of a text and avoid wasting time by repeatedly reviewing the introduction when writing an abstract, we provide a rhetorical structure parsing for each paragraph. In particular, we provide rhetorical relationship recognition for any text with a continuous span. For example, after inputting a text containing three sentences with six elementary discourse units (EDUs: tokens of adjacent text spans, roughly analogous to independent phrases)

<sup>11</sup><https://tldrthis.com/>

<sup>12</sup><https://resoomer.com/en/>

<sup>13</sup><https://app.wordtune.com/read>

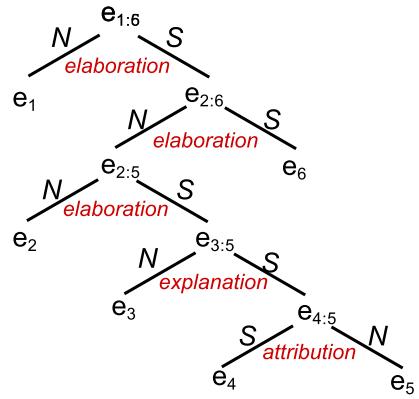


**Figure 1: Pipeline of ALens:** (1) read and comprehend the source text (2) generate the first abstract draft for later revision; (3) refine the abstract draft considering organization, main idea, and lexicon; evaluate the results based on semantic analysis; (4) recognize and learn the style of the reference abstract.

( $e_1 - e_6$ ), the model can output relations for any continuous EDUs. As shown in Figure 2,  $e_2 - e_6$  is an elaboration of  $e_1$ ,  $e_4 - e_5$  is an explanation of  $e_3$ ,  $e_5$  is attributed to  $e_4$ , and so on. All relations and their hierarchy build the structure of the whole text and give a deeper understanding of the text structure.

The identification of rhetorical relations consists of two parts: 1) text segmentation (TS) and 2) relation identification (RI). For text segmentation, we need to segment the input text into EDUs. Each sentence consists of several EDUs. We retrained the model proposed by Heilman et al. [47] to operate the segmentation task with high accuracy. It uses a conditional random field (CRF) model with  $l_2$  regularization[39]. For each token in a sentence, it predicts the whether it is the beginning of a new EDU or not. The CRF regularization parameter was 64.0, adjusted by grid search using a grid of powers of 2 between 1/64 to 64. Compared to the human agreement (HA)[10], the percentages of precision, recall and F1 score were 90.2(TS)/98.5(HA), 83.5(TS)/98.2(HA), and 86.7(TS)/98.3(HA), respectively. The second relation recognition component was modeled as a classification problem, i.e., classifying the relation of two consecutive EDUs into 16-tuple types such as elaboration, contrast, and joint. For this task, we utilized the ZPar model [109] as the relationship parser, which predicts the rhetorical relationships across the text at different levels of granularity. The output of the model is the relationship of EDUs, i.e., the phraseological relationships in a sentence, which is not valuable for abstract writing. Therefore, we modified and retrained the ZPar model to predict rhetorical relations between adjacent sentences at a more appropriate level of granularity. The parser was estimated using multiclass logistic regression with an  $l_1$  penalty.  $l_1$  was adjusted after the grid search and finally set to 0.25. Compared to the human agreement (HA)[53], the parser performs 83.5(RI)/88.7(HA), 68.1(RI)/77.7(HA), and 55.1(RI)/65.8(HA) in terms of span, kernel, and relationship, respectively. The parser model uses a shift-reduced algorithm with a time complexity of  $O(n)$ , where  $n$  is the number of EDUs.

**4.2.2 Summarization Assistance.** It is important to note that providing a relatively reliable first draft of the abstract is an important foundation for the later revision process. However, a requirement for developing NLP models with the ability to summarize research papers is the availability of relevant datasets. We reviewed the literature on corpora and found that the *arXiv* and *PubMed* datasets released in [27] met our requirements. *PubMed* contains 119k article body and abstract pairs of the biomedical literature, while *arXiv*



**Figure 2:**  $e_1$  [ Compare the past eight five-year plans with actual appropriations. ]  $e_2$  [ The Pentagon’s strategists produce budgets ]  $e_3$  [ that simply cannot be executed ]  $e_4$  [because they assume]  $e_5$  [ a defense strategy depends only on goals and threats. ]  $e_6$  [ Strategy, however, is about possibilities, not hopes and dreams. ]  
An example of an RST structure tree from the RST discourse tree bank[22].  $e_i$ ,  $e_{j,k}$ ,  $N$ , and  $S$  denote basic discourse units, spans, nucleus, and satellite, respectively.

contains 203k pairs of articles on different topics. Since different subjects have different terminologies and writing styles, we decided to construct a subset called *arXiv-cs* that contains only computer science (cs) articles by iterating through the abstracts in *arXiv* and match them in *arXiv-dataset* released in [26], which hosts 1.5M metadata of preprinted articles in physics, mathematics and computer science from 1991 to 2019 to determine whether they belong to the computer science domain.

Based on these two datasets (i.e., *arXiv-cs* and *PubMed* datasets consisting of article body and abstract pairs), we developed the following fine-tuning scheme. Since the two datasets belong to the biomedicine and computer science domains, respectively, we developed an abstraction summarization model based on *LongT5-large*<sup>14</sup> [42] for each domain. The training parameters follow those described in the original publication and GitHub; in addition, the maximum input token and maximum output token are set to (4096, 512) on *PubMed* and (16384, 512) on *arXiv-cs*, and the model performance

<sup>14</sup><https://github.com/google-research/longt5>

PubMed				arXiv-cs			
Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
47.34	22.53	28.74	42.79	46.91	19.67	27.41	41.87

TABLE 2: Results of fine-tuning *LongT5* on two datasets in biomedicine and computer science.

Facets of Quality	Linguistic Features	Definition
Understandability	Frequency (COCA spoken, All Words)	The sum of frequency scores of all words occurs in COCA spoken corpus divided by number of words in text with frequency score
	Frequency (COCA spoken, Function Words)	The sum of frequency scores of function words occurs in COCA spoken corpus divided by number of words in text with frequency score
	Frequency (SUBTLEXus, Content Words)	The sum of frequency scores of content words occurs in SUBTLEXus corpus divided by number of words in text with frequency score
	Frequency (SUBTLEXus, All Words)	The sum of frequency scores of all words occurs in SUBTLEXus corpus divided by number of words in text with frequency score
Consistency	Source similarity (ROUGE-3)	Similarity to the source, calculated by ROUGE-3
Fluency	Adjacent sentence similarity (word2vec)	Similarity between two adjacent sentences, calculated by word2vec
	Repeated content lemmas and pronouns	Number of repeated content and third person pronouns divided by number of words
	Binary adjacent sentence overlap (Function Words)	Number of overlapping function words in two adjacent sentences
Diversity	Type-token ratio (All Words)	The number of unique words (types) divided by the total number of words (tokens) in a given segment of language
	MATTR (Function Words)	Moving average type token ratio for function words (50-word window)
	Number of Content Words tokens	Number of Content Words tokens
	MTLD (Function Words)	Average number of function word tokens it takes to reach a given TTR value (.720)
	MTLD (All Words)	Average number of tokens it takes to reach a given TTR value (.720)
	MTLD (Content Words)	Average number of content word tokens it takes to reach a given TTR value (.720)
	Lexical density (Percentage of Content Words)	Percentage of Content Words in the text
	Type-token ratio (Content Words)	The number of unique content words (types) divided by the total number of words (tokens) in a given segment of language
	SD of dependents per nominal subject	Standard deviation of dependents per nominal subject
	SD of dependents per clause	Standard deviation of dependents per clause
Conciseness	SD of dependents per object of the preposition	Standard deviation of dependents per object of the preposition
	Mean length of sentence	Mean length of sentence
	Mean length of clause	Mean length of clause
	Word counts	Word counts

TABLE 3: The table shows the five facets of abstract quality and the corresponding linguistic features used to calculate each facet. For specific definitions of linguistic features, please refer to [29, 60–62].

for both fine-tunings is shown in Table 2. Note that Rouge (Recall-Oriented Understudy for Gisting Evaluation) [65] is a set of metrics for evaluating automated abstracts and machine translations. It measures the “similarity” between an automatically generated abstract or translation and a reference abstract by comparing it with a set of reference abstracts (usually manually generated) and calculating the corresponding score. The results show that these models can be used to generate a relatively reliable abstract for later revisions. Please refer to Table 8 in Appendix A for the hyperparameters we set in this work for details.

**4.2.3 Organization Detection.** To facilitate the organization of the ideas distilled from the source text, we detected the organization in the abstract, thus inducing the user to think about the coverage and arrangement of the content. The organization detection task is considered as a multi-class sentence classification task, where each sentence in the abstract is classified into five types, i.e., *background*, *objectives*, *methods*, *results* and *conclusions* [20, 34, 40, 52]. As for our choice of this classification scheme, the reason is the lack of domain-specific and annotated sentence classification datasets. For the sentence classification task, we found two datasets, *PubMed 200k RCT* [31], containing 200k of type and abstract sentence pairs and *CSAbstract* [6], containing 2k of pairs corresponding to the biomedicine and computer science domains, respectively. The goal

of the classification model is to provide accurate classification to identify sentence intent in the abstract, which can be used to assess the organization of the draft and thus induce self-reflection on how to improve the coverage and arrangement of the content. Following the *BERT-base-uncased*<sup>15</sup>, a model pretrained on *BookCorpus* and *English Wikipedia*<sup>16</sup> proposed in [33], we fine-tuned on these two datasets separately and trained the model with different hyperparameters. For the sentence classification task on *PubMed 200k RCT*, the F1-score of the model is 83.59%, while for the task on *CSAbstract*, the F1-score of the model is 86.37%. These results show that we can embed the BERT model into our system to provide users with organizational analysis. Please refer to Table 9 in Appendix A for the hyperparameters we set in this work for details.

