國立臺灣大學電機資訊學院暨中央研究院

資料科學學位學程

碩士論文

Data Science Degree Program

College of Electrical Engineering and Computer Science

National Taiwan University and Academia Sinica

Master's Thesis

大型語言模型於專家標記任務之局限性

Strong Large Language Models are Weak Expert Annotators

曾郁珉

Yu-Min Tseng

指導教授: 陳信希 博士、王釧茹 博士

Advisor: Hsin-Hsi Chen, Ph.D., Chuan-Ju Wang, Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202501476

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

大型語言模型於專家標記任務之局限性

Strong Large Language Models are Weak Expert Annotators

本論文係 曾郁珉（R11946009）在國立臺灣大學資料科學學位學程完成之碩士學位論文，於民國 113 年 7 月 8 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Data Science Degree Program on 8 July 2024 have examined a Master's thesis entitled above presented by YU-MIN TSENG (R11946009) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

陳信希 （指導教授 Advisor）

張智東 （指導教授 Advisor）

葉丙成

張嘉惠

學位學程主管 Director: 謝尚昀

誌謝

碩班三年的時光晃眼而過，感謝每個陪伴我走過這段旅程的人們，沒有你們的支持、幫助、鼓舞，我無法想像會是如何的掙扎與徬徨。

謝謝我的指導教授陳信希老師，總是給予我滿心信任，在我猶豫是否該繼續在 ESG 題目往下做，或是轉往更廣泛的語言模型研究時，相信並支持我的想法；在我提出需要花錢跑實驗時，也是沒有多問的慷慨給予經費；更是在我成功投稿上論文後，鼓勵我出國參加國際會議進行發表。在這三年中，我看到的是老師對學術的無限熱忱，在力所能及的範圍內，提供給我們適合做研究的環境與幫助，衷心感謝老師讓我成為自然語言實驗室的其中一員。

同時，我也要感謝我的另外一位指導教授王釧茹老師，給予我產學合作的機會，讓我學習到如何將研究帶入業界。謝謝老師總是一再的稱讚我並朝我遞出橄欖枝，這個肯定對我意義重大，讓我燃起一抹希望、相信自己或許真的還不賴，成為我決定想繼續深造的一大動力，很榮幸能獲得老師的肯定以及成為 CFDA 實驗室的其中一員。

我還要特別感謝重吉學長像一盞明燈一樣的指引我，在我研究受阻時，盡可能不改變我想嘗試的方向，分析利弊、提出問題、即時導正；在我對未來迷茫時，給予我多方的建議參考，十分感謝學長亦師亦友的鼓舞及提攜。另外，我也想謝謝安孜學姊及瀚萱學長，總是在禮拜一的 Meeting 中給予我研究上不同的見解及

doi:10.6342/NTU202501476

解惑，我會特別懷念每週吸收研究知識的時光。

感謝我的實驗室夥伴們：謝謝建宏、聖倫、柏君學長每次的幫忙、解惑，以及每次研究上的討論碰撞；謝謝大我一屆的學長學姊們，特別是韋霖學長、恬儀學姊、和羿寧學姊，在一年的期間跟我聊天、討論研究，成為我最好詢問各種碩班事情的好對象；謝謝同屆的承光、彥錞、晉毅、奕崴、家誠的同窗情誼，我永遠記得第一次見面時大家的害羞內向，到後來一同修課、一同趕 Deadline、一同參加畢業典禮，慢慢變更加熟悉彼此互相 cover；謝謝下一屆的學弟學妹們總是跟我天南地北的聊天，成為我的開心果們；謝謝實驗室助理又慈，每次踏入實驗室總先看到你的背影，在忙著幫我們處理實驗室的設備、報銷、和其他雜事，在閒暇時也會跟我一同取暖，聊聊生活、未來、跟心事。謝謝大家的陪伴，我會永遠記得在 301 的點點滴滴、也會記得我們 2023 年夏天在宜蘭 lab 出遊的快樂時光。

最後，感謝我的摯友們，特別是阿青、踢那、小碩，總是聆聽我的煩惱、當我的快樂泉源，很慶幸我的台北碩班時光有你們陪伴；感謝我的家人們：哲仁、呂大媽、曾小弟總是無條件的支持我，是你們在後面為我撐起整個世界，讓我無後顧之憂的為未來努力奮鬥。最後的最後，再次衷心感謝所有陪伴我的、支持我的、相信我的每個人，希望我在這有限的文字內有表達出對你們無限的謝意！

doi:10.6342/NTU202501476

摘要

資料標記旨在對資料進行相關訊息的標記或標籤化，大量研究報告的結果顯示利用大型語言模型作為人類標記者的正向潛在能力。然而，現有研究主要聚焦在經典的自然語言處理任務上，尚未充分探索大型語言模型在需要專家知識領域中作為標記者的表現。在本論文中，我們系統性的評估其在金融、生物醫學及法律三個高度專業化領域的專家級標記者的表現，研究辦法包括單一語言模型以及多語言模型合作在內的綜合方法，以評估其性能和可靠性。實驗結果表明，儘管大型語言模型作為標記者表現出一定的前景，但其在不同領域和任務中的表現有顯著的差異。從成本效益的角度來看，我們的分析表明大型語言模型在使用 vanilla 或 CoT 方法時，能夠有效節約成本並保持中庸的表現。然而，儘管其擁有這些優勢，大型語言模型在高度專業化的任務中尚不能完全替代人類專家。據我們所知，本論文為首篇系統性地評估大型語言模型作為專家級標記者表現的研究，提供在專業領域中的實證結果和初步見解。

關鍵字：大型語言模型、大型語言模型擔任資料標記者、語言智能體

doi:10.6342/NTU202501476

Abstract

Data annotation refers to the labeling or tagging of textual data with relevant information. A large body of work has reported positive results on leveraging large language models as an alternative to human annotators. However, existing studies focus on classic NLP tasks, and the extent to which LLMs as data annotators perform in domains requiring expert knowledge remains underexplored. In this work, we present a systematic evaluation of LLMs as expert-level data annotators across three highly specialized domains: finance, biomedicine, and law. We investigate comprehensive approaches, including single LLMs and multi-agent LLM frameworks, to assess their performance and reliability. Our experimental results reveal that while LLMs show promise as cost-effective alternatives to human annotators, their performance varies significantly across different domains and tasks. From a cost-effectiveness perspective, our analysis indicates that LLMs, particularly when using the vanilla or CoT methods, offer substantial savings compared to traditional human annotation processes. Despite these advantages, LLMs are not yet a

doi:10.6342/NTU202501476

direct substitute for human experts in highly specialized tasks. To the best of our knowledge, we present the first systematic evaluation of LLMs as expert-level data annotators, providing empirical results and pilot insights in specialized domains.

Keywords: Large Language Models, LLMs as Annotators, Language Agents

doi:10.6342/NTU202501476

Contents

doi:10.6342/NTU202501476

doi:10.6342/NTU202501476

doi:10.6342/NTU202501476

List of Figures

doi:10.6342/NTU202501476

doi:10.6342/NTU202501476

List of Tables

doi:10.6342/NTU202501476

Chapter 1

Introduction

Data annotation refers to the task of labeling or tagging textual data with relevant information (Tan et al., 2024). For example, adding topic keywords to social media contents. Typically, data annotation is carried out by crowd-sourced workers (*e.g.*, MTurkers) or specialized annotators (*e.g.*, researchers), depending on the tasks, to ensure high-quality annotations. However, the annotating procedures are often costly, time-consuming, and labor-intensive, particularly for tasks that require domain expertise.

With the rise of large language models (LLMs), a series of works have explored using them as an attractive alternative to human annotators (Choi et al., 2024; Ding et al., 2023; He et al., 2023; Zhang et al., 2023). Empirical results show that, in certain scenarios, LLMs such as ChatGPT and GPT-3.5 even outperform master-level MTurk workers, with substantially lower per-annotation cost (Alizadeh et al., 2023; Bansal and Sharma, 2023; Gilardi et al., 2023; Zhu et al., 2023). However, existing studies mainly focus on classic NLP tasks (*e.g.*, sentiment classification, word-sense disambiguation) on general domain datasets. The extent to which LLMs as data annotators perform in domains requiring expert knowledge remains unexplored.

