Journal of English for Academic Purposes 71 (2024) 101422

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine-tuning

Minjin Kim *, Xiaofei Lu

Department of Applied Linguistics, The Pennsylvania State University, University Park, PA, USA, 234 Sparks Building, PA, 16802, USA

ARTICLE INFO

Handling Editor: Dr Hilary Nesi

Keywords:
Genre
Move-step analysis
ChatGPT
Few-shot learning
Prompt refinement
Fine-tuning

ABSTRACT

Rhetorical move-step analysis has wielded considerable influence in the fields of English for Academic/Specific Purposes. To explore the potential of using ChatGPT for automated move-step analysis, this study examines the impact of few-shot learning, prompt refinement, and base model fine-tuning on its accuracy in move-step annotation. Our dataset consisted of the introduction sections of 100 research articles in the field of applied linguistics that have been manually annotated for move-steps based on a modified version of Swales' (1990) Create-a-Research-Space model, with 80 for training, 10 for validation, and 10 for testing. We formulated an initial prompt that instructed the base model to perform move-step annotation, evaluated it in a zero-shot setting on the validation set, and subsequently refined it with greater specificity. We also fine-tuned the base model on the training set. Evaluation results on the test set showed that few-shot learning and prompt refinement both led to significant albeit relatively small performance improvements, while fine-tuning the base model achieved substantially higher accuracies (92.3% for move and 80.2% for step annotation). Our results highlight the potential of using ChatGPT for discourse-level annotation tasks and have useful implications for EAP pedagogy. They also provide key recommendations for employing ChatGPT in research.

1. Introduction

Genre analysis has emerged as a crucial method for analyzing texts at the discourse level, pivotal in shaping pedagogical approaches aimed at providing novice writers with a foundational understanding of the rhetorical practices in new or unfamiliar genres (Moreno & Swales, 2018; Swales, 1990, 2004; Tardy, 2016). Particularly in the field of English for Academic Purposes (EAP), the art of successful academic writing transcends the mere presentation of content; it critically hinges on adept rhetorical choices, which represent a daunting challenge for novice scholars (Kessler & Polio, 2023; Tardy, 2016). In response, genre analysts and EAP researchers have delved deep into exploring the rhetorical intricacies of prominent academic genres such as published research articles, doctoral dissertations, and conference abstracts (e.g., Cotos et al., 2017; Lu et al., 2021a; Moreno & Swales, 2018; Yoon & Casal, 2020). One of the most common analytical methods adopted in this field is rhetorical move-step analysis, which involves dissecting and describing the rhetorical structure of a genre through manual annotation of 'moves', i.e., distinctive communicative units in a text, and

* Corresponding author. Department of Applied Linguistics, The Pennsylvania State University, University Park, PA, USA, 234 Sparks Building, University Park, PA, 16802, USA.

E-mail addresses: mmk6337@psu.edu (M. Kim), xxl13@psu.edu (X. Lu).

https://doi.org/10.1016/j.jeap.2024.101422
Received 26 February 2024; Received in revised form 11 June 2024; Accepted 16 July 2024
Available online 17 July 2024

'steps', i.e., smaller text segments that build up a move and help fulfill the move's purpose (Biber et al., 2007). Genre analysis research has offered valuable insights into the rhetorical structures of a range of academic (e.g., research articles, grant proposals, and lab reports) and non-academic genres (e.g., business emails) (e.g., Kessler, 2020; Park et al., 2021; Parkinson, 2017). These insights have been argued and shown to have potential for enabling learners to demystify and master the conventions of different types of academic discourse (e.g., Tardy, 2016).

Despite the widespread recognition of the usefulness of rhetorical move analysis, its scale is often limited by the time-consuming nature of manual coding (Casal & Kessler, 2023). Although a variety of Natural Language Processing (NLP) tools have been developed to assess syntactic and lexical features of texts, such as the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle, 2016), fewer automated tools for discourse-level analysis including genre analysis are available. This may not be surprising, as discourse-level annotation often requires a greater level of contextual understanding of the text than lexical or syntactic annotation, making it more labor-intensive to render large amounts of manually labelled training data and more challenging for NLP models to achieve high accuracies in such annotation. Two pedagogically oriented automated genre analysis tools, AntMover (Anthony & Lashkia, 2003) and Research Writing Tutor (RWT; Cotos & Pendar, 2016), utilized traditional supervised learning algorithms with lexical approaches such as bag of clusters and n-gram features, which may fall short in capturing the context-dependent complexity crucial for genre analysis. Models in the Generative Pre-trained Transformer (GPT) series (e.g., ChatGPT), which utilize an extensive transformer-based neural network pre-trained on a vast corpus of text data, can potentially perform better on automated genre analysis, as they can process language with a deeper and more sophisticated understanding of context (Ray, 2023; Wu et al., 2023).

In order to examine the possibility of using ChatGPT for rhetorical move-step analysis, which involves a discourse-level annotation task, this study examines the effect of few-shot learning, prompt refinement, and base model fine-tuning on its accuracy in annotating move-steps. To this end, we sourced the introduction sections of 100 research articles in the field of applied linguistics that have been manually annotated for move-steps based on a modified version of Swales' (1990) Create-a-Research-Space model from the Corpus of Social Science Research Article Introductions (Lu et al., 2021a). Our findings offer useful implications for employing ChatGPT and other large language models (LLMs) in discourse-level annotations and for EAP pedagogy. They also allow us to make important recommendations for using ChatGPT in research in general.

2. Literature review

2.1. Rhetorical move-step analysis and annotation tools

Over the past three decades, the concept of genre has been explored from diverse theoretical perspectives, including Systemic Functional Linguistics (Halliday, 1978), Rhetorical Genre Studies (Hyon, 1996; Paltridge, 1994), and English for Specific Purposes (ESP) (Hyland, 2007; Swales, 1990, 2004). Among these, the ESP approach stands out as particularly influential, which defines a genre as a class of communicative events with purposes recognized and shared by the members of a specific discourse community (Moreno & Swales, 2018; Swales, 1990, 2004). These events and their communicative practices are metaphorically framed as rhetorical moves—discoursal units serving coherent communicative functions—and are further dissected into steps, more granular elements that help achieve the move's objective (Biber et al., 2007). Thus, genre analysis, often referred to as rhetorical move (or move-step) analysis, seeks to unveil the recurring rhetorical structures and linguistic features pivotal in accomplishing rhetorical purposes within a specific discourse community (Casal & Kessler, 2023; Swales, 1990, 2004).

This move-step analysis approach has profoundly impacted the field of EAP by providing insights into the rhetorical and linguistic features of different types of academic writing, and such insights have formed a solid empirical foundation for developing EAP curricula, resources, and teaching methods (e.g., Cortes, 2013; Cotos et al., 2017; Kanoksilapatham, 2005; Lu et al., 2021a; Tessuto, 2015). One well-studied genre among this body of research is that of the research article (RA). For example, Kanoksilapatham (2005) analyzed 180 RAs from three engineering disciplines (civil, software, and biomedical) to identify the organizational structures of the texts with move-step analysis, revealing how sub-disciplinary influences shape the construction of individual RA sections and offering engineering students and practitioners a versatile move-step framework to organize their ideas in alignment with the norms of their disciplinary discourse. In addition to delineating the rhetorical structure of RAs and specific RA sections, some studies investigated the linguistic choices associated with different rhetorical move-steps, which yielded valuable insights for EAP curriculum and material development. Cortes (2013), for example, identified lexical bundles of various lengths in RA introductions and linked those bundles to different moves and steps, resulting in a useful list for academic writers. Lu et al. (2021a) also presented a list of three types of phrase-frames from a corpus of 600 RA introductions, namely, those unique to a specific move-step, those primarily associated with one move-step but appearing in others, and those common across multiple move-steps without a direct link to any single one. The practical applications of move-step analysis have not been limited to the RA genre, extending to various other genres such as conference abstracts, grant applications, business emails, and lab reports (e.g., Casal & Kessler, 2023; Park et al., 2021; Parkinson, 2017; Yoon & Casal, 2020).

The move-step analysis typically involves a three-stage methodology as outlined by Casal and Kessler (2023): 1) development of a rhetorical move-step model to capture patterns of rhetorical activity within the dataset, accommodating data variability, 2) application of the model to segment texts into rhetorical units, and 3) assessment and refinement of the move-step framework to enhance its reliability and validity through inter-coder agreement. Several recent studies have followed these steps in their move-step analysis, albeit with variations tailored to their particular research contexts (e.g., Cotos et al., 2017; Lu et al., 2021a; Yoon & Casal, 2020). In studying established genres, researchers often begin with existing rhetorical move frameworks, such as the Create-a-Research-Space

(CARS) model for RA introductions (Swales, 1990), adapting them to the specific rhetorical activities of the discourse community being studied. There is good consensus among researchers regarding the importance of coding procedure transparency and attention to inter-coder agreement and coding reliability (Casal & Kessler, 2023; Kim et al., 2024). Whereas the methodology is comprehensive and rigorous, the labor-intensive nature of the coding process can potentially limit the scope of move-step analysis studies. Specifically, this situation often compels researchers to work with smaller datasets, potentially compromising the generalizability of their findings across more diverse datasets. Furthermore, it also restricts our ability to conduct large-scale studies that explore the interaction among multiple text-related variables (e.g., discipline, part-genre), writer-related variables (e.g., L1 background, writing expertise), and linguistic features associated with move-steps.