**4.2.4 Evaluation Metrics.** As with abstract writing training, it is critical to provide individual and adaptive feedback during the learning process [16]. The summaries and abstracts synthesize the main ideas of the text and require the ability to understand, express, synthesize, and paraphrase [18, 19, 82]. Abstract writing training may benefit from the evaluation methods used in summary writing training. Computer-assisted summary writing training typically provides formative feedback in the form of assessment scores.

<sup>15</sup><https://github.com/google-research/bert><sup>16</sup>[https://en.wikipedia.org/wiki/English\\_Wikipedia](https://en.wikipedia.org/wiki/English_Wikipedia)

However, this single score is not an appropriate feedback [101], so we tend to score abstracts using different scoring criteria. As a writing task, the scoring rubric for abstract writing should first focus on the scoring dimensions of general writing, i.e., content, content organization, and expression [41]. In addition, from the literature [4, 90, 91] and the writing instructions on the website<sup>17</sup><sup>18</sup><sup>19</sup>, we concluded that a good abstract should be comprehensible, concise, and fluent. In addition, it should be consistent with the source text and use words and phrases that are different from the source text. Linguistic features are considered more appropriate because they capture different aspects of the abstract and thus provide more informative and instructive feedback [14]. Therefore, we use different linguistic features to compute the five aspects. First, we selected 21 linguistic features that were shown to be related to the quality of abstracts to calculate their quality based on [28]. Then, we clustered the linguistic features to measure the five aspects of abstracts as shown in Table 3. Weighted sums of linguistic features were used to measure these facets, where the coefficients were calculated in [28] for the corresponding correlations.

	Krippendorff's $\alpha$	Cohen's kappa	Spearman's $\rho$
Understandability	0.762	0.331	0.420
Consistency	0.434	0.315	0.357
Fluency	0.416	0.377	0.345
Diversity	0.681	0.458	0.563
Conciseness	0.641	0.482	0.555
Perceived quality	0.687	0.574	0.483

**TABLE 4: Correlation between evaluation metrics and human raters.**  
Krippendorff's  $\alpha$  and Cohen's kappa were used to measure the inter-rater reliability between two human raters on the formal quality and two other human raters on the perceived quality of 21 academic abstracts in computer science. Spearman's  $\rho$  was used to measure the correlation between human raters and evaluation metrics. The perceived quality given by the evaluation metrics is the mean of its calculated formal quality.

To verify the validity of the evaluation metrics in Table 3, we randomly selected 21 academic abstracts in the *CSAbstract* dataset [6] and retrieved their original articles. We defined the formal quality of the abstracts as five aspects measured by linguistic features, and the perceived quality was scored by experts directly after reading the abstracts. Following the annotation guidelines in Appendix B, we recruited four senior Ph.D. candidates in computer science to rate the abstracts, two of whom assessed formal quality and two assessed perceived quality. To assess the reliability of the ratings, we used Krippendorff's  $\alpha$  and Cohen's kappa. As shown in Table 4, we obtained inter-rater reliability (IRR) in the interval of (0.4, 0.8) and Cohen's kappa in the interval of (0.3, 0.6), which indicates a moderate agreement among human raters. In addition, we assessed the correlation between the average human raters and the evaluation

<sup>17</sup><https://classroom.synonym.com/list-abstract-qualities-8671549.html>

<sup>18</sup><https://www.brandeis.edu/writing-program/resources/students/handouts/features-of-a-good-abstract-handout.pdf>

<sup>19</sup>[https://www.abstractscorecard.com/uploads/cfp2/images/Abstract\\_Quality\\_Standards\\_Guidelines\\_13.pdf](https://www.abstractscorecard.com/uploads/cfp2/images/Abstract_Quality_Standards_Guidelines_13.pdf)

metrics using Spearman's  $\rho$ . The results showed moderate correlations between human and automatic metrics. Moderate correlations are acceptable because correlations between automatic evaluation metrics and human raters are usually not very high [78, 100].

In addition to experimentally verified correlations, correlations between linguistic features and corresponding quality are also meaningful in Table 3. Usually, words that appear less frequently in the *COCA spoken corpus*<sup>20</sup> or *SUBTLEXus corpus*<sup>21</sup> are uncommon, so they can make the text less easily understood. Consistency refers to the factual alignment between the abstract and the source and *ROUGE-3* is illustrated to be related to consistency [37]. By definition, *ROUGE-3* is used to measure the source similarity, which can be interpreted as high source similarity implies a high consistency. Furthermore, the linguistic features listed in Table 3 correspond to *Fluency* and are intuitively correlated. The implication of *Diversity* is twofold: lexical diversity and syntactic diversity. The first eight features in the row are used to measure lexical diversity [62], and the other features are used to measure syntactic diversity [60]. Finally, conciseness is also associated with the listed features, whose thresholds refer to the reference abstract.

### 4.3 Design of User Interface

Following the design principles mentioned in the formative study, we built *ALens* as a responsive web-based application to demonstrate the academic abstract writing training process. The front-end interface includes a *Rhetorical Structure View*, a *Flow Map*, a *Writing Area*, a *Reference Abstract*, and an *Evaluation Dashboard*. The *rhetorical structure view* displays the parsed rhetorical tree of the original article and is designed to facilitate the user to quickly grasp the main information. After catching the main ideas, users can write their abstracts in the writing area. When users finish their drafts, they can utilize the NLP models to analyze their abstracts. The results of the analysis will be displayed on the *evaluation dashboard*. Based on the evaluation results, users are expected to polish their abstracts. *Reference Abstract* displays the results of the organization of the reference, and the most relevant sentences in the abstract are displayed through the *flow map* and *rhetorical structure view*.

**4.3.1 Rhetorical Structure View.** The Rhetorical Structure View (Figure 3A) is designed to help the user grasp the hierarchy of ideas and thus quickly identify the main ideas (R2). The rectangle (Figure 3A-a1) at the top of the

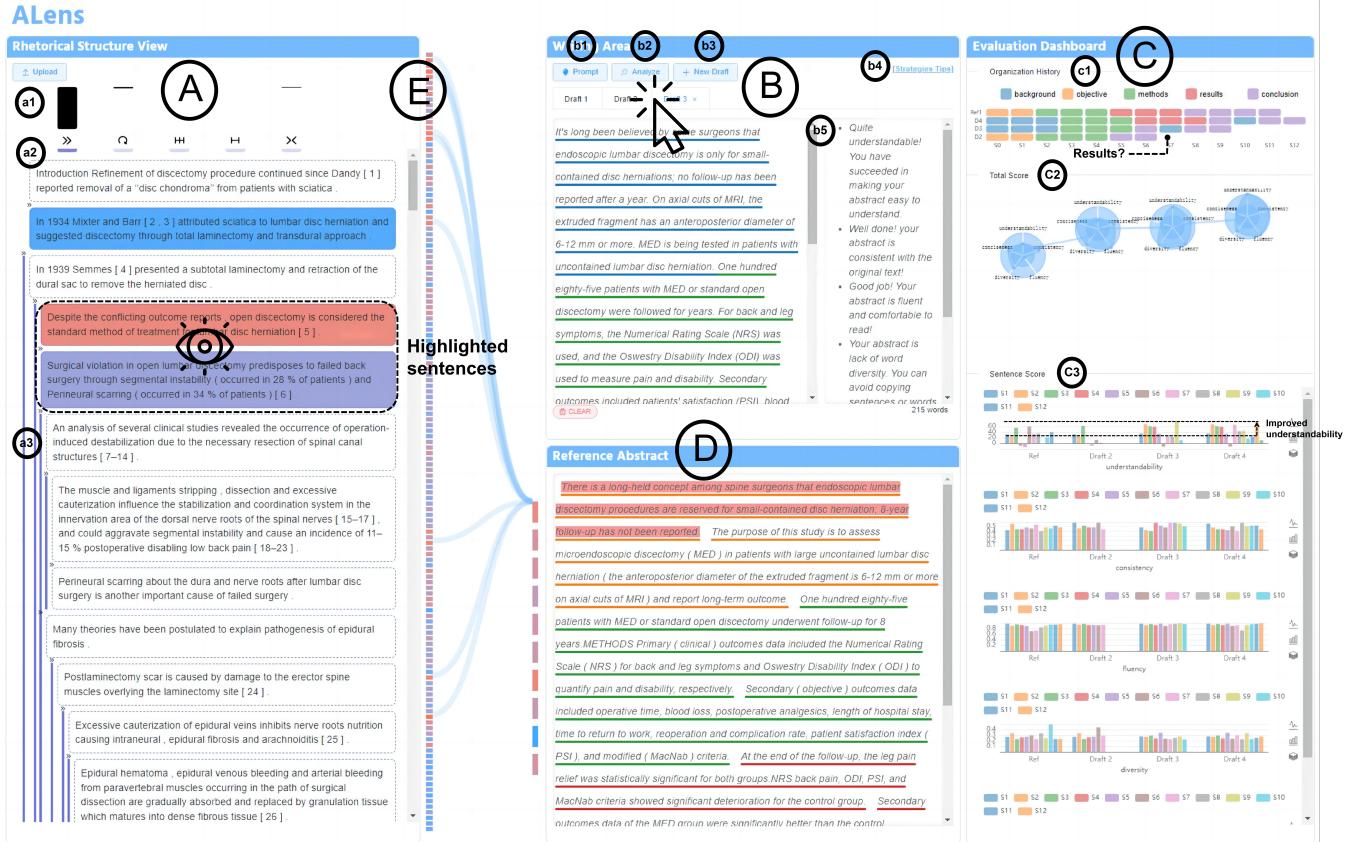
view indicates the number of respective relations by their length. To make the rhetorical relations between sentences more intuitive, these relations are visually encoded with glyphs, as shown in Table 5. Learners can click on these glyphs (Figure 3A-a2) to hide the secondary sentences in a pair of relations. For example, by clicking