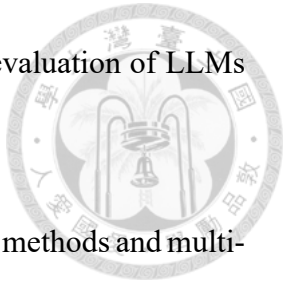
doi:10.6342/NTU202501476

On the other hand, LLMs have exhibited striking performance in a variety of benchmarks, both professional and academic (Achiam et al., 2023; Chen et al., 2021; Hendrycks et al., 2020; Jin et al., 2019; Rein et al., 2023). Leveraging the abundant domain-specific knowledge encoded in the parameters, LLMs could pass exams that require expert-level abilities (Callanan et al., 2023; Choi et al., 2021; Katz et al., 2024; Singhal et al., 2023a,b). These findings prompt the question – Can LLMs apply their parametric knowledge to perform expert-level annotation tasks?

To address this, we investigate three specialized domains: finance, biomedicine, and law. Specifically, we adopt six datasets that *(i)* provide fully-detailed annotation guidelines and *(ii)* are manually labelled by domain experts. We format the annotation task, the guideline, and unlabelled data instances as instructional inputs to the most performant, publicly-available LLMs, and evaluate their annotation results against ground-truth labelled by human experts. Experimental results in our vanilla setting suggest that LLMs show substantial rooms for improvements, with an average of around 35% behind human expert annotators.

Towards a more comprehensive evaluation, we employ a variety of approaches tailored to elicit the capabilities in LLMs, including chain-of-thought (CoT), self-consistency, and self-refine promptings. Additionally, drawing inspiration from how human annotators reach consensus, we introduce a multi-agent annotation framework which incorporates a peer-discussion process for producing annotations. Furthermore, we propose an Environmental, Social, and Governance (ESG) dataset, as well as different language versions to explore the multilingual abilities of LLMs as expert annotators. Lastly, we discuss practical suggestions on leveraging LLMs for expert annotation tasks, from a cost-effectiveness perspective. We summarize our main contributions as follows:

doi:10.6342/NTU202501476



- We present, to the best of our knowledge, the first systematic evaluation of LLMs as expert-level data annotators.
- We explore comprehensive approaches, including prompt-based methods and multi-agent frameworks, across three highly specialized domains.
- We propose DynamicESG, a dataset comprises ESG news articles with multiple language versions to investigate the multilingual ability of LLMs as expert annotators.
- We provide a cost-effectiveness analysis and practical suggestions on leveraging LLMs for expert annotation tasks.

We organize the thesis as follows: Chapter 2 discusses related work. Chapter 3 introduces two types of datasets, namely domain-specific datasets and our proposed dataset, DynamicESG, for the exploration of LLMs as expert annotators. Chapter 4 investigates the capabilities of single LLMs. Chapter 5 explores multi-agent LLMs as annotator groups to simulate the resolution of discrepancies in the annotation process. In Chapter 6, we provide an analysis of multilinguality and cost-effectiveness. Finally, Chapter 7 presents the conclusion, limitations and the future directions.

doi:10.6342/NTU202501476

Chapter 2

Related Work

2.1 LLMs as Annotators

Recent advancements in NLP have leveraged LLMs as effective tools for data annotation, significantly reducing the time and cost associated with traditional human annotation methods. He et al. (2023) introduce AnnoLLM, a system that utilizes LLMs, specifically GPT-3.5, to annotate data through a two-step process of explanation and annotation. AnnoLLM demonstrates the capability to match or even surpass the performance of crowd-sourced human annotators across various NLP tasks. Zhang et al. (2023) propose LLMAAA, which integrates LLMs into an active learning loop, optimizing both the annotation and training processes. This approach not only enhances efficiency but also ensures the reliability of annotations, making it particularly effective for tasks like named entity recognition and relation extraction.

Another innovative approach discussed by Gilardi et al. (2023) and Alizadeh et al. (2023) involves using open-source LLMs for text annotation tasks such as relevance, topic detection, and framing detection. They demonstrate that fine-tuned models can signif-

doi:10.6342/NTU202501476

icantly improve their annotation performance, presenting a cost-effective alternative to commercial models like GPT-3.5. Furthermore, the study by Choi et al. (2024) explores the use of GPT models as multilingual annotators, demonstrating their capability to generate high-quality annotations in multiple languages from a single input. This approach is particularly beneficial for low-resource languages, offering a cost-efficient solution for multilingual dataset construction. Lastly, the work by Bansal and Sharma (2023) highlights the generalization capabilities of LLMs in enhancing NLP models. They introduced a novel sampling strategy for annotating inputs, which significantly improves the accuracy and generalization of models across various domains.

In sum, these studies collectively underscore the potential of LLMs to revolutionize the field of data annotation, providing efficient, scalable, and cost-effective solutions that could potentially replace traditional human-based methods. However, they all focus on general domain annotation tasks, leaving a gap in the exploration of domain-specific applications. While the general domain annotation tasks provide a broad understanding of the capabilities of LLMs, the unique challenges and requirements of specific domains such as legal, medical, or technical fields remain underexplored. Thus, in this thesis, we aim to unveil the capabilities of LLMs as expert annotators.

2.2 Multi-Agent LLM Collaboration

Recently, several research has explored the planning, reflection, and tool-using abilities in LLMs (Shinn et al., 2024; Yao et al., 2024, 2022), potentially paving the way for artificial general intelligence (Bubeck et al., 2023). Thus, LLM-based agent has surged as a fast-growing research field (Wang et al., 2024; Xi et al., 2023). Moreover, the *Multi-Agent*

doi:10.6342/NTU202501476

framework, where multiple LLM-based agents cooperate and communicate with each other to solve complex tasks, has also appeared as a prevalent research direction (Chen et al., 2023c; Du et al., 2023; Hong et al., 2023; Liang et al., 2023b; Tseng et al., 2024).

As identified in prior surveys of LLM-based agents (Guo et al., 2024; Xi et al., 2023), there are two collaboration paradigms in the multi-agent schema: *Cooperative* and *Adversarial*. The cooperative paradigm facilitates information sharing among agents, with some frameworks using message pools to store each agent's current state and ongoing tasks (Chen et al., 2023d; Hong et al.). On the other hand, the adversarial paradigm, including debate, competition, and criticism, enhances the decision-making process and seeks more advantages through adopting opposing perspectives (Chan et al., 2023; Fu et al., 2023).

In this thesis, to enhance annotation performance and mitigate discrepancies during the annotation process, we leverage the benefits of multi-agent LLM collaboration. Specifically, we design multi-agent LLM frameworks within a cooperative paradigm, where multiple LLMs share their information (*i.e.*, annotation and reasoning) with each others. To simulate the real annotation process, we propose two types of discussion processes, namely majority vote and peer discussion, detailed in Section 5.

2.3 AI for Social Good – ESG

The integration of Environmental, Social, and Governance (ESG) factors into AI-based decision-making processes has become increasingly significant, especially in the financial sector. Recent advancements in AI and Natural Language Processing (NLP) have been directed towards understanding and evaluating these non-monetary factors, which are

doi:10.6342/NTU202501476

crucial for sustainable and ethical business practices.

One of the pioneering efforts in this domain is the implications of ESG factors on corporate performance and investment decisions, which have been the focus of an expanding body of research. Foundational research, conducted by MSCI (Giese et al., 2021), identifies governance (G) as the predominant pillar for short-term analysis, representing event risks, while environmental (E) and social (S) indicators become increasingly crucial in the long term, reflecting cumulative risks to performance, such as carbon emissions. Mehra et al. (2022) contribute to the field by introducing ESGBERT, a tool developed by fine-tuning the BERT model for sequence classification and performing a Masked Language Model (MLM) task on an ESG corpus. The experimental results demonstrate ESGBERT's success in learning the ESG context, highlighting its value for various ESG-specific text classification tasks. Further, Raman et al. (2020) automatically generate an ESG relevance score for any corporate discourse, providing a mechanism to gauge and assess the emphasis placed on sustainable business practices. The other interesting study by Lee et al. (2022) points out the challenges associated with calculating ESG scores for small businesses, which often lack extensive data records compared to larger corporations. Distinct from the aforementioned studies, our proposed DynamicESG dataset extends the ESG discourse by focusing on daily news articles to dynamically assess changes in a company's ESG rating. This approach addresses the inherent limitations of the traditional annual update cycle, providing a more nuanced understanding of ESG impacts influenced by real-time events and developments.