Two tools designed to automate genre analysis for instructional purposes, i.e., AntMover (Anthony & Lashkia, 2003) and Research Writing Tutor (RWT; Cotos & Pendar, 2016), could be used to supplement manual analysis. AntMover was the first tool developed to automatically identify the structure of RA abstracts across various disciplines based on the CARS model. Trained using a Naïve Bayes classifier, a type of supervised learning, on a dataset comprising 100 published abstracts in the field of information technology, the tool achieved an average accuracy of 68% across its six-category classification system (Anthony & Lashkia, 2003), which could be considered relatively low for research purposes when compared to other NLP tools in our field. For instance, existing systems for word sense disambiguation (i.e., labeling instances of polysemous words with their specific meanings in context), a task that resembles the move-step annotation task to some extent, have achieved accuracies of around 90% (Lu & Hu, 2022). Although not extensively utilized for research, the tool has been adopted in a few genre-based intervention studies. For instance, Dong and Lu (2020) employed AntMover to facilitate guided genre analysis activities with 30 engineering master students, who were asked to use it to obtain a first-pass annotation of samples in a self-compiled specialized corpus of RA introductions in the students' fields of study. While the activities enhanced the students' genre knowledge and genre-based writing skills, the authors noted that AntMover's six categories did not cover all move-steps in the corpus and asked the students to manually check the output for missed or incorrectly tagged sentences. This study highlights the potential of the automated genre analysis tool for assisting novice academic writers in mastering research article writing while also indicating areas for further enhancement. A more recent development in this area is the RWT (Cotos & Pendar, 2016). Trained using a Support Vector Machine (SVM) algorithm on 650 RA introduction, this tool achieved average accuracies of 72.6% for move classification (3 classes) and 72.9% for step classification (17 classes). Cotos et al. (2020) explored 11 graduate-student writers' interactions with the tool, showing that the move-step tags it produced and its feedback and scaffolding features helped the students better understand the rhetorical structure of RA introductions, identify inconsistencies in their drafts, and implement effective revisions. However, this tool remains inaccessible to the public.

Although AntMover and RWT have exhibited substantial pedagogical value, there is much room for enhancement, particularly since they only rely on features based on bag of clusters and n-grams respectively in modeling. In genre analysis, understanding the context from surrounding sentences and the overall text flow is vital. In other words, accurately identifying specific structural steps often requires recognizing the context provided by adjacent steps (Biber et al., 2007). LLMs, known for their superior understanding of context (Ray, 2023), offer a complex architecture with a vastly larger number of parameters compared to simpler n-gram or bag of clusters approaches (Li et al., 2021; Ray, 2023), providing a promising avenue for significant advancements in move-step annotation accuracy.

2.2. ChatGPT (GPT-3.5) for classification tasks

Text classification, i.e., the procedure of assigning specific class labels to texts, plays an important role in such NLP applications as sentiment analysis, topic labeling, and dialog act classification (Li et al., 2021). Early research has tackled text classification with traditional models such as Naive Bayes, SVM, K-Nearest Neighbors (KNN), and Random Forest paired with different types of linguistic features (e.g., bag of words, n-grams) (Li et al., 2021). While these models have advantages in stability, they require extensive feature engineering and often face performance limitations (Balkus & Yan, 2023; Li et al., 2021), partly due to their neglect of the inherent sequential structure and contextual details in text, which hinders their ability to grasp the meanings of words and other linguistic expressions in context (Brown et al., 2020; Li et al., 2021). In recognition of these limitations, recent NLP research has pivoted towards deep learning models, which enable classifiers to capture complex word characteristics and contextual variation (Liu et al., 2023). OpenAI's GPT models and its consumer-facing service ChatGPT exemplify this shift. With a high level of contextual understanding, these models can generate precise, relevant responses to user prompts, and their performance on specific tasks and/or in specific domains can be further improved through few-shot learning, prompt engineering, or fine-tuning (Brown et al., 2020; Kocoń et al., 2023; Ray, 2023).

Few-shot learning operates by providing the model with K examples of paired contexts and completions, followed by a single context example, from which the model is then tasked to predict the completion, while zero-shot learning relies on a task's natural language description (i.e., prompt) only without any examples (Brown et al., 2020). Fine-tuning the base model with a training dataset, essentially an extended form of few-shot learning, involves increasing the K value to adjust the model's pre-trained parameters on a domain-specific dataset. Few-shot learning and fine-tuning are now both frequently employed to enhance the performance of the base model on targeted classification tasks (Brown et al., 2020; Wei et al., 2022). For example, Loukas et al. (2023) reported that in classifying customer service queries into 77 distinct intent categories, GPT-3.5 in a one-shot setting and GPT-4 in a three-shot setting achieved F1 scores of 74.3% and 82.7%, respectively, outperforming several fine-tuned masked language models (e.g., P-MPNet-v2) in the same settings. Wachowiak and Gromann (2023) reported that in a 12-shot setting, GPT-3 achieved an accuracy rate of 65.15% in detecting the source domains of conceptual metaphors.

Prompt engineering refers to the process of carefully designing and refining the input prompts provided to language models to elicit specific or more accurate responses (Brown et al., 2020). This method could prove especially advantageous for genre analysis, because it could address the difficulty in automatically identifying certain steps whose rhetorical meanings are not explicitly conveyed through functional language, a difficulty discussed in Cotos and Pendar (2016). While the difficulty makes the steps challenging to detect with traditional supervised learning models, we have the potential, by utilizing ChatGPT, refined via prompt engineering, to operationalize some of these challenging steps. Although research into prompt engineering is still in its early stages, one notable study by Fatouros et al. (2023) demonstrated its efficacy. They used ChatGPT to perform sentiment analysis on financial news headlines in a zero-shot setting. By providing questions or statements crafted in a way that guides the model to increasingly better understand the task, they achieved a 35% performance improvement over traditional financial sentiment analysis models, such as FinBERT, showing the promising capabilities of ChatGPT when combined with carefully engineered prompts. Huang et al. (2023) also examined ChatGPT's capability to detect and explain implicit hate speech in hateful tweets through careful prompt designs in a zero-shot setting. They reported that the model correctly identified 80% of implicit hateful tweets.

Within the fields of applied linguistics and language education, a few efforts have been made to explore the potential of ChatGPT in such tasks as automated essay scoring and question generation. For example, Mizumoto and Eguchi (2023) employed the GPT-3 model to score 12,100 essays sourced from the ETS Corpus of Non-Native Written English (TOEFL11) and reported that the model alone could not predict the gold standard levels of the essays with an adequate level of accuracy and that a model combining the GPT-predicted scores with a set of lexical, syntactic, and cohesion features achieved better performance. Lee et al. (2023) used ChatGPT to create an automatic question generation system for English reading comprehension and developed a step-by-step protocol for generating high-quality questions with ChatGPT through multiple validation rounds by experts and teachers. The protocol included the need to limit the passage length to 250 words and to clearly specify question types and formats in the prompt. They also noted some limitations of ChatGPT such as its dependence on the lexicon of the original texts and the restriction of question types to mainly WH-questions. The two studies above demonstrated the potential of using ChatGPT in language assessment and material development, while also pointing to areas for further exploration and improvement.

While previous studies have exhibited the potential of ChatGPT for classification tasks, this potential and the challenges that may arise along with it have not yet been systematically exploited for discourse-level corpus annotation within the fields of applied linguistics and language education, and no study has explored the possibility of using ChatGPT for rhetorical move-step analysis. This analysis also differs from other classification tasks examined in existing studies in terms of the number and functional nature of the classes involved. To address this gap, this study explores the potential of using ChatGPT for rhetorical move-step annotation and examines the impact of prompt refinement, three-shot learning, and fine-tuning on the model's annotation accuracy. The findings are anticipated to shed light on the feasibility of employing ChatGPT for other types of corpus annotation and analysis that can be framed as classification tasks as well as the potential applications of ChatGPT in genre-based EAP pedagogy.

2.3. *Research questions*

This study aims to address the following two research questions:

Table 1
Move-step framework for research article introductions (Lu et al., 2021a).