Rhetorical Relation	Glyph
Background	Q
Contrast	X
Elaboration	»
Joint	H
Sequence	HH

**TABLE 5: Glyphs of rhetorical relations.**

<sup>20</sup><https://www.english-corpora.org/coca/>

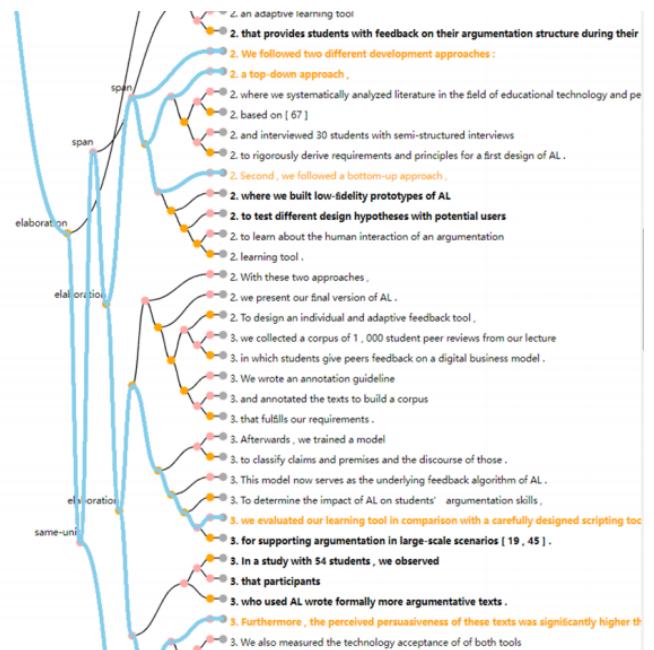
<sup>21</sup><https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus>

**ALens**

**Figure 3:** (A) The rhetorical structure view shows the RST tree of the article. (B) The writing area prepares the user to write an abstract. (C) The evaluation dashboard displays the results of semantic analysis results at different levels. (D)The reference abstract and (E) the flow map reveal the most relevant sentences from the reference abstract source.

on the elaboration glyphs, all elaborated sentences are preserved and the font color of the supporting sentences is lightened. In this way, learners can quickly capture key information, such as the core sentences in the elaboration relation and the secondary sentences in a contrast relation, or they can easily check the context by clicking on the glyphs again. In addition, we use a flattened tree structure (Figure 3A-a3) to display the hierarchy of ideas in the text, which is compact and makes good use of space. Sentences are wrapped as leaf nodes, logical glyphs are on the inner nodes, and the color depth of the rectangle (Figure 3A-a4) represents the number of corresponding relationships in the paragraph, so users can quickly identify the core sentences of each paragraph.

**Design Alternative.** To represent the hierarchy of ideas, we initially designed the rhetorical structure tree, as shown in Figure 4. The leaf nodes are connected to sentences, and pop-up tooltips contain the names of the relations. However, during the design iteration, users commented that there was a large amount of white space on the left side of the tree and it was too cumbersome to move the mouse over the internal nodes to see the relations. In addition, they criticized that sentences in the article were placed side-by-side and the paragraph structure was broken, resulting in low readability. Therefore, we chose the current design to display the relationships visually and minimize the differences from the original natural text and ensure readability.



**Figure 4:** Design alternative of the rhetorical structure tree.

**4.3.2 Writing Area.** The writing area (Figure 3B) supports basic editing functionality, allowing users to write drafts manually or by clicking on the “Prompt” button (Figure 3B-b1), which uses an existing summarization model to predict drafts as prompts (R5). Before writing, a suggested abstract writing strategy is provided by clicking on the “Strategies Tips” (Figure 3B-b4), enabling novices to quickly start abstract writing. After completing a draft, learners can click on the “Analyze” button (Figure 3B-b2) to analyze the abstract draft in six aspects (one for organizational structure and five for linguistic features). The sentences in the writing area are classified into five types and highlighted in different colors. Meanwhile, Figure 3B-b5 provides users with guided steps on how to improve the content and style of their abstracts based on the sentence classification results and the scores of linguistic features. Users can implement feedback by creating new drafts (Figure 3B-b3) to gradually improve the quality of their abstracts.

**4.3.3 Evaluation Dashboard.** The evaluation dashboard (Figure 3C) displays evaluation metrics at different granularities (R1, R3). We design the organization map (Figure 3C-c1) as a row of aligned rectangle tiles – the top row represents the organization scheme of the first draft, and the bottom row represents the most recent. Each tile in the row represents a sentence in the draft, and its color encodes the type of that sentence. After writing several drafts, users are expected to find the best organization scheme, and they are anticipated to identify the writing style of a group of papers by analyzing the best organization scheme for each paper in the group. In addition to the organization scheme, the line chart (Figure 3C-c2) records the overall score of the serialized drafts, with the five linguistic features encoded by the radar plot. However, the whole abstract and its scores for the five aspects may confuse learners [101], as they still need to recognize which parts need to be revised and which parts are already good. Therefore, (Figure 3C-c3) provides a more fine-grained analysis of the abstract. Each row represents a linguistic aspect, and a set of bars in the bar chart represents sentences in that draft. In this way, users can determine which sentences and aspects have not been considered and are poorly written. As a result, they can revise their drafts in a more precise and clear direction.

**4.3.4 Reference Abstract with a Flow Map.** The reference abstract with a flow map (Figure 3D&E) is designed to reveal the writing style of the reference abstracts (R4). Organizational detection is applied to the reference abstracts to explicitly reveal their organizational scheme. Also, we use a flow map to find the most relevant sentences in the source text for each sentence in the reference abstract. We use the sentence transformer [75] to calculate the semantic similarity score of each sentence in the source text with each sentence in the reference abstract. Each square tiles on one side represent a sentence from the source text or the reference. The color depth of each tile on the abstract side is calculated by averaging the first  $k$  similarity scores, where  $k$  can be specified by the user. And the tiles on the source text side represent the similarity score when the user’s mouse is placed over a sentence in the abstract. Meanwhile, the top  $k$  similar sentences in the source text are highlighted and linked to the sentences in the reference. In this way, users can explore the writing patterns of reference abstracts. For example, they may find that the sentence in the reference abstract describing the background may

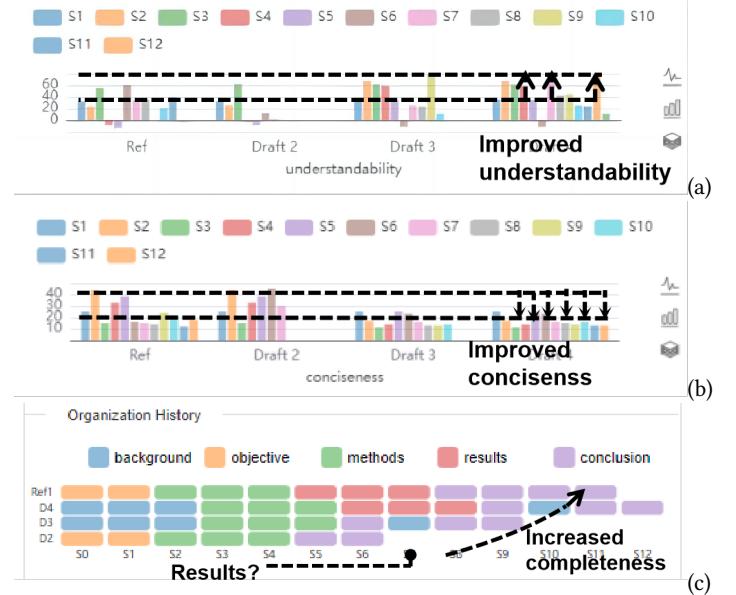
come from the end of the paragraph introducing the background in the source text. In this way, knowledge about the writing style of the abstract can be constructed.

## 5 EVALUATION

We evaluate the effectiveness of *ALens* in two ways. First, we describe two usage scenarios with two target users of *ALens*. Before that, we conducted a 10-minute tutorial with the involved participants to introduce *ALens*. We then asked them to explore with *ALens* for half an hour in a think-aloud manner. Second, we invited 21 participants who had no exposure to our system to conduct a user study to further assess the potency of *ALens*.