To foster advancements in this area, we organize three Multilingual ESG Shared Tasks (ML-ESG) with oversea researchers as part of the 5th, 6th, and 7th Workshop on Financial Technology and Natural Language Processing (FinNLP; Chen et al. 2023a,b,

doi:10.6342/NTU202501476

2024). The ML-ESG datasets includes versions in English, French, Japanese, Korean and Chinese. The ESG issue identification shared task, designed according to the MSCI ESG rating methodology, requires systems to classify news articles into key ESG issues, taking into account the target company and its industry. Other shared tasks has evolved over the years, with subsequent iterations focusing on different aspects such as ESG Impact Type Identification and Impact Duration Inference, reflecting the complex and dynamic nature of ESG factors.

Furthermore, the ML-ESG shared tasks are instrumental in advancing research in multilingual ESG analysis, acknowledging the global nature of ESG issues and the importance of diverse linguistic representation in ESG analysis. These tasks provide valuable insights into the methodologies and performances of various AI models in handling ESG-related data across different languages.

In conclusion, the integration of ESG factors into AI models through tasks like ML-ESG represents a significant step towards leveraging technology for social good. It aligns with the broader goals of sustainable and responsible investing, emphasizing the importance of considering environmental and social impacts in financial decision-making processes. This approach not only enhances the transparency and accountability of corporations but also promotes long-term value-driven investments.

doi:10.6342/NTU202501476

Chapter 3

Datasets

In this thesis, we adopt two types of datasets: various domain-specific datasets and our proposed dataset, DynamicESG, for different purposes.

Firstly, to investigate the expert-level annotation abilities of LLMs, we conduct comprehensive experiments using domain-specific datasets. These datasets, as introduced in Section 3.1, cover three popular domains: finance, biomedicine, and law. Each domain includes two datasets that are annotated by domain experts to ensure the reliability and accuracy of ground-truth labels, resulting in total of six specialized datasets. All datasets are in English and involve single-choice natural language understanding tasks.

On the other hand, we utilize DynamicESG (Tseng et al., 2023), as proposed in Section 3.2, to prevent data contamination issues (Deng et al., 2023; Sainz et al., 2023). Since it was published after October 2023, DynamicESG is not included in the training corpus of LLMs, thus providing a more representative performance evaluation. Furthermore, the ESG datasets encompass five languages, which are valuable for the research community and are also appropriate for our research to further explore the multilinguality.

doi:10.6342/NTU202501476

3.1 Existing Domain-Specific Datasets

3.1.1 Finance

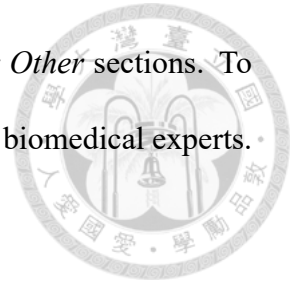
We adopt the REFinD (Kaur et al., 2023) and FOMC datasets (Shah et al., 2023) for financial domain. REFinD is the largest relation extraction dataset over financial documents, comprising 8 entity pairs and 22 relations. The labels have been reviewed by financial experts, ensuring their reliability and accuracy. In this task, annotators are tasked to extract relations between finance-specific entity pairs, such as "[person] *is an employee of* [organization]". FOMC is constructed for identifying sentiments about the future monetary policy stances, annotated by experts with a correlated financial knowledge. The labels of this annotation task are: *Dovish*, *Hawkish*, and *Neutral*, where a Dovish sentence indicates easing and a Hawkish sentence indicates tightening.

3.1.2 Biomedicine

For the biomedical domain, we utilize AP-Relation dataset (Gao et al., 2022) and COVID-19 Research Aspect Dataset (CODA-19) (Huang et al., 2020). AP-Relation is designed for extracting the relationship between Assessment and Plan Subsections in daily progress notes. The Assessment describes the patient and establishes the main symptoms or problems for their encounter, while the Plan Subsection addresses each differential diagnosis or problem with a daily action or treatment plan. The annotation label schemes for different relations are categorized as *Direct*, *Indirect*, *Neither*, or *Not Relevant*. CODA-19 codes each segment aspect of English abstracts in the COVID-19 Open Research Dataset (Wang et al., 2020). In this task, annotators are tasked to label each

doi:10.6342/NTU202501476

segment as *Background*, *Purpose*, *Method*, *Finding/Contribution*, or *Other* sections. To ensure the quality of the labels, we only adopt instances annotated by biomedical experts.



3.1.3 Law

We adopt Contract Understanding Atticus Dataset (CUAD) (Hendrycks et al., 2021) and Function of Decision Section (FoDS) dataset (Guha et al., 2024) in legal domain. CUAD consists of legal contracts with extensive annotations from legal experts, created with a year-long effort by dozens of law student annotators, lawyers, and machine learning researchers. Each law student annotator undergoes 70-100 hours of training before annotating this dataset. The annotation task is to label 41 types out of legal clauses, classified into 5 answer categories, that are considered important in contract review related to corporate transactions. We manually use "*Yes/No*" answer category to construct our annotation task as the identification of 32 types of clauses. FoDS comprises one-paragraph excerpts from legal decisions, annotated by legal professionals who are included as authors. In this task, annotators are tasked to review a legal decision and identify one out of seven function categories that each section (*i.e.*, excerpt) of the decision serves. We provide dataset statistics in Table 3.1 and annotation guidelines in Appendix 7.2.

Domain	**Dataset**	**Instance Type**	**#Instances**	**#Labels**
Finance	REFinD (Kaur et al., 2023)	Sentence	500	22
	FOMC (Shah et al., 2023)	Sentence	500	3
Biomedicine	AP-Rel (Gao et al., 2022)	Pair	73	4
	CODA-19 (Huang et al., 2020)	Paper Abstract	508	5
Law	CUAD (Hendrycks et al., 2021)	Clause	500	32
	FoDS (Guha et al., 2024)	Excerpt	367	7

Table 3.1: The statistics of existing domain-specific datasets.

doi:10.6342/NTU202501476

3.2 Our Proposed Dataset – DynamicESG

3.2.1 Motivation

Corporate Social Responsibility (CSR) has increasingly become a crucial component of company operations. Nowadays, the impacts towards Environmental, Social, and Governance (ESG) serve as a third dimension, beyond return and risk, considered by investors when making corporate and personal investment decisions. Several frameworks and policies have been proposed to assess companies' ESG-related activities, with endeavors to quantify these considerations into scores. Yet, the scoring process necessitates substantial expert involvement and numerous manual annotation procedures for relevant events. To streamline this process and enhance experts' efficiency, we employ the guidelines utilized by experts in Morgan Stanley Capital International (MSCI), a globally recognized authority in formulating financial indexes, to annotate news articles. We anticipate that this dataset will stimulate the proposal of more sophisticated methods.

A shortcoming of the current ESG rating approach is the annual update frequency of ESG ratings. The financial market is dynamic and fast-paced; an annual update cycle is too delayed for decision-making based on the latest information. To address this, we select news articles as the resource for capturing the most recent events and deducing the impact of these events on a company's ESG rating. Another challenge with ESG rating is the recent fragmentation and lack of standardization across different frameworks, complicating the effective navigation of the ESG landscape. In addition to adhering to MSCI ESG guidelines, we incorporate another standard guideline, the Sustainability Account-

doi:10.6342/NTU202501476

ing Standards Board (SASB) Standards[1], to augment our proposed label scheme. This expanded label set enhances the applicability of our proposed dataset to most ESG rating guidelines and allows extension to other ESG-related analyses. We believe that incorporating various perspectives and best practices into a single scheme will yield a more robust and encompassing ESG-rating process.

3.2.2 Overview

We construct this dataset with a meticulous five-step pipeline to guarantee its utility and value. Initially, we gather news data from the Business Today website, a well-regarded Taiwanese magazine known for its content on finance, business, and investment. Subsequently, we align the SASB Standard's ESG issues guideline with that of MSCI, culminating in forty-four ESG categories forming our definitive guideline. Annotators, guided by these categories, are tasked to ascertain three main aspects of the news articles: the impact type, impact duration, and ESG category. We ensure annotation consistency through biweekly discussion sessions, during which disagreements are resolved and answer sets adjusted as required. Lastly, we calculate the agreement and distribution of annotations across the entire dataset.