Move/Step	Description	Tag
Move 1	**Establishing a research territory**	
Step 1	Claiming centrality or value of research area	[M1_S1a]
Step 1	Real-world contextualization	[M1_S1b]
Step 2	Making generalizations about research area	[M1_S2]
Step 3	Reviewing items of previous research	[M1_S3]
Move 2	**Establishing a niche**	
Step 1	Counter-claiming	[M2_S1a]
Step 1	Indicating a gap	[M2_S1b]
Step 1	Question raising	[M2_S1c]
Step 1	Continuing a tradition	[M2_S1d]
Step 1	Pointing out limitations of previous research	[M2_S1e]
Step 2	Providing justification	[M2_S2]
Move 3	**Presenting the present work via**	
Step 1	Announcing present research	[M3_S1]
Step 2	Presenting research questions or hypotheses	[M3_S2a]
Step 2	Advancing new theoretical claims	[M3_S2b]
Step 3	Definitional clarification	[M3_S3]
Step 4	Summarizing methods	[M3_S4a]
Step 4	Explaining a mathematical model	[M3_S4b]
Step 4	Describing analyzed scenario	[M3_S4c]
Step 5	Announcing and discussing results	[M3_S5]
Step 6	Stating the value of present research	[M3_S6]
Step 7	Outlining the structure of the paper	[M3_S7]
Step 8	Rationalizing research focus and design	[M3_S8]
Step 9	Presenting limitations of current study	[M3_S9]

1. To what extent can few-shot learning and prompt refinement improve the performance of the base model of ChatGPT (GPT-3.5) for annotating applied linguistics research article introductions with rhetorical move-steps?
2. To what extent can fine-tuning improve the performance of the base model of ChatGPT (GPT-3.5) for annotating applied linguistics research article introductions with rhetorical move-steps?

3. Methodology

3.1. Data source

To address the two research questions, the present study employed the data from the Corpus of Social Science Research Article Introductions (COSSRAI) (Lu et al., 2021a), which consists of the introduction sections of 600 published research articles from six social science disciplines: Anthropology, Applied Linguistics, Economics, Political Science, Psychology, and Sociology. For each discipline, the corpus compilers sampled the introduction sections of 100 RAs published in five high-impact journals endorsed by two disciplinary experts in 2012–2016, with one sample per issue and four samples per year. Each RA introduction was converted into a plain text file and manually cleaned to remove formulas, figures, tables, parenthetical citation elements, and any textual oddities arising from the text conversion process. The resulting corpus consisted of 513,688 words across 600 introductions (mean = 856, SD = 476).

The corpus was manually annotated for rhetorical move-steps by a team of seven researchers based on a collaboratively refined version of Swale's (1990) CARS model (see Table 1). The annotation proceeded in three stages: framework refinement, inter-annotator agreement assessment, and independent coding. In the first stage, the seven researchers independently coded the same 30 introductions, reconciled discrepancies through extensive group discussion, and modified some steps in the CARS model based on the data. In the second stage, the team was divided into three pairs, with the seventh researcher working as a coordinator. Each pair was assigned a distinct set of 10 texts. The two researchers in each pair first independently annotated the 10 texts, and these annotations were used to assess inter-annotator agreement. The average percent agreement was 91.7% for moves and 72.1% for steps. Each pair attempted to resolve discrepancies through discussion first, and all seven researchers then convened to resolve all remaining discrepancies from the three pairs and made additional minor adjustments to the framework. As a result of the two stages, multiple substeps were added to the CARS model to account for the rhetorical functions that emerged from the data. In the final phase, all remaining texts were split into seven batches, each annotated by one researcher and reviewed by another. Discrepancies were resolved in a series of team meetings.

Throughout the annotation process, the team used the rhetorical chunk as the unit of analysis, conducting a comprehensive examination of linguistic, structural, and content-based cues indicating shifts in the authors' communicative goals and functions. Moves and steps were not treated as formal units. However, following Cotos and Pendar (2016) and Cotos et al. (2017), they employed the sentence as the unit of annotation, assigning rhetorical move-step tags (see the Tag column in Table 1) to the end of each sentence. While the unit of annotation (i.e., sentence) often does not align with the unit of analysis (i.e., rhetorical chunks), sentence-level annotation offers practical advantages in maintaining consistency and improving accuracy in rhetorical chunk identification. In other words, because sentences are universally recognized and easily identifiable units in writing, they offer a clear and consistent basis for marking textual features and communicative functions, while chunk-level annotation may add an additional layer of subjectivity (i.e., potential inconsistencies in identifying rhetorical chunk boundaries) to the coding process. Moreover, detailed sentence-level annotations can highlight specific linguistic and rhetorical cues that signal the start and end of rhetorical chunks. For example, transitions, conjunctions, and topic sentences that may denote boundaries within the text can be more accurately identified when each sentence is meticulously analyzed. This precision at the sentence level makes it easier to aggregate sentences into meaningful chunks based on shared functions. This coding approach resulted in some sentences receiving multiple tags, with the first tag indicating the function realized by the main clause and other tags indicating functions realized by other parts of the sentence.

As the first effort to test the possibility of using ChatGPT for rhetorical move-step annotation, we chose to focus on the field of Applied Linguistics. This decision was informed by the recognition that RA introductions have exhibited not only common characteristics across disciplines but also unique rhetorical and linguistic features specific to each discipline (Hyland, 2015; Lu et al., 2021b). The Applied Linguistics subcorpus contained 100 RA introductions with 63,333 words (mean = 633, SD = 413). We employed stratified random sampling to create two distinct subsets of 10 samples first, each with two introductions from five different journals. One was designated as the validation set (257 sentences) and the other as the testing set (182 sentences). The remaining 80 samples served as the training set (1556 sentences), with a subset of 40 samples (795 sentences) from this set employed to assess the impact of training sample size on the performance. The use of relatively small training sample sizes (40 and 80 RA introductions) in our experiments aligns with the capabilities of ChatGPT (or LLMs in general) to learn from small training samples (Brown et al., 2020) and allows us to compare these capabilities against those of traditional machine learning models. For comparison, AntMover was trained on 80 abstracts and achieved an average accuracy of 68% on a test set of 20 abstracts. Contrastively, RWT was trained on a large training set of 650 introductions and achieved higher accuracies of 72.6% for moves and 72.9% for steps on a test set of 37 introductions. If ChatGPT achieves higher accuracies with fewer training samples, it would confirm a main benefit of LLMs, namely, their efficiency in handling classification tasks such as move-step annotation with smaller training sets compared to traditional machine learning algorithms (Brown et al., 2020). We adopted the standard practice of dividing data into training, validation, and test sets, typically at an 80/10/10 ratio because this approach has been both theoretically and empirically found to ensure sufficient training data while preventing overfitting, contributing to models that are both less error-prone and more generalizable (Gholamy et al., 2018; Pandey et al., 2022).

To further validate our results, we added two more test sets, one with 10 introductions (103 sentences) from Psychology and the other with 10 introductions (151 sentences) from Anthropology. For each set, we randomly sampled two introductions from each of the five journals in the respective field. These fields frequently intersect with applied linguistics and the additional testing can help evaluate the extent to which the model's ability could generalize across closely related fields.

3.2. *Procedure*

To examine the impact of few-shot learning, prompt refinement, and base model fine-tuning on the accuracy of ChatGPT for annotating rhetorical move-steps in RA introductions, we followed the procedure summarized in Fig. 1. We first designed an initial prompt to guide the GPT-3.5 base model to annotate unlabeled RA introductions with rhetorical move-steps. This initial prompt included the definitions of moves and steps and detailed explanations of the goals of rhetorical move-step analysis, the unit of annotation (i.e., sentences), and the move-step framework and tags to be used in the annotation as outlined in Table 1. It further provided explanations for genre (i.e., RAs), part-genre (i.e., introductions), and discipline (i.e., applied linguistics) of the texts to be annotated. The prompt then guided the model to segment paragraphs into sentences and annotate each sentence according to the specified framework. As some sentences in the corpus were annotated with multiple tags to account for their potential multi-functionality (Cotos et al., 2017; Lu et al., 2021a), the prompt also included the possibility of assigning multiple tags to a sentence with multiple rhetorical functions. We evaluated the performance of the base model with the initial prompt on the validation set, scrutinized the confusion matrix to identify potential areas of ambiguity for the model, and refined the initial prompt in an effort to help the model address those areas. The full initial and refined prompts can be found in the Appendix.

The base model was subsequently evaluated on the test set in four settings: 1) with the initial prompt in a zero-shot setting, 2) with the initial prompt in a three-shot setting, 3) with the refined prompt in a zero-shot setting, and 4) with the refined prompt in a three-shot setting. In the three-shot setting, we provided the model with three annotated samples (totaling 38 sentences) randomly selected from the training set. Finally, we proceeded to fine-tune the base model using the training sets (40 samples and 80 samples) with the refined prompt, aiming to enhance its annotation accuracy and adaptability to our specific domain. To this end, we utilized the gpt-3.5-turbo-1106 model recommended by OpenAI. We fine-tuned the model using the OpenAI API in Python and conducted evaluations within the Playground system.[1] The OpenAI Playground is a web-based interface that simplifies the process of constructing and testing predictive language models (OpenAI, 2024). This user-friendly platform allows users to select and load the desired model and enter system and user messages through a chat-like interface. It offers a more interactive experience than coding in Python similar to chatting with ChatGPT, but with additional features such as the ability to adjust parameters, load specific models, maintain consistent prompts, and save results. Given an input text, the system outputs the text along with the assigned tags. We opted to use this platform for testing both the base and fine-tuned models, rather than batch-processing texts via the API in Python, because this approach may be more accessible to researchers and teachers who are not familiar with coding, making it easier for future researchers and teachers to replicate or further test our method.