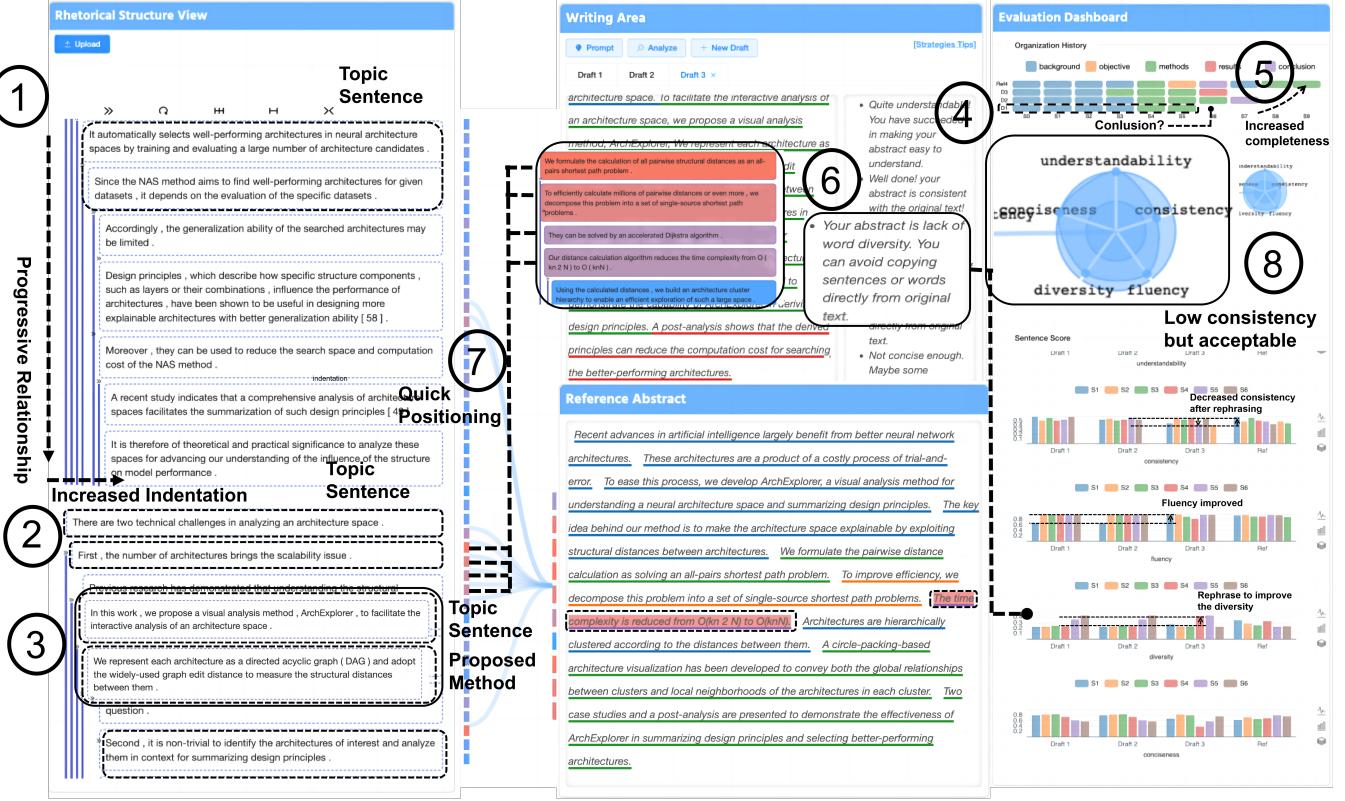
### 5.1 Usage Scenario I

In this subsection, we describe how Anker, a third-year undergraduate student, used *ALens* to train his academic abstract writing skills. Anker is from the Department of Biomedical Engineering and has been starting his research career for about four months. Prior to using *ALens*, he had no experience writing academic abstracts for journals and conferences, but he did have experience writing abstracts for course essays. We chose a paper from biomedical science and trained him in writing academic abstracts in a related field.



**Figure 5:** Anker’s training procedure: (a) The results of comprehensibility and references for each sentence in the three analysis drafts are on the far left. (b) The conciseness of each sentence in the three analyzed drafts and the results of the analysis of the reference are on the leftmost side. Relatively short bars imply relatively concise sentences. (c) History of the content organization of the three analyzed drafts and the organization of the reference.

First, he uploaded a prepared text file and then parsed the article into a rhetorical tree, as shown in Figure 3(A). We observed that he started to consciously select sentences while reading the article, explaining “as far as I know, a typical abstract should present the purpose of the work, what problems it tries to solve, the research methods used, and the conclusions”, and then he created a new tab



**Figure 6:** (1) The indentation before the sentence implies a progressive relationship in each paragraph. (2) Topic sentence and two juxtaposed challenges in the second paragraph. (3) The conclusion was omitted in the first draft. (4) The omission of the conclusion in the first draft was found, and content related to the assessment was added to improve the completeness of the abstract. (5) Based on the hint from *ALens*, Jimmy rewrote the abstract and diversity was improved. (6) Descriptions of efficiency were quickly highlighted and positioned. (7) Jimmy confirmed that the low consistency of the abstract could be caused by the rephrasing of the reference abstract, so he considered that low consistency was acceptable.

to parse the copied sentences and clicked on the “Analyze” button (Figure 3(B)(b2)). Looking at the conciseness (Figure 5(b)) and understandability (Figure 5(a)) indicators (Figure 5) in the evaluation dashboard (Figure 3(C)(c3)), he found that his second draft was not easy to understand and not concise enough, mainly because of the problems with the last three sentences. In addition, he found that the last three sentences were too long to read, so he broke them into shorter sentences, added some connectives, and then clicked the “Analyze” button. The result of the analysis showed that the sentences in his draft were more concise and easier to read. Then he decided to look at the reference abstract. To his surprise, he found that he had missed the results of the experiment in the *organization history view* (Figure 5(c)). He said that “*this is the first time I know that the experimental results needed to be independent of the conclusions, which I had previously thought already contained the results.*” Anker concluded with a quick review of the results section of the paper, grasping the comprehensive description of the results from the lengthy description of the results under the auxiliary rhetorical relations.

## 5.2 Usage Scenario II

In this subsection, we describe how Jimmy, a first-year graduate student, used *ALens* to train his academic abstract writing skills.

Jimmy comes from the Department of Computer Science and Engineering, and he has been starting his graduate career for about four months. Prior to using *ALens*, he had no experience writing academic abstracts for journals or conferences, but he did have experience writing abstracts for course papers. We selected a paper from IEEE TVCG and trained him in writing academic abstracts in a related field. First, as shown in Figure 6, he uploaded the introduction section of the prepared article, and the system parsed it into a rhetorical tree. He first looked at each paragraph of the rhetorical tree, and based on the visualized structural information, he found that overall, the indentation of sentences in each paragraph was increasing (Figure 6(1)), which meant that each paragraph was largely in a progressive relationship, so he inferred that the first sentence of each paragraph was the main idea sentence, and the sentences that were indented too much were most likely not the abstract alternatives that needed to be focused on (Figure 6(2)). Next, he looked at the first paragraph, which was mainly about the general background of the article, and he found that the paragraph mainly revolved around the first two sentences (**Topic Sentence** in Figure 6), so he copied the first two sentences directly into the abstract. He continued with the second paragraph, which focused on two “technical challenges” in the related area (**Topic Sentence** in Figure 6). From the sentences with the same level of indentation, he

intuitively found these two juxtaposed challenges, which he considered to be more important, and therefore copied them. In addition, he thought that transitions were also important logical relationships, but when he looked at the transitions and found the content after “however”, he thought that it did not need to be included in the abstract because the content of the transition had already been mentioned in the first challenge. Moreover, when he wrote the first draft, he thought that the transitions all fell under the above two technical challenges. He turned to move on to the third paragraph, which focuses on the method proposed by the authors (Figure 6(3)), which he thought was the focus of the article and needed more space to discuss. By looking at the rhetorical tree, he found that the content of the third paragraph mainly revolved around the first sentence, so he extracted the first sentence. He further looked at the rhetorical tree and found that part of the content is the author’s proposal to use the DAG method (Figure 6(3)) to characterize the problem, so he also extracted the DAG-related description words into the abstract. He was eager to use the above-extracted content as the first version of the abstract and clicked the “Analyze” button to see the results.

According to the analysis result, *ALens* thought that the abstract of this version lacked conclusive content (Figure 6(4)), and also the first sentence did not have enough fluency. In response to the first problem, Jimmy revisited the rhetorical tree and found that he had overlooked the assessment-related content, so he added the relevant content and revised the first sentence. After analyzing again, he found that the completeness of this version of the abstract was much improved (Figure 6(5)), but the fluency of the first sentence was not significantly improved. He checked the hints (Figure 6(6)) and learned that he might have retained too much content from the original text, so he made a proper rephrase of the abstract and then re-analyzed it. At this point, he found that the diversity of this version had improved, but the consistency with the article had decreased. He further guessed that he might have modified the original text, so the moderate decrease in consistency between sentences was acceptable. Overall, Jimmy was satisfied with this version of the abstract and did not intend to continue revising it.

He then clicked “Show Reference Abstract” to see the gap between his written abstract and the reference abstract for further learning, and the system displayed the relevant indicators. He found that the organization of the reference abstract was clearly different from the version he finally submitted. Specifically, he found that the reference abstract consisted mainly of one type of sentence, i.e., the objective, so he looked at the relevant sentences of the reference abstract based on the color indicators and found a description for efficiency (Figure 6(7)), i.e., “*Our distance calculation algorithm reduces the time complexity from O(kn<sup>2</sup>N) to O(knN).*”, “*which is indeed what I missed in my abstract,*” said Jimmy. He clicked on that sentence, and the rhetorical tree on the left side of the system showed the original content most relevant to that sentence by highlighting it. “*I can locate the relevant sentence in the original text very quickly.* (Figure 6(7))” He looked at the most relevant part of the original text and found that it was indented to the same degree as the DAG method proposed by the author, so he inferred that they were true of the same importance, and “*that was ignored by me.*”

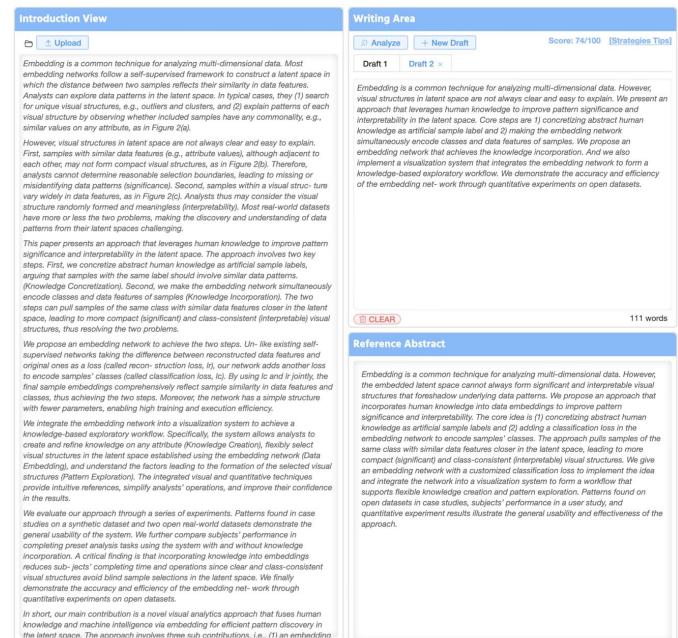
He finally observed the total score of the reference abstract and found that the reference abstract scored very high in all categories

except consistency . Using the “quick locate” function (Figure 6(7)), he compared each sentence of the reference abstract with the original text and found that the reference abstract had made a lot of rephrasing to the original text, so the consistency with the original text was not very high. At the same time, he noticed that the first sentence of the reference abstract only differed from the key content of the original text by one word, so this sentence had the highest consistency. “*This confirms my inference in my writing,*” that is, rephrasing causes a slight decrease in consistency (Figure 6(7)).