Our annotations span three aspects: impact type (risk/opportunity), impact duration, and ESG key issues. The impact type helps infer if a given news item will augment the ESG rating. Impact duration aids in understanding the duration of an event's influence on the ESG rating. Differentiating key issues across industries is crucial as the MSCI guidelines suggest varying weights for each issue. It is crucial to ascertain the issue addressed by the news; for example, the carbon emissions issue has a 15.6% weighting in the oil and

[1] https://sasb.org/standards/materiality-map/

doi:10.6342/NTU202501476

gas drilling industry, contrasting with a mere 4.7% weighting in the specialized finance industry. Based on this annotation scheme, we can capture the influence of sentiment, temporal, and topic dimensions for dynamically understanding the possible change of a company's ESG rating.

To sum up, we propose the DynamicESG dataset, a unique resource designed to dynamically glean ESG ratings from news articles. Our work underscores the importance of timely, multi-faceted analysis in understanding the implications of news narratives on ESG ratings and provides a foundation for future explorations in this domain.

3.2.3 Data Collection

After surveying numerous ESG news websites, we select the ESG page of Business Today[2] as our data source, owing to its reputation for comprehensive, diverse content spanning finance, business, and investment topics. We gather a total of 2,472 news articles covering a broad spectrum of ESG topics from this source, ranging from January 1, 2011, to December 31, 2022. Figure 3.1 shows the number of collected news articles across the years, demonstrating an increasing attraction to ESG topics.

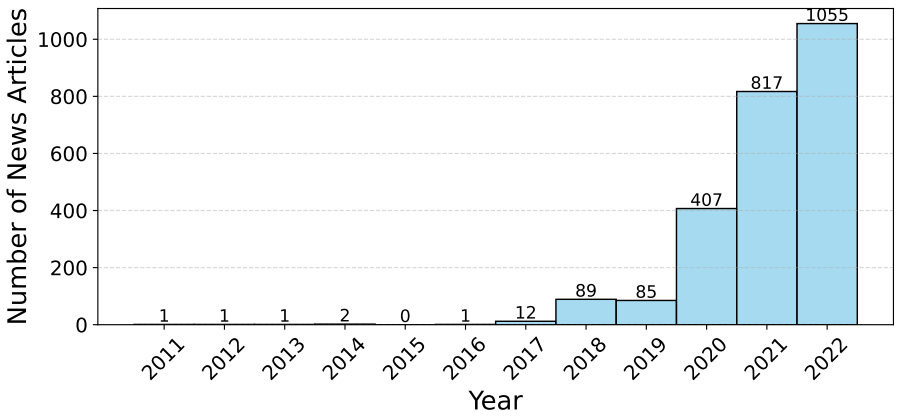


Figure 3.1: The histogram of collected news articles across years.

[2] https://esg.businesstoday.com.tw/catalog/180686/

doi:10.6342/NTU202501476

3.2.4 Task Design

Various methodologies and guidelines for ESG ratings have been introduced globally to assess the implementation outcome of ESG practices by companies and organizations. These frameworks provide a standardized approach for evaluating and comparing the ESG performance of different entities. We select the MSCI and SASB Standards as our primary guidelines among all the existing ESG frameworks.

Regarding ESG ratings, MSCI has outlined 35 key issues across three topics: Environmental, Social, and Governance. Additionally, MSCI's ESG rating methodology categorizes the expected time frame for risk or opportunity to materialize into two distinct periods: short-term ("less than 2 years") and long-term ("more than 5 years"). The methodology also differentiates between the importance of ESG key issues across industries, assigning the highest weight to short-term, high-importance factors and the lowest weight to long-term, low-importance factors. Based on this, MSCI assesses the ESG ratings of companies and organizations annually, using data from corporate or social disclosures, open government information, and other sources. SASB aims to improve the reporting of ESG issues. SASB has identified 26 significant issues spanning five aspects: Environment, Social Capital, Human Capital, Business Model & Innovation, Leadership & Governance. Furthermore, SASB has released the Materiality Map that details the relative importance of each issue across 77 industries.

Though SASB does not assess ESG ratings, Taiwan's Financial Supervisory Commission requires listed companies to report in alignment with SASB's ESG issues. Upon conducting a comparative analysis of two guidelines, we discover that MSCI places greater emphasis on environmental aspects, whereas SASB encompasses elements related to busi-

doi:10.6342/NTU202501476

Topic	Theme	# of Key Issues
Environmental (E)	Climate Change	4
	Natural Capital	3
	Pollution & Waste	3
	Environmental Opportunity	3
Social (S)	Human Capital	5
	Product Liability	9
	Stakeholder Opposition	2
	Social Opportunity	4
Governance (G)	Corporate Governance	8
	Corporate Behavior	3

Table 3.2: Labels in ESG category identification task.

ness models and innovation that MSCI overlooks. Therefore, for the purpose of generalization of ESG issues, we merge SASB's 26 important issues with MSCI's 35 key issues to formulate our final set of 44 ESG category guidelines for the ESG category task. Like MSCI, our guidelines are structured into three topics and ten themes but encompass 44 key ESG issues instead of 35. The statistics of our guidance are provided in Table 3.2, and the detailed illustration of each ESG key issue is in Appendix 7.2.

Guided by this methodology, we crafted three tasks to help classify news articles:

- **Impact Type Identification**: This single-choice question aims to determine the type of impact a news article might have on the company. The possible labels are "*Opportunity*", "*Risk*", and "*Cannot Distinguish*".
- **Impact Duration Inference**: This single-choice question seeks to determine the duration of the impact a news article might have on the company. Based on the distinction between short-term and long-term defined above, we present three labels: "*Less than 2 years*", "*2 to 5 years*", and "*More than 5 years*".
- **ESG Category Identification**: This multiple-choice question is designed to identify the ESG categories related to a news article. The labels include the 44 ESG key

doi:10.6342/NTU202501476

issues, and an additional "*None of the Above*" choice.

3.2.5 Dataset Agreement & Statistics

We employ three measures to evaluate the agreement among annotators, namely Cohen's Kappa (Cohen, 1960), Fleiss Kappa (Fleiss, 1971), and Krippendorff's α (Krippendorff, 2011). Table 3.3 presents the agreement for all tasks. As per Landis and Koch (1977), a Fleiss Kappa value exceeding 0.20 indicates fair agreement among annotators. During our biweekly meetings, we make discussions on the instances that get different labels from annotators to clarify the rationale. After excluding articles that do not align with our objective, we obtain 2,220 instances from the 2,472 collected news articles. Table 3.4 shows the distribution of different labels in the proposed dataset.

In the released DynamicESG, we furnish the news headlines accompanied by a URL, and the annotations corresponding to the proposed three tasks. Due to copyright concerns, users are required to gather the news content themselves. To facilitate this, we provide the web crawler code that enables the reconstruction of the full dataset along with the annotations[3]. The annotations are released under the CC BY-SA 4.0 license. We conduct experiments to ensure the DynamicESG usability in Appendix 7.2.

[3]DynamicESG: https://github.com/ymntseng/DynamicESG

doi:10.6342/NTU202501476

Task		Cohen	Fleiss	Krippendorff
Impact Type		0.49	0.49	0.54
Impact Duration		0.21	0.21	0.29
ESG Category	Topic (3)	0.52	0.49	0.55
	Theme (10)	0.35	0.34	0.37
	Key Issues (44)	0.28	0.28	0.28

Table 3.3: The agreement among annotators of the DynamicESG. Fleiss Kappa value exceeding 0.20 indicates fair agreement among annotators (Landis and Koch, 1977).