Additionally, we conducted two further stages of validation. First, we tested the base model in a three-shot setting with the refined prompt and also assessed the 80-introduction fine-tuned model (using the same refined prompt) with the same test set on three separate occasions to evaluate result consistency. For each test, we initiated a new session to ensure that the model did not retain any previous information. Subsequently, we assessed the performance of the 80-introduction fine-tuned model using additional test sets from Psychology and Anthropology, to further validate our results. OpenAI confirms that their models are not trained with data input through the API and Playground services (OpenAI, 2024).

3.3. *Evaluation*

To evaluate the performance of the base model in different settings and of the fine-tuned models, we computed the precision, recall, and F1 score for each move and step and the overall accuracy and weighted average precision, recall, and F1 score for all moves and steps. For each move or step, precision was the ratio of true positive (TP) predictions to the total number of true and false positive (FP) predictions, recall was the ratio of TP predictions to the total number of TP and false negative (FN) predictions, and the F1 score was the harmonic mean of precision and recall, i.e., (2 * Precision * Recall)/(Precision + Recall). The harmonic mean is a type of average of rates or ratios that balances them in a way that does not disproportionally bias toward any of them. In the case of F1 score, it combines precision and recall into a single measure that balances both, ensuring that neither is disproportionately emphasized. For all moves and steps, the overall accuracy was the ratio of correct predictions to the total number of predictions, and the weighted average precision, recall, and F1 score were computed by first weighting the value for each move or step by the number of actual instances of that move or step, then summing those weighted values, and finally normalizing the summed weighted values by the total number of instances. For all metrics, the values range from 0 to 1, with a higher value indicating better performance. These metrics, along with the confusion matrix, were computed using the *sklearn.metrics* package in Python. In the small proportion (10 out of 182) of cases where multiple labels were assigned to a single sentence, the sentence was considered to have been correctly annotated when at least one predicted tag matched the primary tag (i.e., the first tag) of the sentence to facilitate comparison with previous results. The significance of performance differences between different experiment settings was tested with McNemar's test, a well-established statistical test for

[1] https://platform.openai.com/playground (The parameters were set as follows in our experiment: temperature = 1, maximum token length = 2048, Top P = 1, frequency_penalty = 0, presence_penalty = 0, stop sequences = none, epochs = 4, batch size = 1, learning rate multiplier = 2.).

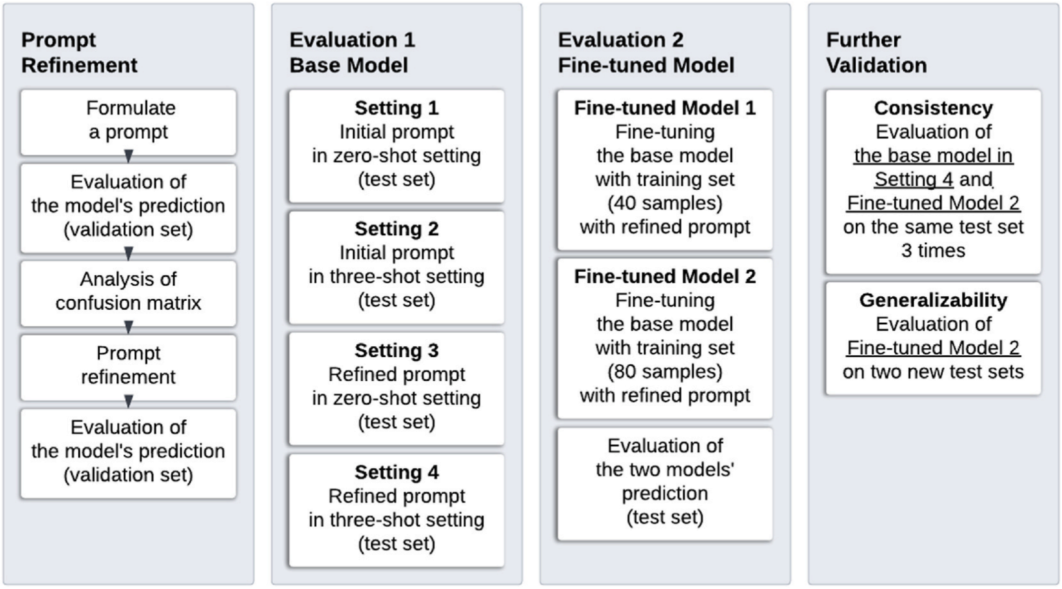
Prompt Refinement
Formulate a prompt
Evaluation of the model's prediction (validation set)
Analysis of confusion matrix
Prompt refinement
Evaluation of the model's prediction (validation set)
Evaluation 1 Base Model
Setting 1 Initial prompt in zero-shot setting (test set)
Setting 2 Initial prompt in three-shot setting (test set)
Setting 3 Refined prompt in zero-shot setting (test set)
Setting 4 Refined prompt in three-shot setting (test set)
Evaluation 2 Fine-tuned Model
Fine-tuned Model 1 Fine-tuning the base model with training set (40 samples) with refined prompt
Fine-tuned Model 2 Fine-tuning the base model with training set (80 samples) with refined prompt
Evaluation of the two models' prediction (test set)
Further Validation
Consistency Evaluation of the base model in Setting 4 and Fine-tuned Model 2 on the same test set 3 times
Generalizability Evaluation of Fine-tuned Model 2 on two new test sets

Fig. 1. Evaluation procedure of ChatGPT's accuracies in move-step annotation.

scenarios like ours based on the Chi-squared (χ^2) distribution (e.g., Dietterich, 1998; Kavzoglu, 2017). This test was performed using the *statesmodels* package in Python.

4. Results

4.1. Impact of prompt refinement and three-shot learning

In order to answer our first research question, we first evaluated the performance of the base model with the initial prompt in a zero-shot setting on the validation set. For move classification, the model achieved an accuracy of 28.5% and weighted average precision, recall, and F1 scores of 62.0%, 28.5%, and 32.9%, respectively. For step classification, the model achieved an accuracy of 11.3% and weighted average precision, recall, and F1 scores of 39.2%, 11.3%, and 13.4%, respectively.

Following this evaluation, we examined the confusion matrices for both move and step classifications to refine the initial prompt. The recall rates for Moves 1 (20.5%) and 2 (28.5%) were both rather low, indicating that these moves were frequently misclassified as other moves. The analysis of the move-classification matrix and the misclassified instances helped us identify potential sources of the confusion. For example, in Example 1, the model classified an M1_S3 (Reviewing items of previous research) sentence as M3_S5 (Announcing and discussing results). Although the framework in the prompt indicated that Move 3 involves discussion of the present study by prefacing the Move description with "Presenting the present work via", the model may have not associated that part of the description as constraining the subsequent descriptions of the steps. Thus, we added an explicit phrase in the step description to specify the boundary of the move (i.e., M3_S5 – Announcing and discussing the results/principal outcomes of the present research). In a similar fashion, we refined the descriptions of all moves and steps to clearly indicate the boundaries of each move and the scope of each step. At the move level, we revised "Establishing a research territory" as "Establishing the broader research territory within which the present study is situated" for Move 1, and "Establishing a niche" as "Identifying a niche in previous research" for Move 2. Examples of step description revisions included expanding "indicating a gap" to "indicating a gap in previous research" and "providing justification" to "providing justification of the present research area based on previous research".

Example 1
True Code: M1_S3 (Reviewing items of previous research)
Predicted Code: M3_S5 (Announcing and discussing results)

> Meara & Buxton argue that vocabulary recognition from various frequency bands is an efficient way of quantifying the number of words L2 learners actually know.

Misclassifications at the step level were most common between M1_S2 and M1_S3, as illustrated in Example 2. In the manual coding, M1_S3 was designated for units focusing on a specific study while establishing the research territory. To address this confusion, we refined the description of M1_S3 as "Reviewing a specific previous research study (reviewing one specific study, and human names typically indicate specific studies referenced within the article)" in an effort to clarify for the model that this step should be applied to units reviewing a particular previous research study and that the presence of author names could be a helpful indicator of this step. Overall, the prompt refinement process resulted in step descriptions with increased specificity. With the refined prompt (see the

Appendix), the performance of the base model on the validation improved. For move classification, it achieved an accuracy of 37.9% and weighted average precision, recall, and F1 scores of 58.7%, 37.9%, and 43.1%, respectively. For step classification, it achieved an accuracy of 14.5% and weighted average precision, recall, and F1 scores of 30.2%, 14.5%, and 16.7%, respectively. McNemar's tests indicated a significant improvement in move classification accuracy ($\chi^2 = 7.374, p = 0.007$) but not step classification accuracy ($\chi^2 = 1.641, p = 0.200$).

Example 2
True Code: M1_S2 (Making generalizations about research area)
Predicted Code: M1_S3 (Reviewing items of previous research)

> Research within the field of interlanguage pragmatics (ILP) has established that L2 pragmatics can be effectively taught.