### 5.3 User Study

We quantitatively conducted a laboratory experiment to evaluate the performance of *ALens* and compared it to a baseline training system.

**Baseline Training System.** To evaluate *ALens*, we built a baseline system, as shown in Figure 7, which simulates the traditional way of writing an abstract. We controlled for similarities and differences between *ALens* and the baseline. Specifically, they both follow the same approach to abstract writing and share many features. First, both tools have a “Strategies Tips” button to learn general abstract writing strategies. In addition, a new draft button has been implemented to facilitate the iterative process of writing an abstract. The reference abstract appears when clicking on the “Show Reference” and disappears when clicking on the “New Draft” button. The difference between the baseline and *ALens* is that in the baseline system, users reflect on their abstracts based on the same overall score and reference abstract, but in *ALens*, they can obtain additional supporting information from the RST, classification, and metrics for revision.



**Figure 7: The baseline system supports users in reading articles (Introduction View), writing abstracts (Writing Area), and checking the reference abstract (Reference Abstract). The baseline can provide an overall score on the quality of the abstract.**

**Hypotheses.** To answer **RQ4**, we investigated perceived usefulness and usability between participants who used *ALens* and those who used the baseline system. In particular, we use a Wilcoxon rank sum tests to assess whether there are significant differences in the means of the constructs. Research on self-regulated learning theory suggests that personal feedback can help them learn better [11]. In the learning process, self-reflection is important for effective learning, which can trigger the creation of new knowledge through self-regulated learning [112]. Writing, as a creative process, is highly dependent on engagement [48]. It has been argued that when students are attentive and engaged in the writing process, they are able to write in a more cohesive manner [48]. Therefore, we propose the following hypothesis to answer **RQ5**. **H1:** *Individual feedback helps users to write abstracts in a more appropriate style than the baseline system.* Here, style is defined by content organization and language style. **H2:** *Compared to the baseline system, ALens helps users to construct knowledge for academic abstract writing.* **H3:** *Compared to the baseline system, ALens improves user satisfaction with the final draft of the abstract.* **H4:** *Compared to the baseline system, ALens enables users to be more involved in the writing process.*

**Experiment Setup.** To test our hypotheses, we designed a laboratory experiment in which participants were asked to read and comprehend the introduction of a given article in the field of computer science, write an abstract based on the introduction, learn the writing style of the given reference abstract, and revise the abstract they wrote at least once. Since academic abstract writing is highly relevant to the field, and students outside the field usually encounter obstacles in reading and understanding articles, we recruited 21 students from the Department of Computer Science of a local university via social media. Participants were randomly assigned to an experimental and control group. The experimental group used *ALens*, while participants in the control group used the baseline system. After random assignment, there are 12 students in the experimental group and 9 students in the control group. Participants in the experimental group (9 males and 3 females) had an average age of 21.17 ( $SD = 1.64$ ) and they had an average of 0.33 ( $SD = 0.65$ ) of academic abstract writing experience. In the control group, there were 7 males and 2 females. Their mean age was 20.89 ( $SD = 1.54$ ) and they had written an average of 0.22 ( $SD = 0.67$ ) academic abstracts. Upon completion, each participant received a \$20 stipend for their contribution. We designed an experiment with three phases, namely, a **pre-test phase**, a **short-term abstract writing training phase**, and a **post-test phase**. During the training phase, the experimental group used *ALens* and the control group used the alternative baseline system for reading, writing, and learning.

**Pre-test Phase.** To test whether our initial random grouping was indeed random, we first tested participants' acceptance of new information technology, feedback seeking, self-confidence, and academic abstract writing skills through a 16-question pretest. First, we asked participants four questions about the acceptance of new information technology for writing assistance, referring to the approach proposed by Agarwl et al. [2]. Second, we based on Ashford et al. [8] and asked them questions about the ability to actively seek feedback on academic paper writing. In addition, based on Ashford et al. [8], we tested their ability to control their mental ability states with the aim of knowing whether they were

overconfident in reporting their experimental results. Samples items for the constructs are “*I don't believe in myself*”; “*I feel that I am a valuable person on an equal footing with others*”; “*I seem to have a real inner strength when dealing with things. I have a very solid foundation, which makes me very confident in myself*.” Fourth, studies [44, 55, 87] have shown that juniors with task-specific writing practice (e.g., academic writing) statistically performed better in writing than students with only general writing training. Therefore, we indirectly understand their ability to write academic abstracts by asking them about their experience in writing academic abstracts such as past academic writing achievements and problems pointed out by reviewers or instructors during research submissions.

**Short-Term Abstract Writing Training Phase.** Before this phase began, we gave a brief introduction to our system and let them play with it for about 5 minutes. To test whether our system enables users to construct knowledge of academic abstract writing, we developed a short-term abstract writing training phase as follows. First, participants were asked to read the introduction of a computer science paper from the IEEE Transactions on Visualization and Computer Graphics (TVCG), since the introduction of the TVCG paper usually contains all the information of the article. Our co-author, an expert in the field of visualization, confirmed this fact. Participants were asked to spend at least 5 minutes reading the introduction of about 800 words to ensure that they had a basic understanding of the article. They were then asked to spend at least 10 minutes writing the abstract. Subsequently, users were allowed to check the reference abstract and learn the writing style of that abstract for a minimum of 5 minutes. The experimental group was then asked to use *ALens* to check the organization of the abstract and the placement of the core sentences, while students in the control group were allowed to analyze the abstract based on their knowledge. Then, if they were in the experimental group, they needed to revise their first drafts based on what they had learned about writing from the reference abstract and based on the evaluation metrics. Finally, they could progressively embellish the abstract until they were satisfied with the draft. All drafts from the training process were collected by both systems and sent for post-evaluation.

**Post-test Phase.** In this phase, we first measured users' intention to use our system, as well as usability and usefulness after technology acceptance testing [96]. In addition, we measured user satisfaction with their first and final drafts, and measured perceived engagement. The sample items for the five constructs are “*I will use the system for abstract writing training if it were released;*” “*I can write abstracts in the appropriate style;*” “*I feel I can learn to use the system quickly;*” “*I am satisfied with the first draft I wrote using the system;*” “*I focused on the writing itself and the time passes quickly for me.*” We used a 7-point Likert scale (7: very sure, 1: not very sure, 4 for neutral statements) for participants to assess.

**Measurement.** Technology acceptance, user satisfaction, and engagement were used to evaluate the system from the user's perspective and to test hypotheses **H3** and **H4**. In addition, we tested hypotheses **H1** and **H2** by measuring the quality of abstracts from the two groups. We measured the quality of the drafts in two ways: 1) perceived quality and 2) formal quality. In particular, for perceived quality, we invited two senior researchers in the field of visualization to help us evaluate the abstracts on a Likert scale of

	Content Integrity	Content Organization	Comprehensibility	Consistency	Fluency	Diversity	Conciseness	Perceived quality
Krippendorff's $\alpha$ between two raters	0.953	0.903	0.682	0.834	0.732	0.694	0.754	0.676
Cohen's kappa between two raters	0.807	0.851	0.637	0.702	0.643	0.578	0.602	0.631

TABLE 6: IRR between human raters. Two human raters for formal quality and two other human raters for perceived quality.

Pre-test Phase				
Group	New technology acceptance	Self-confidence	Feedback seeking	Times for writing abstract
Mean ALens	5.83	5.42	5.25	0.33
Mean Baseline	5.89	5.33	5.00	0.22
SD ALens	0.58	0.90	0.87	0.65
SD Baseline	0.60	0.87	1.00	0.67
Asymp.Sig. (2-sided)	0.831	0.874	0.433	0.500
Training Phase				
Group	First Draft Formal quality	First Draft Perceived quality	Second Draft Formal quality	Second Draft Perceived quality
Mean ALens	4.08	4.50	5.39	5.58
Mean Baseline	4.22	4.44	4.56	4.78
SD ALens	0.74	0.90	0.99	0.79
SD Baseline	1.15	0.73	0.74	0.67
Asymp.Sig. (2-sided)	0.776	0.874	0.041	0.025
Post-test Phase				
Group	First Draft Satisfaction	Second Draft Satisfaction	Engagement	Knowledge Construction
Mean ALens	3.50	5.33	4.33	5.33
Mean Baseline	3.67	4.67	4.44	3.78
SD ALens	1.09	0.49	0.49	0.98
SD Baseline	0.71	0.71	0.53	0.67
Asymp.Sig. (2-sided)	0.625	0.026	0.613	0.001

TABLE 7: Results of statistical analyses of the ALens and baseline systems on the Likert scale (1: low, 7: high).

1 – 7 (7: very good, 1: very poor). Both of them have 5 years of research experience. We used their average score as the final score of the draft. For another, we analyzed the formal quality of the first and final drafts. We defined the formal quality of the abstracts as the following seven aspects: *content integrity*, *content organization*, *comprehensibility*, *consistency*, *fluency*, *diversity*, and *conciseness*. The first two metrics are used to assess the classification results in terms of content integrity and content organization. The other five metrics are the same as the five metrics previously mentioned in Table 3. We used a Likert scale of 1 – 7 (7: very good, 1: very poor) to create criteria for these seven aspects. The rating guideline can be referred to Rating Guideline in Appendix B. We then annotated the 42 (21 \* 2) drafts ourselves and used our average score as the final score for the draft in that area. Krippendorff's  $\alpha$  in Table 6 shows the resulting inter-rater agreement reliability (IRR) scores. We obtained Krippendorff's  $\alpha$  scores between (0.67, 0.96) for the seven metrics, indicating considerable agreement between the two raters. In addition, Cohen's kappa was between (0.57, 0.71), showing the same result. Therefore, we conclude that the rated abstracts for the seven metrics should be reliable.