	Impact Type			Impact Duration			ESG Category		
	Opportunity	Risk	X	2 years <	2 to 5 years	> 5 years	E	S	G
Train	70.25%	7.60%	3.01%	19.95%	14.22%	46.56%	25.92%	33.45%	27.12%
Dev	7.86%	0.79%	0.39%	2.29%	1.61%	5.28%	2.12%	2.36%	1.84%
Test	8.78%	0.92%	0.39%	2.52%	1.83%	5.73%	2.72%	2.24%	2.20%

Table 3.4: DynamicESG Dataset Statistics.

doi:10.6342/NTU202501476

Chapter 4

Single LLMs as Expert Annotators

In this section, we explore the capabilities of single state-of-the-art (SOTA) models as expert-level annotators. We use six datasets across three domains: finance, biomedicine, and law, as introduced in 3.1. The methodologies range from vanilla prompts to prompts that elicit reasoning abilities. To investigate whether these SOTA models can perform as annotators *out-of-the-box*, we minimize efforts in prompt engineering. Instead, we develop uniform prompt templates that are easily generalizable across domains and datasets.

4.1 Methodology

4.1.1 Vanilla

The vanilla method refers to standard direct-answer prompting, where instructional input consists of the annotation task, guideline, and the sample to be annotated are given to the LLMs. LLMs are tasked to conduct annotation as a domain expert of relevant fields. The vanilla prompt also serves as the base of other sophisticated approaches (described

doi:10.6342/NTU202501476

below). We provide the template in Figure 4.1.

4.1.2 CoT

Prompting with chain-of-thought (CoT) improves LLMs' complex reasoning ability significantly (Wei et al., 2022). Specifically, we employ zero-shot CoT (Kojima et al., 2022), where a trigger phrase "*Let's think step by step*" augments the prompt to elicit reasoning chain from LLMs and leads to a more accurate answer. We provide the CoT prompt template based on the vanilla one in Figure 4.2.

4.1.3 Self-Consistency

Self-consistency (Wang et al., 2022) further improves upon CoT via a sample-and-marginalize decoding procedure, which selects the most consistent answer rather than the greedily decoded one. That is to say, the self-consistency prompt is the same with CoT prompt as illustrated in Figure 4.2. Concretely, we sample five diverse reasoning paths with temperature 0.7, and take the majority vote to determine the final answer.

4.1.4 Self-Refine

Self-refine (Madaan et al., 2024) method includes three steps: generate, review, and refine. An LLM first generates an initial answer with reasoning (*i.e.*, draft). Then, the model reviews its draft and provides feedback. Lastly, the LLM refines the draft by incorporating its feedback, and outputs an improved answer. Note that the same LLM is used in all steps. We provide the three-step self-refine prompt template from Figure 4.3 to Figure 4.5.

doi:10.6342/NTU202501476

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the label in the format of: 'The label is ...'. Do not include any reasoning or explanation.

Figure 4.1: The vanilla prompt template. The words "*domain*", "*guideline*", and "*instance*", enclosed by placeholders, can be easily replaced by different annotation tasks.

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: '**Let's think step by step.** ... The label is ...'.

Figure 4.2: The CoT prompt template. Texts in red, "*Let's think step by step*", indicate the trigger phrase of zero-shot CoT (Kojima et al., 2022).

doi:10.6342/NTU202501476

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

Figure 4.3: The self-refine prompt template (1/3).

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

{[model response from step 1.]}

Review your previous reasoning and annotation and find potential problems. For example, whether the annotation guideline is violated, whether the reasoning is not conclusive.

Figure 4.4: The self-refine prompt template (2/3). Texts in blue are the placeholder of the model draft from step 1 (*i.e.*, the model response of the Figure 4.3 prompt).

doi:10.6342/NTU202501476

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

{**[model response from step 1.]**}

Review your previous reasoning and annotation and find potential problems. For example, whether the annotation guideline is violated, whether the reasoning is not conclusive.

Review:
{**[model response from step 2.]**}

Based on the problems you found in the above review, improve your annotation quality and reasoning and output in the format of: 'Let's think step by step ..... The label is ...'.

Figure 4.5: The self-refine prompt template (3/3). Texts in green are the placeholder of the model feedback from step 2 (*i.e.*, the model response of the Figure 4.4 prompt).

doi:10.6342/NTU202501476

4.2 Experimental Results

We report our main results in Table 4.1. We compare four SOTA LLMs, including GPT-3.5-Turbo (OpenAI, 2023), GPT-4o (OpenAI, 2024), Gemini-1.5-Pro (Reid et al., 2024), and Claude-3-Opus (Anthropic, 2024), and report their annotation accuracy. We use labels annotated by human experts from the corresponding dataset as ground-truth answers.

As observed, under the vanilla method (upper block), GPT-4o records the best overall performance. Claude-3-Opus and Gemini-1.5-Pro achieve similar scores, while GPT-3.5-Turbo performs notably worse. However, all LLMs show substantial rooms for improvements, with an average of 32.2% $\sim$ 43.3% behind human expert annotations. The best single score (GPT-4o on CUAD dataset) still lacks around 20%. The results suggest that naive standard prompting is *not* feasible to obtain satisfactory annotation quality from LLMs in tasks involving domain expertise. Considering that these specialized domains are often relevant to high-risk sectors (*e.g.*, medial application), it is crucial to ensure the annotated data has a higher precision and accuracy.

To further compare the degree of expert-level performance reached by SOTA LLMs, we present a comparison in the bar plot in Figure 4.6. For MMLU benchmark (Hendrycks et al., 2020), we report models scores from the HELM (Liang et al., 2023a) website divided by human-expert score (89.8) from Hendrycks et al. (2020). For annotation tasks, we calculate the average annotation accuracy of GPT-4o, Claude-3-Opus, and Gemini-1.5-Pro with vanilla method across three domains. As illustrated, although these models perform at near human expert-level on MMLU, they still struggle to apply intrinsic domain knowledge in annotation tasks, or even lack the domain knowledge.

doi:10.6342/NTU202501476

To probe the capabilities of LLMs more further, we experiment the most performant model of the vanilla method, GPT-4o, with three methods: CoT, self-consistency (SC), and self-refine (SR), proposed to improve LLMs factual knowledge and reasoning capabilities. The results are present in Table 4.1 lower block. As observed, in general, all methods exhibit improved results, with an average of 1% $\sim$ 2% accuracy gain. However, comparing with the huge performance boosts of how these methods typically benefit general domain datasets, their efficacy on expert-level annotation tasks is relatively low. This might imply that the models inherently lack necessary knowledge and reasoning capability to perform as expert annotators.

Model / Method	Finance		Biomedicine		Law		Avg.
	REFinD	FOMC	AP-Rel	CODA-19	CUAD	FoDS	
GPT-3.5-Turbo	47.4	60.4	58.9	64.4	71.8	37.1	56.7
GPT-4o	**67.2**	**67.6**	65.8	**79.3**	**82.2**	44.4	**67.8**
Gemini-1.5-Pro	64.6	**67.6**	54.8	73.2	80.6	42.8	63.9
Claude-3-Opus	61.2	63.6	**71.2**	65.6	80.8	**46.9**	64.9
GPT-4o	67.2	67.6	65.8	79.3	82.2	44.4	67.8
CoT	71.0 (↑3.8)	*68.2 (↑0.6)	*68.5 (↑2.7)	*81.1 (↑1.8)	79.8 (↓2.4)	43.9 (↓0.5)	68.7
Self-Consistency	*72.4 (↑5.2)	*70.4 (↑2.8)	*68.5 (↑2.7)	78.9 (↓0.4)	*†82.4 (↑0.2)	*†45.0 (↑0.6)	**69.6**
Self-Refine	70.0 (↑2.8)	*69.2 (↑1.6)	*69.9 (↑4.1)	*81.5 (↑2.2)	78.0 (↓4.2)	*45.5 (↑1.1)	69.0

Table 4.1: Performance of SOTA LLMs as annotators (accuracy) and a comparison of GPT-4o with different advanced techniques for expert-level annotation tasks. An asterisk (*) indicates that the method is statistically significant with p-value < 0.05 than the vanilla method. A dagger (†) indicates that the self-consistency method is statistically significant with p-value < 0.05 than the CoT method.