The base model's performance was subsequently evaluated on the test set with the initial and refined prompts in zero-shot and three-shot settings to assess the impact of prompt refinement and few-shot learning, both separately and in tandem. In the three-shot settings, we provided the model with three manually annotated introduction sections (totaling 38 sentences) randomly sampled from the training set. The results of this evaluation are presented in Table 2 and visualized in Fig. 2 (for move classification) and Fig. 3 (for step classification).

With the initial prompt in a zero-shot setting, the base model achieved accuracy and F1 scores of 42.9% and 49.5% for move classification and of 17.0% and 20.1% for step classification. Integrating three-shot learning with the initial prompt led to some performance improvement, with accuracy and F1 scores increased to 46.2% and 51.4% for move classification and to 18.7% and 19.6% for step classification. However, the difference in accuracy was not statistically significant for either move ($\chi^2 = 0.543, p = 0.461$) or step classification ($\chi^2 = 0.121, p = 0.728$). Replacing the initial prompt with the refined prompt also led to some performance improvement, with accuracy and F1 scores increased to 50.5% and 54.0% for move classification and to 19.8% and 22.8% for step classification, but the change in accuracy was again insignificant for either move ($\chi^2 = 3.521, p = 0.061$) or step classification ($\chi^2 = 0.410, p = 0.522$). Finally, combining the refined prompt with three-shot learning led to the greatest performance improvement, with accuracy and F1 scores increased to 52.7% and 58.3% for move classification and to 25.3% and 27.2% for step classification. Compared to the accuracy of base model with the initial prompt in a zero-shot setting, the improvement was statistically significant for both move ($\chi^2 = 4.817, p = 0.028$) and step classification ($\chi^2 = 5.600, p = 0.018$). Notably, the base model achieved a higher precision than recall in all four experimental settings. Three-shot learning and prompt refinement led to recall improvement for both move and step classification. However, they resulted in lower precision for move classification and mixed changes in recall for step classification.

4.2. *Impact of fine-tuning GPT3.5*

As shown in Table 2, the fine-tuned models demonstrated substantially improved performance for move (Fine-tuned Model 1: precision = 92.5%, recall = 91.2%, F1 score = 91.8%, accuracy = 91.2%; Fine-tuned Model 2: precision = 92.2%, recall = 92.3%, F1 score = 92.2%, accuracy = 92.3%) and step classification (Fine-tuned Model 1: precision = 78.5%, recall = 75.3%, F1 score = 76.2%, accuracy = 75.3%; Fine-tuned Model 2: precision = 80.6%, recall = 80.2%, F1 score = 80.2%, accuracy = 80.2%). The accuracies were significantly higher than that achieved by the base model with the refined prompt in a three-shot setting for both move (Fine-tuned Model 1: $\chi^2 = 59.046, p < 0.001$; Fine-tuned Model 2: $\chi^2 = 59.568, p < 0.001$) and step classification (Fine-tuned Model 1: $\chi^2 = 76.676$, $p < 0.001$; Fine-tuned Model 2: $\chi^2 = 91.000, p < 0.001$). The two fine-tuned models' performance did not differ significantly for either move ($\chi^2 = 0.050, p = 0.823$) or step ($\chi^2 = 3.064, p = 0.080$).

Table 3 presents the performance of the best-performing model – Fine-tuned Model 2 (with 80 samples for training) on each move. Notably, for Moves 1 and 3, the model achieved precision, recall, and F1 scores between 93.8% and 100%. However, its performance in classifying Move 2 was lower, with the three metrics hovering around 70%. Table 4 presents the performance of the fine-tuned model on each step. For seven steps, the model achieved precision, recall, and F1 score all over 80%. Meanwhile, it performed the worst on M1_S1a (Claiming centrality or value of research area), M2_S1c (Question raising), M2_S2 (Providing justification), and M3_S2a (Presenting research questions or hypotheses) with precision, recall, and F1 scores all at or below 50%.

To assess the stability and consistency of the base and fine-tuned models with the same inputs under identical conditions (using the

Table 2
Classification task evaluation results.

		Initial 0 shot	Initial 3 shot	Refined 0 shot	Refined 3 shot	Fine-tuned 1	Fine-tuned 2
Precision	Move	77.1%	73.6%	73.9%	72.7%	92.5%	92.2%
	Step	48.4%	59.7%	54.0%	43.2%	78.5%	80.6%
Recall	Move	42.9%	46.1%	50.5%	52.7%	91.2%	92.3%
	Step	17.0%	18.7%	19.8%	25.3%	75.3%	80.2%
F1 score	Move	49.5%	51.4%	54.0%	58.3%	91.8%	92.2%
	Step	20.1%	19.6%	22.8%	27.2%	76.2%	80.2%
Accuracy	Move	42.9%	46.2%	50.5%	52.7%	91.2%	92.3%
	Step	17.0%	18.7%	19.8%	25.3%	75.3%	80.2%

Note. Initial = the initial prompt, 0shot = zero-shot, Refined = the refined prompt, 3shot = three-shot. Fine-tuned 1 = the fine-tuned model with 40 training samples, Fine-tuned 2 = the fine-tuned model with 80 training samples.

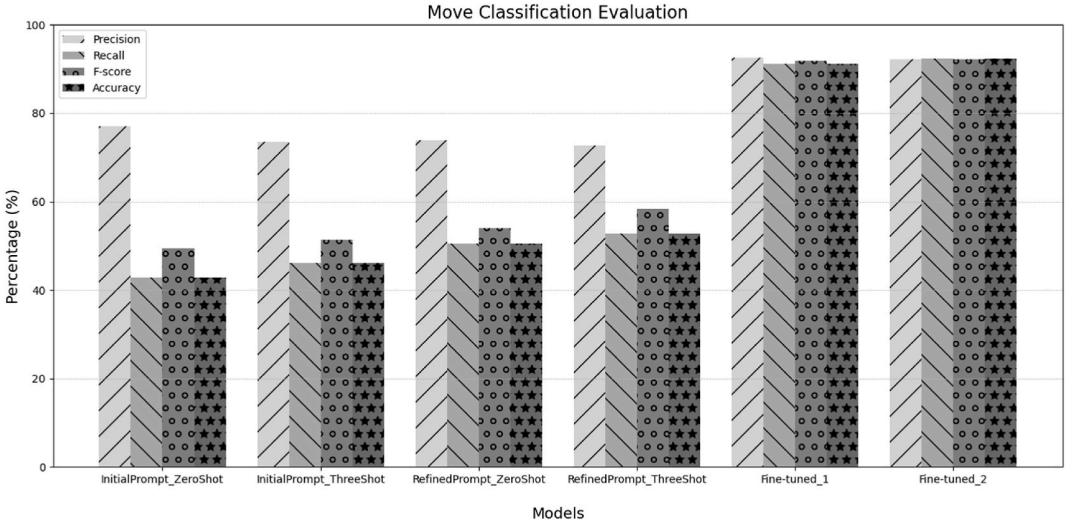

Fig. 2. Move classification evaluation.

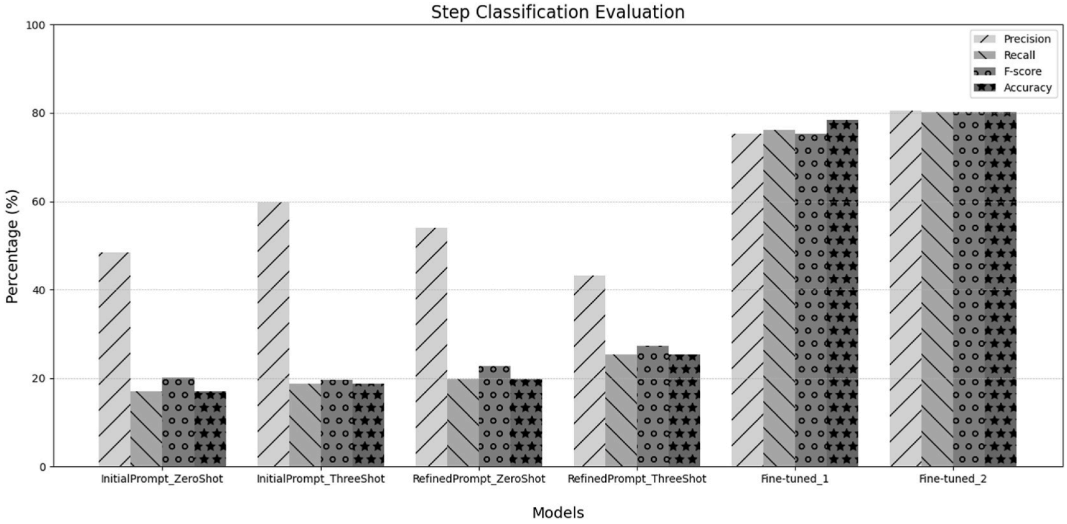

Fig. 3. Step classification evaluation.

Table 3
Precision, recall, and F1 score of Fine-tuned Model 2 for move classification.