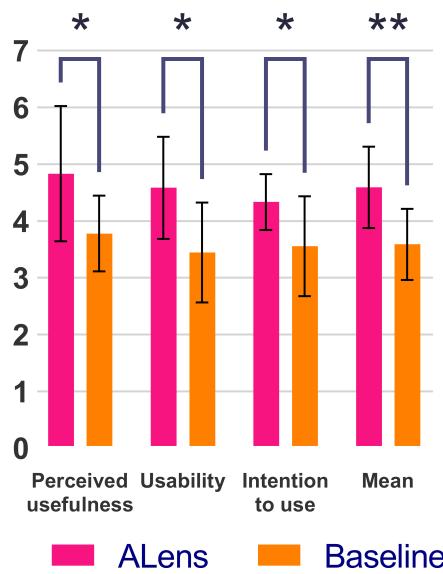
To answer research questions RQ4 – RQ5, we first ensured that our random assignment was successful and controlled for potential effects of small samples, and we compared the differences between the two groups on the four aspects in the pre-test phase as shown in Table 7. The two-sided asymptotic significance was greater than 0.05 for all four constructs, which ensured that the two groups did not differ significantly on these four constructs. Since the distribution of the two groups was not always normal, we performed post hoc Wilcoxon rank sum tests for pairwise group comparisons.

**5.3.1 RQ4: What is the technology acceptance level among junior researchers?** Figure 8 shows the average user ratings for technology-related questions. Wilcoxon rank sum tests show significant differences in perceived usefulness (*Asymp.Sig.* = 0.015), usability (*Asymp.Sig.* = 0.016), and intention to use (*Asymp.Sig.* = 0.039) when abstract writing skills were trained with different systems. We averaged the Likert scores for perceived usefulness, intention to use, and usability to compare technology acceptance. We find that ALens obtained a higher acceptance than the baseline system. Technology acceptance of a learning tool is an essential basis for further user learning. A positive technology acceptance provides a promising result for using this tool as an adaptive feedback application.

## 5.4 Results

### 5.4.1 RQ5: How effective is ALens in helping users write abstracts compared to the baseline system?

**Writing in a More Appropriate Style.** To test whether users write abstracts in a more appropriate style, we collected 42 draft abstracts written by participants and assessed their formal and perceived quality. From Table 7, we can see that there was no significance between the first drafts of the two groups on formal quality (*Asymp.Sig.* = 0.776) and perceived quality (*Asymp.Sig.* = 0.874). However, there is a statistically significant difference (formal quality: *Asymp.Sig.* = 0.041; perceived quality: *Asymp.Sig.* = 0.025) between the second draft of the two groups, which demonstrates the effectiveness of ALens in training users to write abstracts in a more appropriate style. Therefore, **H1 is accepted**.



**Figure 8: The technology acceptance of ALens and the baseline system on the Likert scale (1: low, 7: high). (\* : Asymp.Sig. < .05; \*\* : Asymp.Sig. < .01)**

**Abstract Writing Knowledge Construction.** To test whether the users gained knowledge about abstract writing, we asked them three quantitative problems and one qualitative problem. The three quantitative questions were: 1) “I have a general understanding of what an abstract should include when using the system”; 2) “I have a general understanding of how an abstract should be organized when using the system”; and 3) “I am now more familiar with the language style of the abstract.” The qualitative question was “do you have new insights into abstracts? Can you talk about them?” The results of the three quantitative questions ( $\text{Asymp.Sig.} = 0.001$ ) (Table 7) and the positive feedback from the qualitative question indicate a promising result that users can gain knowledge about abstract writing. Sample answers to the qualitative questions were “background is important and is usually described in one sentence”; “it is important to balance the proportion of the padding information”; “I first realize that there is a trade-off between diversity and consistency. While the phrases in the abstract should not be exactly the same as in the manuscript, preserving the original expressions is not a bad thing, given the consistency”. Therefore, **H2 is accepted**.

**Increased Satisfaction Level.** The results of the post-test phase showed that *ALens* and the alternative tool enabled users to improve their satisfaction with their drafts (Table 7 post-test phase). Although iteration usually makes things better, the difference in satisfaction ( $\text{Asymp.Sig.} = 0.026$ ) between the two groups became more significant in the second draft, implying that *ALens* could increase their satisfaction significantly. **H3 is accepted**.

**Higher Involvement in the Writing Process.** The statistical results showed that there was no significance ( $\text{Asymp.Sig.} = 0.613$ ) between the two groups in terms of engagement. Therefore, **H4 is rejected**. We believe that users feel distracted during the long reading and writing process, and the interaction time for learning writing style and analyzing their drafts is relatively shorter.

**Qualitative Feedback.** We also included some open-ended questions in the survey to get some suggestions for improvement. For example, we asked, *which part(s) of the system need(s) improvement and why?* In general, most participants were positive about *ALens*, especially the flow map, sentence-level evaluation, “Strategies Tips”, the sentence classification function, and the “Prompt” function. However, some participants also made constructive suggestions. In general, they complained about sometimes misleading rhetorical trees, incorrect sentence classification results, and confusing scores on the evaluation dashboard.

## 6 DISCUSSION AND LIMITATION

### 6.1 Validity and Technology Acceptance Evaluation

The outcomes of our user study demonstrated that the provision of adaptive formative feedback on students’ abstract drafts significantly contributes to their ability to compose abstracts in a more appropriate style, encompassing content organization and language usage. This improvement was verified through assessments of both formal and perceived quality, where the final drafts of both student groups surpassed the initial drafts in terms of overall quality. We posit that this effect can be explained by the principles of self-regulated learning theory. However, upon comparing the quality levels between drafts from the same batch, we observed a notable increase in the discrepancy between the two student groups. This underscores the significance of delivering feedback in the correct proportion and granularity, as learner uptake and self-regulation are heavily influenced by these factors, as discussed by Bandura [11]. In contrast to the alternative tool, *ALens* offers a range of personalized feedback, thereby motivating students to modify their writing behaviors. The short-term enhancements witnessed in the user study regarding academic abstract writing provide compelling evidence that self-regulation fosters participants’ motivation to acquire writing skills and construct pertinent knowledge. Furthermore, to effectively implement our study in a real-world scenario for abstract writing training, we conducted a validation of system technology acceptance, yielding promising results.

### 6.2 LLM Impact on L2 Students’ Academic Abstract Writing

The recent development of Large Language Models (LLMs) has had a positive impact on L2 students’ ability to learn academic abstract writing. For instance, these models can be used to automatically generate concise and accurate abstracts, helping students quickly grasp the core points and conclusions of their papers. They can provide targeted writing guidance, assisting students in organizing the structure and content framework of their papers. Additionally, these models can identify and correct language errors and grammar issues in students’ writing, providing real-time feedback and suggestions to improve the language quality of their papers. By analyzing the abstracts generated by these models, students can learn excellent writing styles and structures, enhancing their academic expression abilities.

However, it’s important to note that the improvement in these learning abilities is independent of the models themselves and

requires students to invest additional time in comparison and comprehension. Otherwise, if students overly rely on the models to complete their abstract writing tasks, they may lose their ability to think independently and solve problems [30, 105]. Academic writing requires deep thinking and independent research, and excessive dependence on large language models may result in students lacking a profound understanding of the issues and the ability to think critically. Furthermore, the content generated by large models may not always be accurate and reasonable. Students need to possess critical thinking and judgment skills when using these models in order to correctly evaluate and apply the generated content. If students lack these skills, they may blindly accept the suggestions and guidance from the models, which can affect the quality of their writing and their academic expression abilities.

### 6.3 Design Implications

We conducted a formative study aimed at gaining insights into the challenges faced by L2 junior students/researchers during the process of writing academic abstracts. Based on our findings, we derived design requirements that are relevant to this context. To the best of our knowledge, *ALens* represents one of the pioneering studies that have successfully established validated design requirements for an adaptive learning tool targeting academic abstract writing. Our research holds the potential to serve as a source of inspiration for individuals interested in the development of tools for training metacognitive skills. Instructors and developers of such tools can leverage our design requirements and discoveries to create their own training resources tailored specifically for enhancing academic abstract writing abilities.

The majority of existing computer-assisted writing tools primarily focus on assisting students in producing well-crafted writing pieces through iterative feedback and instructional support [84]. Similarly, in the present study, *ALens* and other similar tools explicitly incorporate knowledge construction as a key objective during their design process. While these tools anticipate users to enhance their writing skills through computer-assisted guidance, informed by the principles of self-regulated learning theory [112], the specific factors contributing to user progress have not been adequately measured or fully elucidated within the theoretical framework. However, as an adaptive training tool, *ALens* not only aids L2 junior researchers in achieving satisfactory writing outcomes but also endeavors to facilitate their exploration and acquisition of writing and stylistic knowledge. Notably, rather than providing a singular evaluation outcome such as scores or reviews, *ALens* offers a carefully designed pipeline that enables users to delve into the underlying reasons behind the evaluation results. In addition to the usage scenarios delineated in this study for L2 junior researchers, we envision *ALens* being of assistance to **experienced researchers** in analyzing abstracts outside their specialized domains. For instance, when sociologists or psychologists intend to contribute to the CHI community, *ALens* can be a potential option for comparing abstract writing methodologies in both engineering and humanities, subsequently facilitating the transfer of writing knowledge across fields. In addition to promoting interdisciplinary learning, *ALens* can be employed to acquire the abstract writing style employed by scientists or a specific research group of interest.