doi:10.6342/NTU202501476

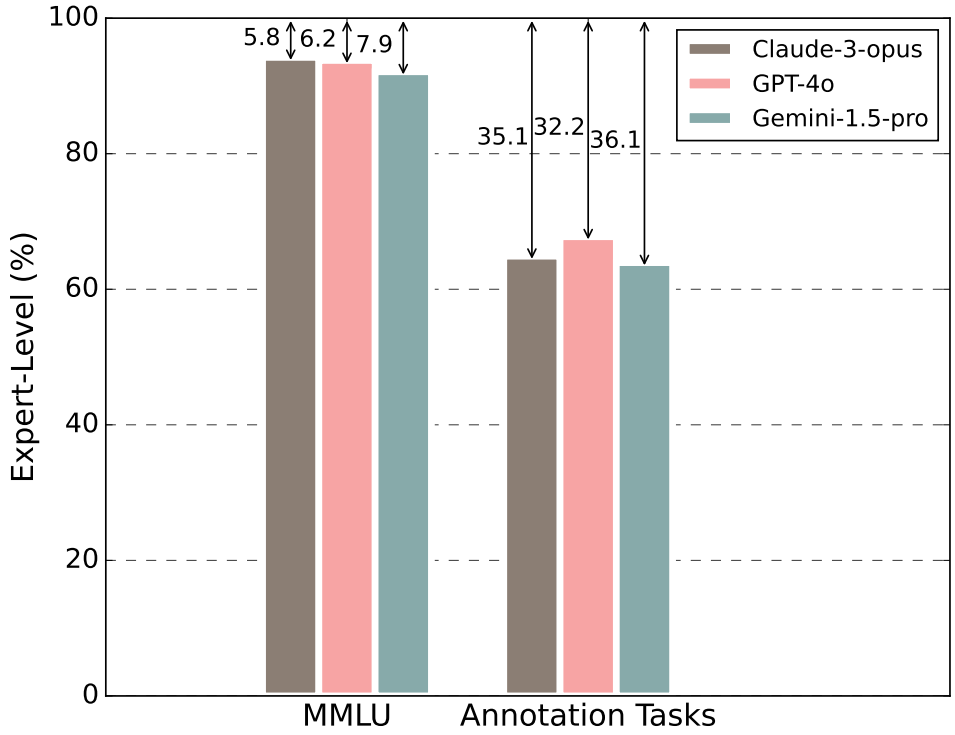


Figure 4.6: The degree of expert-level performance reached by SOTA LLMs.

doi:10.6342/NTU202501476

Additionally, we provide a comparison of performance in both vanilla and CoT settings across different domains and SOTA LLMs in Table 4.2. For better comparison, we also present the difference between the two settings in the table.

In the finance domain, three LLMs show a slight improvement when applying CoT method, except for Gemini-1.5-pro on the REFinD dataset. In the biomedicine domain, we observe a significant difference between the two datasets for Gemini-1.5-Pro and Claude-3-Opus, where the CoT method enhances performance on CODA-19 but decreases performance on AP-Rel. As for the law domain, all three LLMs exhibit lower performance when applying the CoT method, suggesting that LLMs might not be capable of reasoning effectively on law datasets.

From an average perspective, while applying CoT method, only GPT-4o boosts performance, whereas both Gemini-1.5-Pro and Claude-3-Opus show a decline in performance. However, the effectiveness of the CoT method varies across different models and datasets, indicating that its utility might be context-dependent. The table highlights the strengths and weaknesses of each model and method across varied specialized domains, suggesting that the choice of model and method could be crucial depending on the specific requirements of the task at hand.

Model / Method	Finance		Biomedicine		Law		Avg.
	REFinD	FOMC	AP-Rel	CODA-19	CUAD	FoDS	
GPT-4o	67.2	67.6	65.8	79.3	**82.2**	44.4	67.8
CoT	**71.0** (↑3.8)	68.2 (↑0.6)	68.5 (↑2.7)	81.1 (↑1.8)	79.8 (↓2.4)	43.9 (↓0.5)	**68.7**
Gemini-1.5-Pro	64.6	67.6	54.8	73.2	80.6	42.8	63.9
CoT	63.4 (↓1.2)	**68.6** (↑1.0)	46.6 (↓8.2)	**82.7** (↑9.5)	73.2 (↓7.4)	40.3 (↓2.5)	62.5
Claude-3-Opus	61.2	63.6	**71.2**	65.6	80.8	**46.9**	64.9
CoT	66.4 (↑5.2)	65.2 (↑1.6)	60.3 (↓10.9)	71.9 (↑6.3)	77.8 (↓3.0)	42.2 (↓4.7)	64.0

Table 4.2: Performance of SOTA LLMs in both vanilla and CoT settings.

doi:10.6342/NTU202501476

To prevent data contamination issues in the aforementioned six domain-specific datasets, we conduct the same experiments on the proposed English version of the ESG dataset and report the results in Table 4.3. As observed, GPT-4o performs moderately, achieving a similar average score of approximately 64% compared to the performance in Table 4.2. Additionally, GPT-4o only boosts performance by 0.3% with the CoT method and the self-consistency method, which differs from the results in the financial domain shown in Table 4.1.

It is noteworthy that the label scheme of the impact type consists of only two categories: Opportunity and Risk. For GPT-4o, this task, which lies on the two side of the spectrum and is more similar to general tasks, might be easier to distinguish. Consequently, GPT-4o achieves a high accuracy score of 87.2% with the CoT method.

On the other hand, the other two tasks are more difficult for GPT-4o due to the time factors in the impact duration tasks and the large number of label schemes in the key issue identification task. Obviously, the performance on the ESG datasets still lags behind human expert annotations, indicating that there is still ample room for improving the abilities of single LLMs as expert annotators.

	Impact Type	**Impact Duration**	**Key Issue**	**Avg.**
GPT-4o	85.3	**63.2**	44.3	64.3
CoT	*__87.2__ (↑1.9)	61.8 (↓1.4)	*45.0 (↑0.7)	**64.6**
Self-Consistency	*86.7 (↑1.4)	61.8 (↓1.4)	*__45.3__ (↑1.0)	**64.6**

Table 4.3: Performance of GPT-4o on the three tasks in the English ESG datasets. An asterisk (*) indicates that the method is statistically significant with p-value < 0.05 than the vanilla method.

doi:10.6342/NTU202501476

Chapter 5

Multi-Agent LLMs as Expert Annotators

When it comes to data annotation, a common scenario during the annotation process is the disagreement among multiple annotators. One typical way of resolving such discrepancy is by discussing with others to reach a consensus, thereby constructing a higher-quality annotated dataset. Motivated by this, we investigate the capabilities of multi-agent LLMs as expert annotators, where they simulate this typical resolution process to test whether LLMs could also construct a higher-quality and more accurate dataset in this manner. Our multi-agent annotation framework consists of three performant LLMs: GPT-4o, Gemini-1.5-Pro, and Claude-3-Opus.

5.1 Methodology

5.1.1 Majority Vote

Majority vote (MV) represents a minimal form of discussion, reducing the process to simply selecting the majority output as the final annotation. We apply two settings for

doi:10.6342/NTU202501476

MV: vanilla and CoT, as introduced in 4.1.

5.1.2 Peer-Discussion

Peer-Discussion consists of three steps: (*1*) Generate initial annotation, (*2*) Check annotations, (*3*) Discuss and re-annotate, as illustrated in Figure 5.1. Initially, each agent generates their own annotation through CoT prompting given the same annotation task, guideline, and instance. Next, we check if consensus has been reached (*i.e.*, all annotations are the same labels). If consensus is achieved, the instance is successfully annotated and the annotation process is complete. Otherwise, we incorporate all agents' reasoning and labels to generate a "*Discussion History*". Subsequently, agents are required to re-annotate the instance, given the same input and the discussion history. Thus, we iteratively repeat the same check-consensus-discuss-re-annotate procedure until achieving consensus or reach the maximum discussion round (hyperparameter). In our experimental settings, we set the maximum discussion round to 2. We provide the prompt templates in Figure 5.2 and Figure 5.3.