Move	Precision	Recall	F1 score	Support
M1	95.6%	93.9%	94.7%	115
M2	71.4%	68.2%	69.8%	22
M3	93.8%	100.0%	96.8%	45
Weighted average	92.2%	92.3%	92.2%	182

Note. Support denotes the number of actual occurrences.

same model parameters and prompt), we conducted three additional tests with the refined prompt for both the base model (three-shot setting) and Fine-tuned Model 2, initiating each test in a new session to ensure no retention of previous data. The outcomes, as shown in Table 5, were relatively consistent. The base model achieved F1 scores ranging from 56.2% to 63.5% for move classification and from 23.2% to 29.7% for step classification. Meanwhile, the fine-tuned model displayed greater consistency with F1 scores ranging from 91.9% to 93.0% for move classification and from 78.6% to 81.7% for step classification.

Subsequent validation was conducted using the two additional test sets. For the Anthropology test set, the model achieved a precision of 84.5%, recall of 86.7%, F1 score of 85.0%, and accuracy of 87.7% for move classification, along with a precision of 72.6%,

Table 4
Precision, recall, and F1 score of Fine-tuned Model 2 for step classification.

Step	Precision	Recall	F1 score	Support
M1_S1a	28.6%	33.3%	30.8%	6
M1_S1b	85.7%	100.0%	92.3%	6
M1_S2	88.6%	84.9%	86.7%	73
M1_S3	82.8%	80.0%	81.4%	30
M2_S1b	61.5%	66.7%	64.0%	12
M2_S1c	0.0%	0.0%	0.0%	3
M2_S1e	66.7%	66.7%	66.7%	3
M2_S2	40.0%	50.0%	44.4%	4
M3_S1	94.1%	88.9%	91.4%	18
M3_S2a	50.0%	100.0%	66.7%	1
M3_S3	100.0%	90.9%	95.2%	11
M3_S4a	66.7%	66.7%	66.7%	6
M3_S5	100.0%	100.0%	100.0%	1
M3_S7	80.0%	100.0%	88.9%	4
M3_S8	66.7%	100.0%	80.0%	4
Weighted Average	80.6%	80.2%	80.2%	182

Note. Support denotes the number of actual occurrences; not all steps in the framework were present in the test set.

Table 5
Outcome consistency.

		Move				Step			
Model	Attempts	1	2	3	4	1	2	3	4
Base	Precision	72.7%	76.5%	82.4%	74.5%	43.2%	60.9%	57.8%	53.1%
	Recall	52.7%	51.6%	59.3%	52.2%	25.3%	19.2%	25.8%	23.1%
	F1 score	58.3%	56.2%	63.5%	56.5%	27.2%	23.2%	29.7%	24.4%
	Accuracy	52.7%	51.6%	59.3%	52.2%	25.3%	19.2%	25.8%	23.1%
Fine-tuned	Precision	92.2%	92.7%	93.1%	92.1%	80.6%	80.4%	83.6%	81.4%
	Recall	92.3%	92.3%	93.4%	91.8%	80.2%	77.4%	81.9%	81.3%
	F1 score	92.2%	92.5%	93.0%	91.9%	80.2%	78.6%	81.7%	80.9%
	Accuracy	92.3%	92.3%	91.8%	93.4%	80.2%	77.5%	81.9%	81.3%

Note. The outcomes of the 1st attempt are the ones reported in Tables 2–4.

recall of 75.3%, F1 score of 72.2%, and accuracy of 75.3% for step classification. For the Psychology test set, the model achieved a precision of 92.0%, recall of 91.2%, F1 score of 91.1%, and accuracy of 91.2% for move classification, along with a precision of 75.5%, recall of 71.6%, F1 score of 71.7%, and accuracy of 71.6% for step classification.

5. Discussion

The findings for the first research question uncovered several intriguing aspects pertaining to the available domain (i.e., genre analysis here) adaptation methods for LLMs such as ChatGPT. First, our analysis of the confusion matrix of the classifications on the validation set by the initial prompt in a zero-shot setting showed that the lack of prompt specificity could hurt the model's classification accuracy and an important goal of prompt refinement should be to improve prompt specificity. This analysis echoes Giray's (2023) point that when designing prompts for LLMs, academic writers and researchers should be aware that insufficient specificity could lead to ambiguity and result in diminished accuracy. With a higher level of specificity, our refined prompt helped improve the accuracy of the model's classifications. Although the improvement turned out to be not statistically significant, our analysis and results nevertheless support Fatouros et al.'s (2023) conclusion from their experiments to exploit ChatGPT for sentiment analysis that adept prompt engineering and thorough prompt evaluation are critical before model deployment.

Second, our findings showed that three-shot learning alone did not significantly improve the performance of the initial prompt, but combining it with prompt refinement led to further, statistically significant improvement in the model's performance for both move and step classification. The overall accuracy, however, was still low even for both move (52.7%) and step classification (25.3%) with both domain adaptation methods. Loukas et al. (2023) reported much higher F1 scores for GPT-3.5 in a one-shot setting and GPT-4 in a three-shot setting in classifying customer service queries into 77 intention categories (over 70% and 80% respectively), and Lossio-Ventura et al. (2024) also reported a much higher accuracy (87%) for ChatGPT in a zero-shot setting in a three-level (positive, neutral, negative) sentiment analysis task. The much higher performance of ChatGPT on these tasks suggests that rhetorical move-step annotation may represent a more challenging and/or complex classification task for ChatGPT, as it involves a deeper understanding of a larger context to determine the specific functions of sentences, while the intention classification and sentiment analysis tasks both deal with short texts or individual sentences as standalone units.

The accuracy levels achieved by ChatGPT for move and step classification also appeared lower than those reported for supervised learning models such as AntMover (Anthony & Lashkia, 2003) and RWT (Cotos & Pendar, 2016), although it is important to contextualize these differences. AntMover was trained on 554 sentences and achieved an average accuracy of 68% across six step-categories; RWT was trained on 650 RA introductions (15,460 sentences) and achieved overall accuracies of 72.6% for move classification and 72.9% for step classification. In contrast, the three-shot setting provided ChatGPT with three annotated RA introductions totaling 38 sentences only. Overall, these performance differences show that more rigorous domain adaptation than prompt refinement and few-shot learning is necessary to enable ChatGPT to successfully handle the domain-specific task of rhetorical move-step annotation.

The findings from our second research question revealed that fine-tuning the base model led to substantially higher accuracies, with 91.2% (Fine-tuned Model 1) and 92.3% (Fine-tuned Model 2) for move classification and 75.3% (Fine-tuned Model 1) and 80.2% (Fine-tuned Model 2) for step classification. These accuracies surpassed those reported for AntMover, even with a much larger number of step categories. They also surpassed the accuracies reported for the RWT (Cotos & Pendar, 2016), even though the size of the training data (40 RA introductions with 795 sentences for Fine-tuned Model 1 and 80 RA introductions with 1556 sentences for Fine-tuned Model 2) was much smaller than that used to train the RWT (650 RA introductions with 15,460 sentences). These findings echo those of Brown et al. (2020), who observed that GPT models require considerably less training data to perform well in NLP tasks than traditional NLP models. The potential of GPT models to effectively tackle challenging domain-specific classification tasks with small-scale domain-specific training data can dramatically alleviate researchers of laborious efforts in manual coding.

The higher accuracies of both Fine-tuned Model 1 and Fine-tuned Model 2 than those achieved by existing systems highlight the efficiency of fine-tuned models. Whereas there was no significant performance difference between the two fine-tuned models, accuracy did trend up from Model 1 to Model 2, suggesting that using larger training sets could potentially lead to further accuracy improvement. This observation sets the stage for future research to identify the optimal balance between maximizing accuracy and minimizing training sample size in the context of fine-tuning ChatGPT for move analysis, similar to explorations of such a balance in other domains (e.g., Pecher et al., 2024).

Despite the substantial performance improvement, Fine-tuned Model 2 encountered some difficulties in accurately classifying Move 2 and the following four steps: M1_S1a (Claiming centrality or value of research area), M2_S1c (Question raising), M2_S2 (Providing justification), and M3_S2a (Presenting research questions or hypotheses). The model's lower performance on these moves and steps could be attributed to two reasons. First, the training data for Move 2 and these four steps were relatively scarce. In our training dataset, Move 1 and Move 3 were represented about 4 and 2.25 times more than Move 2, respectively, and M2_S1c, M2_S2, and M3_S2a were each represented with fewer than 50 instances. Cotos and Pendar (2016) similarly pointed out the difficulty in correctly classifying underrepresented steps in their training dataset. Second, some steps may have fewer explicit linguistic cues than other steps, making it more difficult to automatically identify them. Although not numerous, the training dataset contained 71 instances of M1_S1a (Claiming centrality or value of research area), yet the F1 Score for this step was only 30.8%. This step was similarly challenging to AntMover (accuracy = 28%) (Anthony & Lashkia, 2003), suggesting a lack of explicit linguistic cues for the model to accurately identify this step.