### 6.4 Limitation

Our study is subject to several limitations. First, there were concerns raised by four participants regarding the potentially misleading nature of the rhetorical tree employed in our research. Specifically, the performance of the tree structure in capturing relations such as “Joint” and “Sequence” was viewed as suboptimal. It is worth noting that our RST parser primarily operated at the sentence and paragraph levels. Due to the limited availability of sentence-level rhetorical structure datasets and the usage of the RST Discourse Treebank by Lynn and Marcu [22], which is annotated at the EDU level, we merged the EDUs to obtain corresponding sentences. During the user study, it became apparent that the reason behind the participants’ positive perception of the system’s performance was primarily rooted in their modest expectations regarding full automation. As one participant stated, “*the tip itself is just a reference, after all, I still need to read through INTRODUCTION.*” Second, five participants expressed dissatisfaction with the sentence classification model’s effectiveness in correctly identifying categories such as “objective” and “background”, as well as “background” and “conclusion”. This issue can be attributed to the lack of high consistency between the annotated criteria and the original abstract. Moreover, the limited dataset available for abstract sentence classification restricts the generalizability of the model. Third, participants exhibited confusion regarding the assigned scores. We believe that a disparity exists between the human understanding of words and the descriptive metrics employed for evaluation purposes. Last, although our user study indicated that *ALens* contributed to the improvement in abstract quality, it remains uncertain whether this enhancement can be solely attributed to increased editing. The study solely focused on written abstracts, with some observed editing behaviors, while the complete extent of editing actions was not documented or measured. In future investigations, we plan to record comprehensive edit logs throughout the entire process and analyze the effectiveness of editing behaviors.

Furthermore, during the user study, we successfully confirmed the immediate favorable impact of *ALens* on participants’ composition of academic abstracts. However, the long-term learning effects necessitate further validation. To address this, we plan to conduct a field experiment aimed at examining the efficacy and acceptance of *ALens* in a practical setting. This experiment will involve the formation of two distinct groups: a control group that will receive feedback exclusively from a tutor, and an experimental group that will receive feedback from both the tutor and *ALens*. By comparing the outcomes of these two groups, we aim to ascertain the long-term effectiveness of *ALens* in facilitating and enhancing the process of academic abstract writing.

## 7 CONCLUSION AND FUTURE WORK

This study introduces *ALens*, an innovative automated feedback learning tool designed to enhance academic abstract writing training by integrating visualization and interactive elements. A comparative analysis between *ALens* and a baseline system was conducted in a user study. The findings revealed that participants utilizing *ALens* exhibited improved abstracts in terms of content organization and language style. Notably, *ALens* demonstrated promising levels of technology acceptance and validity, thereby indicating

its potential for practical application in academic abstract training. Furthermore, the outcomes of both the formative study and the user study offer valuable insights for informing the development of abstract writing training tools.

For further research, we present two prospective scenarios aimed at enhancing the efficacy of our RST model and sentence classification. Regarding RST, our intention is to adopt the approach outlined in [58] to refine our RST parser. This involves initially training the parser on automatically annotated data and subsequently fine-tuning it using the RST-DT corpus introduced by Carlson et al. [23]. Additionally, we contemplate leveraging the capabilities of pre-trained language models, following the methodology proposed by Yu et al. [106] in their work on RST. To improve the precision of sentence classification, we propose the utilization of a similar data augmentation technique. Specifically, we will commence by training our Bert model on automatically annotated data, and subsequently fine-tune it using a dataset that is specific to the relevant domain. In terms of the evaluation dashboard’s confusing score, our objective is to enhance the comprehension of these scores by providing annotated exemplars and explanations. This approach aims to elucidate the underlying meaning of the scores and facilitate the development of actionable step-by-step guides for achieving higher scores. Furthermore, we are considering the exploration of a transformer-based model with the aim of predicting more accurate scores.

## ACKNOWLEDGMENTS

This work is partially supported by the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI) and Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Ministry of Education.

## REFERENCES

- [1] Asad Abdi, Norisma Idris, Rasim M Alguliyev, and Ramiz M Aliguliyev. 2016. An automated summarization assessment algorithm for identifying summarizing strategies. *PloS one* 11, 1 (2016), e0145809.
- [2] Ritu Agarwal and Elena Karahanna. 2000. Time flies when you’re having fun: Cognitive absorptions and beliefs about information technology usage. *MIS quarterly* (2000), 665–694.
- [3] JoAnn Grif Alspach. 2017. Writing for publication 101: Why the abstract is so important. , 12–15 pages.
- [4] Chittaranjan Andrade. 2011. How to write a good abstract for a scientific paper or conference presentation. *Indian journal of psychiatry* 53, 2 (2011), 172.
- [5] Linda Argote, Paul Ingram, John M Levine, and Richard L Moreland. 2000. Knowledge transfer in organizations: Learning from the experience of others. *Organizational behavior and human decision processes* 82, 1 (2000), 1–8.
- [6] Daniel King Bhavana Dalvi Dan Weld Arman Cohan, Iz Beltagy. 2019. Pretrained Language Models for Sequential Sentence Classification. In *EMNLP*.
- [7] Natasha Artemeva. 2008. Toward a unified social theory of genre learning. *Journal of business and technical communication* 22, 2 (2008), 160–185.
- [8] Susan J Ashford. 1986. Feedback-seeking in individual adaptation: A resource perspective. *Academy of Management journal* 29, 3 (1986), 465–487.
- [9] Anis Ashrafzadeh and Vahid Nimehchisalem. 2015. Vocabulary Knowledge: Malaysian Tertiary Level Learners’ Major Problem in Summary Writing. *Journal of Language Teaching & Research* 6, 2 (2015).
- [10] Ngo Xuan Bach, Minh Le Nguyen, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 160–168.
- [11] A Bandura. 1997. Social cognitive theory of self-regulation organizational behavior. (1997).
- [12] Jasmin Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607* (2020).
- [13] Diane Dewhurst Belcher and George Braine. 1995. *Academic writing in a second language: Essays on research and pedagogy*. Greenwood Publishing Group.
- [14] Elizabeth Bernhardt. 2010. *Understanding advanced second-language reading*. Routledge.
- [15] Robert A Bjork, John Dunlosky, and Nate Kornell. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology* 64 (2013), 417–444.
- [16] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)* 21, 1 (2009), 5–31.
- [17] Florian Bohm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214* (2019).
- [18] Ann L Brown, Joseph C Campione, and Jeanne D Day. 1981. Learning to learn: On training students to learn from texts. *Educational researcher* 10, 2 (1981), 14–21.
- [19] Ann L Brown and Jeanne D Day. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior* 22, 1 (1983), 1–14.
- [20] Jochen WL Cals and Daniel Kotz. 2013. Effective writing and publishing scientific papers, part II: title and abstract. *Journal of clinical epidemiology* 66, 6 (2013), 585.
- [21] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *thirty-second AAAI conference on artificial intelligence*.
- [22] Daniel Marcus Carlson, Lynn and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. <https://catalog.ldc.upenn.edu/LDC2002T07>.
- [23] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*. Springer, 85–112.
- [24] Chiou Sheng Chew, Wen-Chi Vivian Wu, Norisma Idris, Er Fu Loh, and Yan Piaw Chua. 2020. Enhancing summary writing of ESL learners via a theory-based online tool: system development and Evaluation. *Journal of Educational Computing Research* 58, 2 (2020), 398–432.
- [25] Shao Joyce Chin. 2016. Investigating the summary writing performance of university students in Taiwan. In *20th conference of English teaching and learning in the ROC*.
- [26] Colin B. Clement, Matthew Bierbaum, Kevin P. O’Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. *arXiv:1905.00075 [cs.IR]*
- [27] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685* (2018).
- [28] Scott A Crossley, Minkyung Kim, Laura Allen, and Danielle McNamara. 2019. Automated summarization evaluation (ASE) using natural language processing tools. In *International Conference on Artificial Intelligence in Education*. Springer, 84–95.
- [29] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48, 4 (2016), 1227–1237.
- [30] Li Deng and Yang Liu. 2018. *Deep learning in natural language processing*. Springer.
- [31] Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071* (2017).
- [32] Michael Derntl. 2014. Basics of research paper writing and publishing. *Int. J. Technology Enhanced Learning* 6, 2 (2014), 105.
- [33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [34] Mauro Bittencourt Dos Santos. 1996. The textual organization of research paper abstracts in applied linguistics. *Text & Talk* 16, 4 (1996), 481–500.
- [35] Dylan B Dryer. 2013. Scaling writing ability: A corpus-driven inquiry. *Written Communication* 30, 1 (2013), 3–35.
- [36] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications* 165 (2021), 113679.
- [37] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.
- [38] Muhammad Fareed, Almas Ashraf, and Muhammad Bilal. 2016. ESL Learners’ Writing Skills: Problems, Factors and Suggestions. *Journal of Education and Social Sciences* 4 (10 2016), 81–92. <https://doi.org/10.20547/jess0421604201>
- [39] Conditional Random Fields. 2001. Probabilistic models for segmenting and labeling sequence data. In *ICML* 2001.
- [40] Nancy Frey, Douglas Fisher, and Ted Hernandez. 2003. What’s the gist? Summary writing for struggling adolescent writers. *Voices from the Middle* 11, 2 (2003),