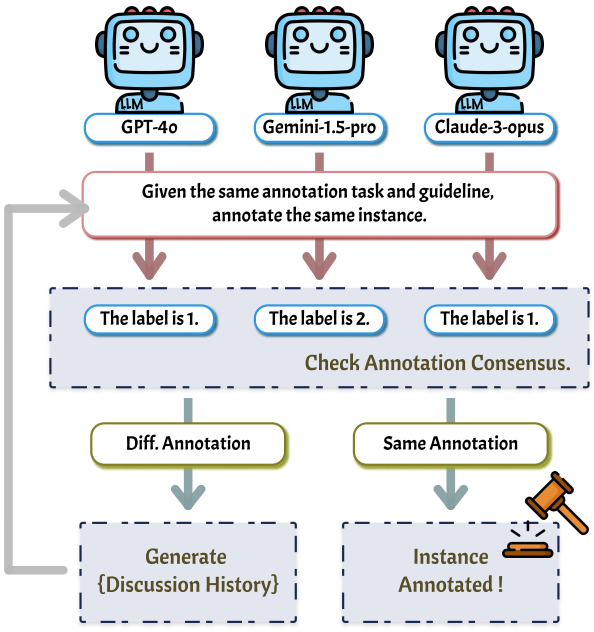


Figure 5.1: An illustration of multi-agent peer-discussion method.

doi:10.6342/NTU202501476

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

Figure 5.2: The multi-agent peer-discussion prompt template (1/2).

You are a [domain] expert tasked to annotate a [domain] dataset. Please follow the annotation guideline below.

Annotation Guideline:
{[guideline]}

{[instance_type]}:
{[instance]}

Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

Discussion History:
{[discussion_history]}

You need to consider the above discussion history carefully. You can maintain your point of view and annotation if others' reasons are not concrete or cannot convince you.
Please strictly follow the guideline and output the reasoning and the label in the format of: 'Let's think step by step. ... The label is ...'.

Figure 5.3: The multi-agent peer-discussion prompt template (2/2). Texts in purple are the placeholder of the *discuss history*, which generated when agents do not reach a consensus.

doi:10.6342/NTU202501476

5.2 Experimental Results

We present the results of multi-agent frameworks across three domains in Table 5.1. As observed, the performance of the three multi-agent settings is competitive, with the majority vote in the vanilla setting showing the best performance among the three. Additionally, Figure 5.4 compares these results with the average performance of single agents (*i.e.*, the single-LLM setting discussed in Section 4).

As shown, multi-agent frameworks slightly outperform the average of single agent settings. However, the most performant singe LLMs, primarily GPT-4o, whether in vanilla or in CoT settings, still exhibit superior results in most cases. Surprisingly, multi-agent with vanilla consistently outperforms the average of single-vanilla LLMs, but underperforms compared to the best single-vanilla LLM. Similarly, multi-agent with CoT also consistently outperforms the average of single-CoT LLM, but underperforms compared to the best single-CoT LLM.

For each domain, different multi-agent settings achieve the highest performance. In the finance domain, peer discussion yields the best results, while in the biomedicine domain, the majority vote with CoT is most effective. In the law domain, the majority vote with vanilla performs the best. These results demonstrate that the effectiveness of various methods varies across different models and datasets, indicating that their utility might be context-dependent. However, it is noteworthy that these results still lag significantly behind human annotators.

On the other hand, multi-agent with vanilla-MV appears to be a better, cheaper, and more stable methods in the multi-agent framework. Though multi-agent with vanilla-MV

doi:10.6342/NTU202501476

is still inferior to the best single-vanilla and single-CoT LLM, it may be a more suitable approach when we are unable to infer which model to adopt in advance.

Model / Method	Finance		Biomedicine		Law		Avg.
	REFinD	FOMC	AP-Rel	CODA-19	CUAD	FoDS	
MV-Vanilla	67.0	**68.2**	**67.1**	73.8	**82.6**	**46.3**	**67.5**
MV-CoT	68.4	67.6	64.4	79.9	81.4	42.8	67.4
Discussion-CoT	**72.0**	66.4	57.5	**81.3**	**82.6**	45.0	**67.5**

Table 5.1: Performance of multi-agent LLMs as annotators.

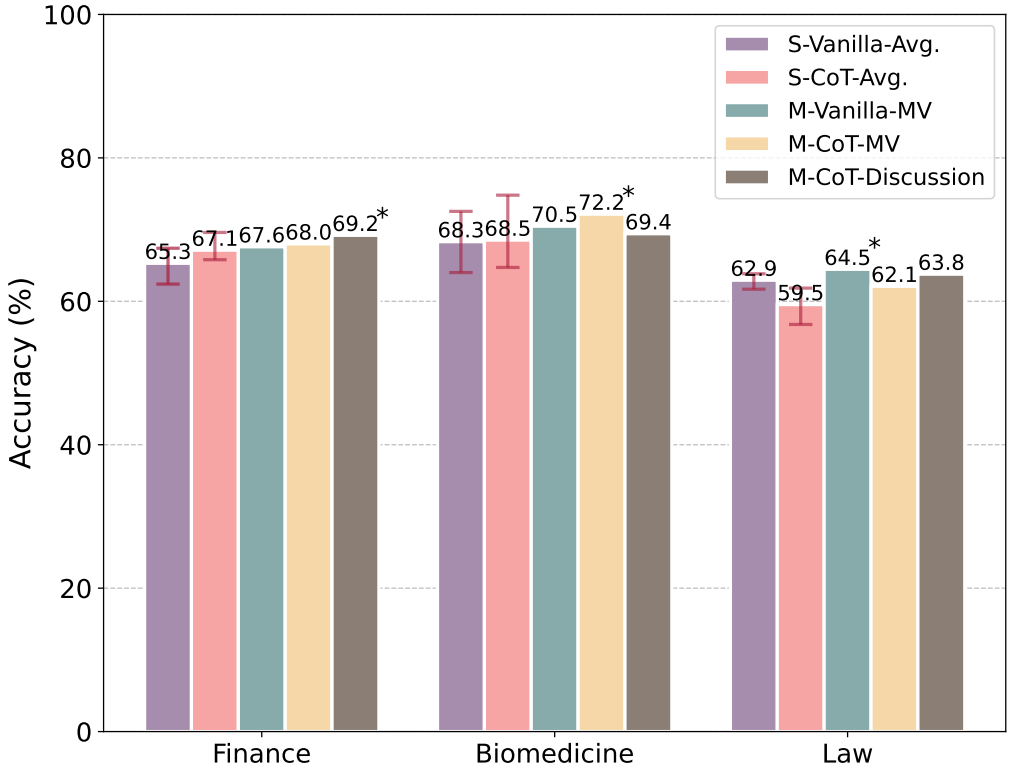


Figure 5.4: The performance comparison of different single LLM settings (S) and multi-agent frameworks (M) across three domains. For the two single agent settings (*i.e.*, *S-Vanilla-Avg.* and *S-CoT-Avg.*), the numbers on the figure represent the average performance of three single LLMs: GPT-4o, Gemini-1.5-Pro, and Claude-3-Opus, while the red bars indicate the range of performance. An asterisk (*) indicates that the method is statistically significant with p-value < 0.05.

doi:10.6342/NTU202501476

To investigate the discussion process in the multi-agent peer-discussion setting, we present the marginal performance of each LLM over two rounds in Figure 5.5. As observed, the peer-discussion process benefits most datasets, leading to improved annotation performance (i.e., increasing *Majority Vote* lines in the figure). For the AP-Rel dataset, however, peer discussion alternately hurt performance, which might be due to the smaller number of instances.

Especially, Gemini-1.5-Pro exhibits improvement in the peer-discussion process across different domains and datasets, except for the FOMC dataset. Initially, Gemini-1.5-Pro achieves the best performance among the three LLMs on this dataset, however, the peer-discussion process degrades its performance. This suggests that Gemini-1.5-Pro might not be able to maintain its correct annotations and is easier to follow other's annotations.