ChatGPT's capability to assign multiple codes to a single sentence addresses a limitation pointed out by Cotos and Pendar (2016) in existing tools such as AntMover and the RWT that can only assign a single move or step category to each sentence. This is a desirable feature as some sentences in a text may serve multiple rhetorical functions. In the training set, 80 of the 1556 sentences (5.14%) were multifunctional. In the test set, 13 out of the 182 sentences (or 7.14%) were multifunctional, among which 10 were assigned multiple tags by the fine-tuned model. Of the 10 sentences, the first predicted tag matched the primary tag in seven cases, and the second predicted tag matched the primary tag in three cases. Additional research is still needed to further improve the model's performance on multifunctional sentences and to refine the methodology for aggregating sentences as rhetorical chunks based on sentence-level annotations.

The results on outcome consistency demonstrate that with unchanged prompts and parameters, the performance of the models remained largely stable, particularly with Fine-tuned Model 2, which showed better stability than the base model. Subtle changes in parameter settings or prompts can cause ChatGPT to produce different outputs, a characteristic of its non-deterministic nature (Reiss, 2023). Therefore, maintaining consistent parameters and prompts is essential to minimize this randomness. Reiss (2023) highlighted inconsistencies in ChatGPT's zero-shot capabilities for text annotation in a task of classifying 234 website texts into news or not news, noting variability under different parameter settings or with slight alterations in prompts as well as some inconsistency even with repeated identical inputs. In contrast, our experiments showed that both the base and fine-tuned models maintained relatively stable accuracy levels, with the fine-tuned model demonstrating particular consistency. This difference may stem from our use of a 3-shot setting and fine-tuning, whereas Reiss tested a zero-shot setting. Fine-tuning likely reduced output variability by focusing the model's "knowledge" and response patterns on a narrower subset of styles. For example, we used highly consistent data for fine-tuning, which helps stabilize the model's behavior across similar inputs. In contrast, the base ChatGPT model, trained on a diverse range of internet texts, is optimized for generalization rather than specific task consistency. Additionally, the "temperature" parameter impacts output randomness, where lower settings yield more deterministic outputs, and higher settings increase diversity and unpredictability. Our initial tests used a default temperature of 1, but further reductions could decrease randomness. The impact of temperature settings on output reliability and validity requires further exploration. Therefore, for researchers using this approach for move analysis, several considerations are crucial. First, uniformity in parameters, model settings, and prompts, along with targeted fine-tuning, can significantly reduce inconsistency. Second, it is desirable for researchers to meticulously document all parameters and prompts to ensure replicability. Third, validation of output consistency is strongly recommended to confirm model reliability. Despite the minor inconsistencies observed, the significantly improved accuracy, efficiency (requiring less data to achieve comparable accuracy), and easy accessibility to researchers and teachers highlight the considerable potential of using ChatGPT for genre analysis and justify continued exploration of this approach.

The results of the tests conducted with the two additional test sets from Anthropology and Psychology demonstrated that the model's capability for move annotation can reliably generalize to these two related disciplines, as it achieved accuracies exceeding 87% on both. For step classification, the model achieved accuracies exceeding 71% on both test sets. The slightly lower accuracy for steps may show that similar communicative functions are executed differently in different disciplines, with linguistic cues varying from one discipline to another to fulfill the same communicative objectives. These variations underscore the need for further research to refine this tool, ensuring it effectively captures the nuanced differences across academic fields. Notably, the results on these two test sets surpassed that of AntMover and were comparable to that of the RWT with a much smaller training set, illustrating the model's efficiency.

The capabilities of the fine-tuned model have useful applications in genre analysis research. Genre researchers can fine-tune the base model with substantially smaller manually annotated training sets than previously required to automate rhetorical move-step annotation for different genres using appropriate move-step frameworks. This advancement has the potential to expedite the processes of genre analysis research and expand the scale of such research in terms of sample sizes. Building on reliable move-step frameworks established through manual analyses, genre researchers can capitalize on automated move-step annotations in large-scale follow-up studies to more systematically explore such issues as disciplinary variation and form-function mappings, among others. Meanwhile, researchers concerned with the accuracy of the model's annotations could also use them as a useful first pass or second opinion in their in-depth move-step analysis of texts. Our results further demonstrate the potential of using ChatGPT for discourse-level annotations in general, and future studies could explore other types of discourse-level annotations, such as annotations of speech acts, pragmatic functions, and cohesive relations.

Methodologically, a significant advantage of this approach is the relative ease of creating domain-specific fine-tuned models compared to traditional machine learning models, which require extensive feature engineering and coding skills that may be less accessible to many researchers or teachers. Researchers can build and deploy their models using either the OpenAI API with a Python script or in Playground, a user-friendly platform where they can build and easily interact with their fine-tuned model and obtain tagged outputs for their input texts. This method not only simplifies the process but also enhances the accessibility of the tool to the field. Our findings suggest several important methodological recommendations for researchers looking to replicate or build upon this study. First, regarding training sample size, our results suggest that 3-shot learning may be insufficient for move analysis; while 40 training samples might be adequate for move classification and some steps that are frequently present in the training set, categories lacking sufficient training data or those with subtle linguistic cues may require larger datasets. Meanwhile, our results showed that with 80 training samples, the performance of the model improved but only slightly. Further research is needed to explore the optimal balance between minimal training data and maximal accuracy. Moreover, when applied to different genres, the number of categories (different levels of move and step) may influence the optimal training set size, as indicated by the different accuracies in our model's move and step classifications. Therefore, researchers should determine the optimal sample size for each genre and may also consider exploring how the number of categories interacts with training size. Second, crafting precise prompts through a meticulous refinement process is essential to ensure that steps or moves are clearly defined and any potentially overlapping definitions are distinctly distinguished. Researchers can utilize confusion matrices to pinpoint areas of the prompts that require refinement and experiment with various prompts to identify those that yield the most accurate results. Future research should investigate the optimal level of specificity by testing different versions of prompts to prevent overfitting and enhance accuracy. This approach will not only improve the framework but also enhance the clarity and understanding for human annotators. Third, using consistently formatted inputs can significantly enhance consistency in model outputs. Once an optimal prompt has been established, maintaining a uniform format for prompts, training data, and user input is crucial, as this consistency directly influences the precision of the outputs. Our experiments showed that such consistency was vital for reducing variability in the model's behavior and ensuring reliable results. Fourth, robust validation tests are crucial to ensure the reliability of the model's outputs. It is important to repeatedly test the model with identical inputs, prompts, and parameter settings, and to accurately document the range of observed accuracies to assess the model's consistency. As mentioned earlier, researchers should clearly report all parameters used for training and testing the model and prompts so that future researchers can replicate and build upon their research. Additionally, using large test sets, when feasible, is recommended to enhance the validity of the findings and achieve greater rigor.

In terms of genre-based writing pedagogy, the model can serve as a valuable supplementary resource in guided corpus-based genre analysis activities designed to enhance learners' genre competence. The automated annotations provided by the model can be a useful starting point for novice academic writers as they engage with genre analysis of academic texts. Directly working on genre analysis of such texts immediately following instructor explanations of the move-step framework can be rather overwhelming to learners new to the task. Instructors can ask learners to interact with the automated tool with authentic texts to see the framework in action, reducing initial intimidation. As they become adequately familiar with the framework, learners can collaboratively investigate the accuracy of the automated annotations. They can also be guided to identify move-step sequences, frequent move-steps, and linguistic features associated with different move-steps in automatically annotated expert texts. Such hands-on experiences can help deepen learners' understanding of the move-step framework and genre-specific practices within their disciplines. Additionally, learners can also use the model to analyze their own texts and assess how their own writing aligns with typical rhetorical structures of the target genre.

6. Conclusion

This study investigated the potential of utilizing ChatGPT for rhetorical move-step analysis and the impact of prompt refinement, few-shot learning, and fine-tuning on its accuracy for move and step classification. Our results showed that prompt refinement and few-shot learning helped improve the performance of the model but fine-tuning achieved more competitive and satisfactory accuracy with relatively small training sets. The power to accurately automate move-step annotation has important implications for more comprehensive and larger-scale genre research. Fine-tuned models can also function as accessible and interactive tools to facilitate the design of corpus-based genre analysis activities in genre-based writing pedagogy. Our study also allowed us to offer methodological recommendations for using ChatGPT in research.

As an early attempt to adopt the GPT model for rhetorical move-step analysis, however, our study has several limitations, some of which can be addressed in future research. First, we limited our investigation to a single discipline (with additional test sets from two related disciplines) to obtain an initial assessment of ChatGPT's ability to tackle this highly complex domain-specific task. Future studies can expand the scope of the evaluation using more data from diverse disciplines to understand the generalizability of the effectiveness of the model and methodology. Second, for prompt refinement, our exploration was confined to two prompts. Subsequent studies could adopt an iterative prompt refinement strategy with increasingly more specific prompts to ascertain the optimal balance between prompt specificity and generality to maximize accuracy while preventing overfitting, as Giray (2023) advised. Third, we did not examine the effects of parameter setting or training data size on the performance of the fine-tuned model. Future studies could experiment with different parameter settings, sample sizes, and numbers of move-step categories to shed light on the optimal parameter configuration and training sample size to maximize efficiency and accuracy. Lastly, we assessed model consistency by repeating the tests three times, but the consistency of both the base model and the fine-tuned model may warrant further exploration with varying temperature settings and more repetitions.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Minjin Kim: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xiaofei Lu:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors have no conflict of interest to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jeap.2024.101422.