- 43–49.
- [41] Liang Guo, Scott A Crossley, and Danielle S McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18, 3 (2013), 218–238.
- [42] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916* (2021).
- [43] Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications* 121 (2019), 49–65.
- [44] Aceng Hasani, Aan Hendrayana, and Arip Senjaya. 2017. Using Project-Based Learning in Writing an Educational Article: An Experience Report. *Universal Journal of Educational Research* 5, 6 (2017), 960–964.
- [45] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.
- [46] Yulan He, Siu Cheung Hui, and Tho Thanh Quan. 2009. Automatic summary assessment for intelligent tutoring systems. *Computers & Education* 53, 3 (2009), 890–899.
- [47] Michael Heilman and Kenji Sagae. 2015. Fast rhetorical structure theory discourse parsing. *arXiv preprint arXiv:1505.02425* (2015).
- [48] Alan Hirvela and Qian Du. 2013. “Why am I paraphrasing?”: Undergraduate ESL writers’ engagement with source-based academic writing and reading. *Journal of English for Academic Purposes* 12, 2 (2013), 87–98.
- [49] Tasos Hovardas, Olia E Tsivitanidou, and Zacharias C Zacharia. 2014. Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education* 71 (2014), 133–152.
- [50] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.
- [51] Norisma Idris, Sapiyan Baba, and Rukaini Abdullah. 2011. Identifying students’ summary writing strategies using summary sentence decomposition algorithm. *Malaysian Journal of Computer Science* 24, 4 (2011), 180–194.
- [52] Mehrdad Jalalian. 2012. Writing an eye-catching and evocative abstract for a research article: A practical approach. *Electronic Physician* 4, 3 (2012), 520–524.
- [53] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 13–24.
- [54] Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3622–3631.
- [55] Karla M Johnstone, Hollis Ashbaugh, and Terry D Warfield. 2002. Effects of repeated practice and contextual-writing experiences on college students’ writing skills. *Journal of educational psychology* 94, 2 (2002), 305.
- [56] Suvarna Satish Khadilkar. 2018. The art and craft of making a draft: writing a good-quality scientific paper!, 151–154 pages.
- [57] Margaret R Kirkland and Mary Anne P Saunders. 1991. Maximizing student performance in summary writing: Managing cognitive load. *Tesol Quarterly* 25, 1 (1991), 105–121.
- [58] Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1600–1612.
- [59] Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*. Springer, 115–126.
- [60] Kristopher Kyle. 2016. Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. (2016).
- [61] Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods* 50, 3 (2018), 1030–1046.
- [62] Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly* 18, 2 (2021), 154–170.
- [63] Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch. 2013. *Handbook of latent semantic analysis*. Psychology Press.
- [64] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [65] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
- [66] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).
- [67] Jennifer A Livingston. 2003. Metacognition: An Overview. (2003).
- [68] Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics* 39, 2 (2013), 267–300.
- [69] William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.
- [70] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse* 8, 3 (1988), 243–281.
- [71] Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130* (2021).
- [72] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.
- [73] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-first AAAI conference on artificial intelligence*.
- [74] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
- [75] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://arxiv.org/abs/2004.09813>
- [76] Henry L Roediger III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.
- [77] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144.
- [78] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–39.
- [79] Françoise Salager-Meyer. 1994. Hedges and textual communicative function in medical English written discourse. *English for specific purposes* 13, 2 (1994), 149–170.
- [80] Abigail See. 2021. *Neural Generation of Open-Ended Text and Dialogue*. Stanford University.
- [81] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).
- [82] Arie S Spiegel and Peter F Delaney. 2016. Does writing summaries improve memory for text? *Educational Psychology Review* 28, 1 (2016), 171–196.
- [83] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 353–363.
- [84] Carola Strobl, Emilie Ailhaud, Kalliopi Benetos, Ann Devitt, Otto Kruse, Antje Proské, and Christian Rapp. 2019. Digital support for academic writing: A review of technologies and pedagogies. *Computers & education* 131 (2019), 33–48.
- [85] Yao-Ting Sung, Chia-Ning Liao, Tao-Hsing Chang, Chia-Lin Chen, and Kuo-En Chang. 2016. The effect of online summary assessment and feedback system on the summary writing on 6th graders: The LSA-based technique. *Computers & Education* 95 (2016), 1–18.
- [86] Thanatkun Tangpermpoon. 2008. Integrated approaches to improve students writing skills for English major students. *ABAC journal* 28, 2 (2008).
- [87] Luu Trong Tuan. 2010. Enhancing EFL Learners’ Writing Skill via Journal Writing. *English Language Teaching* 3, 3 (2010), 81–88.
- [88] MS Tullu and S Karande. 2017. Writing a model research paper: A roadmap. *Journal of postgraduate medicine* 63, 3 (2017), 143.
- [89] Milind S Tullu. 2019. Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key. *Saudi journal of anaesthesia* 13, Suppl 1 (2019), S12.
- [90] Oleg Vasilyev and John Bohannon. 2020. Is human scoring the best criteria for summary evaluation? *arXiv preprint arXiv:2012.14602* (2020).
- [91] Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836* (2020).
- [92] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.
- [93] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.
- [94] Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714* (2019).

- [95] Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284* (2019).
- [96] Jan Vom Brocke, Wolfgang Maas, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future work and enterprise systems. *Business & Information Systems Engineering* 60, 4 (2018), 357–366.
- [97] Jan Vom Brocke, Alexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the association for information systems* 37, 1 (2015), 9.
- [98] David Wade-Stein and Eileen Kintsch. 2004. Summary Street: Interactive computer support for writing. *Cognition and instruction* 22, 3 (2004), 333–362.
- [99] Adrian Wallwork. 2016. *English for writing research papers*. Springer.
- [100] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228* (2020).
- [101] Sara Cushing Weigle. 2002. *Assessing writing*. Cambridge University Press.
- [102] Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 10890–10905.
- [103] Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. Sumqe: a bert-based summary quality estimation model. *arXiv preprint arXiv:1909.00578* (2019).
- [104] Yu-Fen Yang. 2016. Transforming and constructing academic knowledge through online peer feedback in summary writing. *Computer Assisted Language Learning* 29, 4 (2016), 683–702.
- [105] Hyunsook Yoon and Alan Hirvela. 2004. ESL student attitudes toward corpus use in L2 writing. *Journal of second language writing* 13, 4 (2004), 257–283.
- [106] Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST Discourse Parsing with Second-Stage EDU-Level Pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4269–4280.
- [107] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.
- [108] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
- [109] Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational linguistics* 37, 1 (2011), 105–151.
- [110] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795* (2020).
- [111] Wei Zhu. 2004. Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of second language Writing* 13, 1 (2004), 29–48.
- [112] Barry J Zimmerman and Dale H Schunk. 2001. *Self-regulated learning and academic achievement: Theoretical perspectives*. Routledge.

## A HYPERPARAMETERS OF NLP MODELS

Hyperparams	PubMed	arXiv-cs
Size of the intermediate feed forward layer in each T5Block	2816	3072
Size of the encoder layers	1024	768
Maximum sequence length	16384	4096
Number of attention heads	16	12
Number of hidden layers	24	12
Vocabulary size	32100	32100
Size of the key, query, value	64	64
Number of decoder layers	24	12
Dropout rate	0.1	0.1
Activation function	relu	relu

TABLE 8: Hyper-parameters of two summarization models based on *LongT5* on the *PubMed* and *arXiv-cs* datasets.

Hyperparams	PubMed 200k RCT	CSAbstract
Number of attention heads	12	12
Number of hidden layers	6	6
Size of the encoder layers	3072	3072
Activation function	glue	glue
Maximum sequence length	512	512
Dropout rate	0.2	0.2
Vocabulary size	30522	30522

TABLE 9: Hyper-parameters of sentence classification model, *BERT*, on the *PubMed 200k RCT CSAbstract* dataset.

## B RATING GUIDELINE

In the realm of comprehensive writing tasks, the abstract writing scoring rubric should encompass the dimensions outlined in the rubric for independent writing, while also incorporating specific requisites for abstract composition. The five dimensions inherent to independent writing encompass *content integrity*, *content organization*, *language expression*, *communicative function*, and *writing conventions* [101]. These dimensions collectively address various aspects of writing, ranging from coherence to linguistic proficiency. For abstract writing, the evaluation framework extends beyond the five dimensions and accentuates the importance of *comprehensibility*, *fluency*, and *conciseness* as fundamental markers of communicative efficacy and adherence to writing conventions. In addition to these dimensions, the assessment also underscores the necessity for abstracts to exhibit alignment with the source text. The formal quality assessment of abstracts is thus underpinned by these seven key aspects [79]. For an elaborate exposition of the scoring criteria, kindly refer to the table (Figure 9) delineated in our user study.

Annotation Guideline for Academic Abstract Writing	Grade	6~7	3~5	1~2
	Content Integrity	Includes all content points in the source material	Contains most of the content points in the material	Obvious omission of most of the key points in the source material
	Content Organization	Accurate, appropriate use of rich articulation and compact structure of the written abstract	A simple articulation technique is used, and the abstracts written are generally compact	The abstracts written for the purpose of using the articulation technique are confusingly organized
	Comprehensibility	Smoothly written, with terminology based on general domain knowledge and explanations of emerging concepts	The text is well organized and based on general domain knowledge, but does not provide timely explanations of emerging concepts	Obscure wording throughout, no explanation of emerging concepts
	Consistency	All information is taken from the original text and is logically coherent	All the information is from the original text but some of the logic is broken	Contains some information that is not in the original text
	Fluency	Accurate and coherent presentation of the core points of the source text	A generally accurate and coherent expression of the core points of the original text	The wording is incomprehensible and the writing is incoherent
	Diversity	1. Use a wide range of vocabulary, sentence patterns and grammatical knowledge accurately and appropriately 2. No direct references to original text	1. Can use basic vocabulary, sentence patterns and grammar to express meaning, but there are obvious errors and inappropriate use 2. Generally uses own language to summarize, with a few phrases or sentences copied from source material	1. Simple use of vocabulary, sentence patterns and grammar knowledge, with many errors 2. Inability to summarize in their own language, with most content copied directly from the original text
	Conciseness	Long and short sentences are staggered, without overly complex and roundabout sentences and the word count is within a reasonable range	Long and short sentences are staggered, the number of words is within a reasonable range but some syntax is too complex and the sentences are long	Abstract exceeds word count while using a lot of redundant expressions