References

Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Transactions on Professional Communications, 46*(3), 185–193. https://doi.org/10.1109/TPC.2003.816789

Balkus, S. V., & Yan, D. (2023). Improving short text classification with augmented data using GPT-3. *Natural Language Engineering*, 1–30. https://doi.org/10.1017/S1351324923000438

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins Publishing.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Casal, J. E., & Kessler, M. (2023). Rhetorical move-step analysis. In M. Kessler, & C. Polio (Eds.), *Conducting genre-based research in applied linguistics* (pp. 82–104). Routledge. https://doi.org/10.4324/9781003300847-7.

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12* (1), 33–43. https://doi.org/10.1016/j.jeap.2012.11.002

Cotos, E., Huffman, S., & Link, S. (2017). A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes, 46*, 90–106. https://doi.org/10.1016/j.esp.2017.01.001

Cotos, E., Huffman, S., & Link, S. (2020). Understanding graduate writers' interaction with and impact of the Research Writing Tutor during revision. *Journal of Writing Research, 12*(1), 187–232. https://doi.org/10.17239/jowr-2020.12.01.07

Cotos, E., & Pendar, N. (2016). Discourse classification into rhetorical functions for AWE feedback. *CALICO Journal, 33*(1), 92–116. https://doi.org/10.1558/cj.v33i1.27047

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Dong, J., & Lu, X. (2020). Promoting discipline-specific genre competence with corpus-based genre analysis activities. *English for Specific Purposes, 58*, 138–154. https://doi.org/10.1016/j.esp.2020.01.005

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. *Machine Learning with Applications, 14*, Article 100508. https://doi.org/10.1016/j.mlwa.2023.100508

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. *International Journal of Intelligent Technologies and Applied Statistics, 11*(2), 105–111. https://doi.org/10.6148/IJITAS.201806_11(2).0003

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Halliday, M. A. K. (1978). *Language as social semiotics*. Edward Arnold.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023* (pp. 294–297). https://doi.org/10.1145/3543873.3587368

Hyland, K. (2007). Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing, 16*(3), 148–164. https://doi.org/10.1016/j.jslw.2007.07.005

Hyland, K. (2015). Genre, discipline and identity. *Journal of English for Academic Purposes, 19*, 32–43.

Hyon, S. (1996). Genres in three traditions: Implications for second language teaching. *Tesol Quarterly, 30*, 693–722. https://www.jstor.org/stable/3587930.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes, 24*(3), 269–292. https://doi.org/10.1016/j.esp.2004.08.003

Kavzoglu, T. (2017). Chapter 33—object-oriented random Forest for high resolution land cover mapping using quickbird-2 imagery. In P. Samui, S. Sekhar, & V. E. Balas (Eds.), *Handbook of neural computation* (pp. 607–619). Academic Press. https://doi.org/10.1016/B978-0-12-811318-9.00033-8.

Kessler, M. (2020). A text analysis and gatekeepers' perspectives of a promotional genre: Understanding the rhetoric of Fulbright grant statements. *English for Specific Purposes, 60*, 182–192.

Kessler, M., & Polio, C. (2023). Introduction. In M. Kessler, & C. Polio (Eds.), *Conducting genre-based research in applied linguistics* (pp. 1–10). Routledge. https://doi.org/10.4324/9781003300847-1.

Kim, M., Qiu, X., & Wang, Y. (2024). Interrater agreement in genre analysis: A methodological review and a comparison of three measures. *Research Methods in Applied Linguistics, 3*(1), 100097. https://doi.org/10.1016/j.rmal.2024.100097

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion, 99*, Article 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Unpublished doctoral dissertation]. Georgia State University. https://doi.org/10.57709/8501051.

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, 1–33. https://doi.org/10.1007/s10639-023-12249-8

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., … He, L. (2021). A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*. https://doi.org/10.48550/arXiv.2008.00364

Liu, J., Yang, S., Peng, T., Hu, X., & Zhu, Q. (2023). ChatICD: Prompt learning for few-shot ICD coding through ChatGPT. In *2023 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 4360–4367). https://doi.org/10.1109/BIBM58861.2023.10385482

Lossio-Ventura, J. A., Weger, R., Lee, A. Y., Guinee, E. P., Chung, J., Atlas, L., Linos, E., & Pereira, F. (2024). A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: Sentiment analysis of COVID-19 survey data. *JMIR Mental Health, 11*, Article e50150. https://doi.org/10.2196/50150

Loukas, L., Stogiannidis, I., Malakasiotis, P., & Vassos, S. (2023). Breaking the bank with ChatGPT: Few-shot text classification for finance. In C.-C. Chen, H. Takamura, P. Mathur, R. Sawhney, H.-H. Huang, & H.-H. Chen (Eds.), *Proceedings of the fifth workshop on financial technology and Natural Language processing and the second multimodal AI for financial forecasting* (pp. 74–80). https://aclanthology.org/2023.finnlp-1.7.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods, 54*(3), 1444–1460. https://doi.org/10.3758/s13428-021-01675-6

Lu, X., Yoon, J., & Kisselev, O. (2021a). Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes, 61*, 63–83. https://doi.org/10.1016/j.esp.2020.10.001

Lu, X., Yoon, J., Kisselev, O., Casal, J. E., Liu, Y., Deng, J., & Nie, R. (2021b). Rhetorical and phraseological features of research article introductions: Variation among five social science disciplines. *System, 100*, 102543. https://doi.org/10.1016/j.system.2021.102543

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Moreno, A. I., & Swales, J. M. (2018). Strengthening move analysis methodology towards bridging the function-form gap. *English for Specific Purposes, 50*, 40–63. https://doi.org/10.1016/j.esp.2017.11.006

OpenAI. (2024). *OpenAI playground (May 4 version) [Large language model]*. https://platform.openai.com/playground.

Paltridge, B. (1994). Genre analysis and identification of textual boundaries. *Applied Linguistics, 15*(3), 288–299. https://doi.org/10.1093/applin/15.3.288

Pandey, R. K., Dahiya, A. K., Pandey, A. K., & Mandal, A. (2022). Optimized deep learning model assisted pressure transient analysis for automatic reservoir characterization. *Petroleum Science and Technology, 40*(6), 659–677. https://doi.org/10.1080/10916466.2021.2007122

Park, S., Jeon, J., & Shim, E. (2021). Exploring request emails in English for business purposes: A move analysis. *English for Specific Purposes, 63*, 137–150. https://doi.org/10.1016/j.esp.2021.03.006

Parkinson, J. (2017). The student laboratory report genre: A genre analysis. *English for Specific Purposes, 45*, 1–13. https://doi.org/10.1016/j.esp.2021.03.006

Pecher, B., Srba, I., & Bielikova, M. (2024). *Fine-tuning, prompting, in-context learning and instruction-tuning: How many labelled samples do we need?*. arXiv preprint arXiv:2402.12819. https://doi.org/10.48550/arXiv.2402.12819.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reiss, M. V. (2023). *Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark*. arXiv preprint arXiv:2304.11085. https://doi.org/10.48550/arXiv.2304.11085.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications*. Cambridge University Press.

Tardy, C. M. (2016). *Beyond convention: Genre innovation in academic writing*. University of Michigan Press.

Tessuto, G. (2015). Generic structure and rhetorical moves in English-language empirical law research articles: Sites of interdisciplinary and interdiscursive cross-over. *English for Specific Purposes, 37*, 13–26. https://doi.org/10.1016/j.esp.2014.06.002

Wachowiak, L., & Gromann, D. (2023). Does GPT-3 grasp metaphors? Identifying metaphor mappings with generative language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1018–1032). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.58.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., … Le, Q. V. (2022). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652, 2022*. https://doi.org/10.48550/arXiv.2109.01652

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Yoon, J., & Casal, J. E. (2020). Rhetorical structure, sequence, and variation: A step-driven move analysis of applied linguistics conference abstracts. *International Journal of Applied Linguistics, 30*(3), 462–478. https://doi.org/10.1111/ijal.12300

Minjin Kim is a doctoral candidate in Applied Linguistics at The Pennsylvania State University. Her research interests include corpus linguistics, second language acquisition, English for Academic Purposes, and intelligent computer-assisted language learning. Her recent work has appeared in *English for Specific Purposes*, *The Modern Language Journal*, and *Research Methods in Applied Linguistics*. ORCID: https://orcid.org/0000-0001-9935-7867

Xiaofei Lu is Professor of Applied Linguistics and Asian Studies at The Pennsylvania State University. His research interests are primarily in corpus linguistics, English for Academic Purposes, second language writing, and intelligent computer-assisted language learning. He is the author of *Corpus Linguistics and Second Language Acquisition: Perspectives, Issues, and Findings* (2023, Routledge).