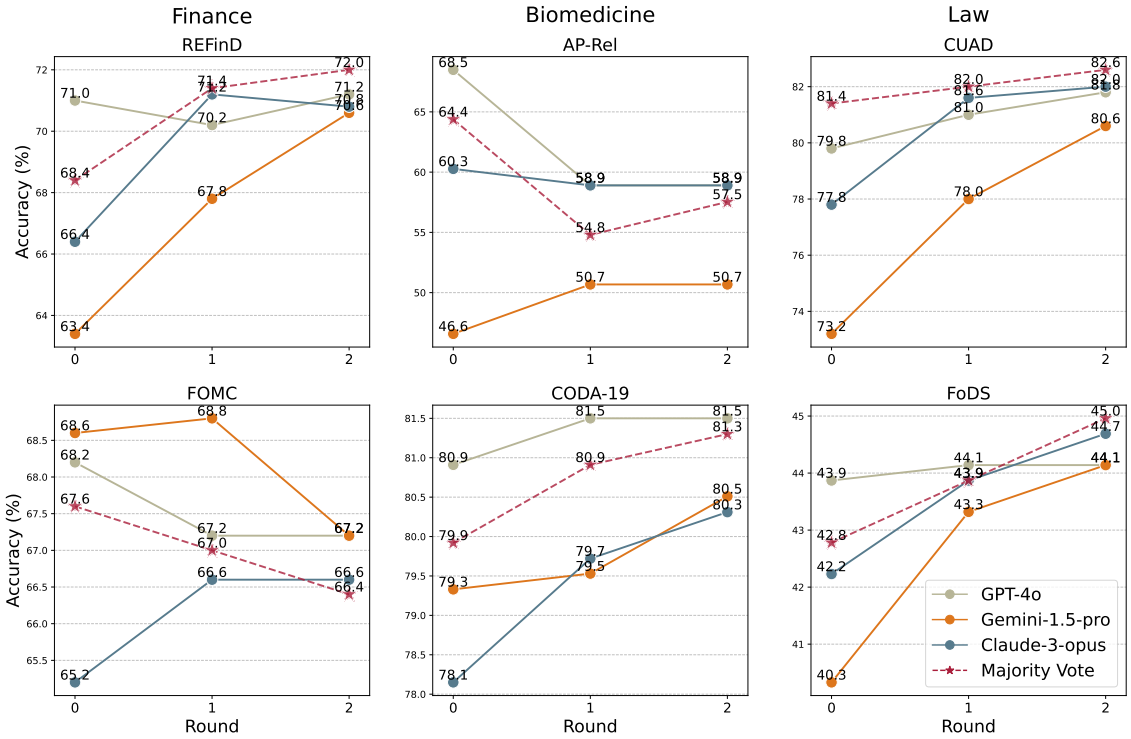


Figure 5.5: Marginal performance of each LLM during multi-agent peer-discussion process. The performance in Round 0 indicates the LLMs' initial annotation performance, while the performance in Round 1 and Round 2 indicates the LLMs' annotation performance after one or two rounds discussion, respectively.

doi:10.6342/NTU202501476

Chapter 6

Discussion

6.1 Analysis of Multilinguality

To further explore the multilinguality of LLMs as Expert Annotators, we conduct experiments on ML-ESG datasets.

For impact type and impact duration tasks, we compare the English, French, Japanese, Korean, and Chinese versions. Though the tasks descriptions are the same, the Chinese and Korean label scheme is slightly different between various versions as described in Table 6.1. We present the results of GPT-4o vanilla and CoT settings in Table 6.2.

As shown, the results indicate that GPT-4o generally offers better performance in English compared to French, given the same label schemes. Surprisingly, for the Japanese version on the impact type task, GPT-4o outperforms all other languages. As for the Korean and Chinese versions, GPT-4o underperforms compared to the other three languages. Note that the results cannot be directly compared to each other due to the different label schemes.

doi:10.6342/NTU202501476

Task	Language	Label Scheme
Impact Type	EN	(2) Opportunity, Risk
	FR	
	JP	
	KR	(3) Opportunity, Risk, Cannot Distinguish
	CN	(5) Opportunity, Risk, Cannot Distinguish, Not Related to Company but Related to ESG Topic, Not Related to ESG Topic
Impact Duration	EN	(3) < 2 Years, 2 to 5 Years, > 5 Years
	FR	
	JP	
	KR	
	CN	(5) < 2 Years, 2 to 5 Years, > 5 Years, Not Related to Company but Related to ESG Topic, Not Related to ESG Topic

Table 6.1: Label schemes of two tasks in ESG datasets across various languages.

Task	Model / Method	EN	FR	JP	KR	CN
Impact Type	*GPT-4o*	85.3	78.0	92.0	67.5*	55.1*
	CoT	87.2	80.5	96.5	76.0*	51.9*
Impact Duration	*GPT-4o*	63.2	65.8	58.5	36.0	24.4*
	CoT	61.8	58.2	60.6	35.5	22.0*

Table 6.2: Performance of GPT-4o on various language versions of ESG tasks. An asterisk (*) indicates that the language has slightly different label schemes compared to others.

doi:10.6342/NTU202501476

While the annotation guidelines, label schemes, and the number of instances for the English and French ESG key issue identification tasks are the same, we compare the performance on both datasets using various model settings with the CoT prompting. Figure 6.3 demonstrates the results. As observed, the performance of the French dataset underperforms in all settings, whether using various single LLMs or multi-agent configurations. Notably, Gemini-1.5-Pro exhibits the largest gap between the English and French datasets, with an absolute difference of 10% and a relative difference of 26.8%. The results suggest that LLMs might not be capable of performing domain annotation tasks in languages other than English, such as French.

	EN	FR	Absolute Diff. (%)	Relative Diff. (%)
GPT-4o	**45.0**	38.7	↓ 6.3	↓ 14.1
Gemini-1.5-Pro	37.3	27.3	↓ 10.0	↓ 26.8
Claude-3-Opus	43.7	**40.7**	↓ 3.0	↓ 6.9
Majority Vote	44.0	38.0	↓ 6.0	↓ 13.6
Peer-Discussion	44.0	39.3	↓ 4.7	↓ 10.6

Table 6.3: CoT Performance of English and French ESG key issue identification tasks (accuracy) with the absolute and relative differences.

doi:10.6342/NTU202501476

6.2 Cost-Effectiveness Analysis

We aggregate our empirical results and compile a cost-effectiveness illustration in Figure 6.1. The cost denotes per-instance annotation cost. In sum, GPT-4o with vanilla or CoT method presents as the best cost-effective options. GPT-4o with SC achieves the best overall performance at the expense of tripling the cost. An intermediate option would be multi-agent vanilla-MV, which demonstrates competitive performance and could be a more robust option when access to different LLMs are available. Despite LLMs do not present as a direct alternative for annotation tasks requiring domain expertise, their collective performance of over 50% and profoundly lower cost present a promising human-LLM hybrid annotation schema in the future.

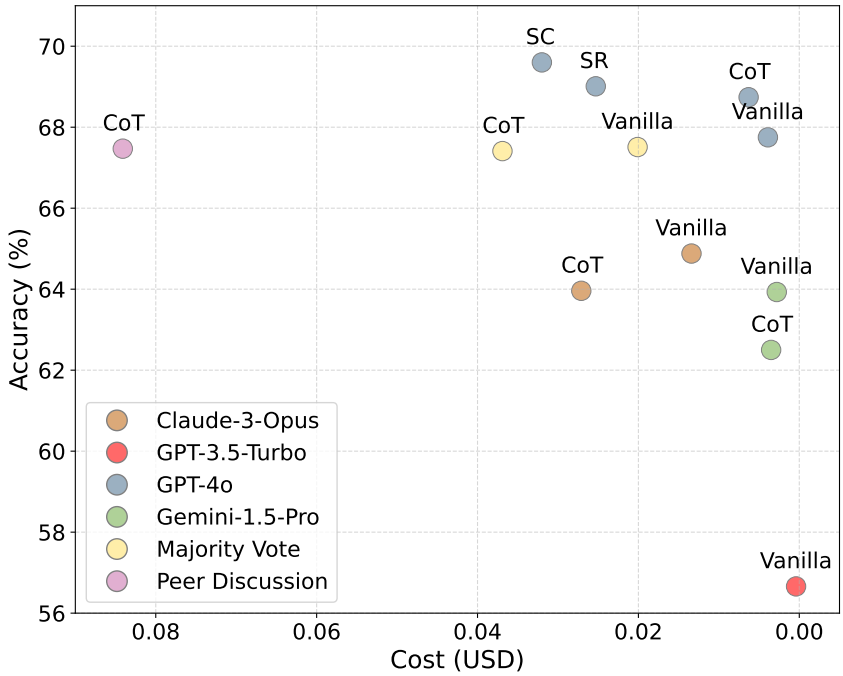


Figure 6.1: An illustration of the cost-effectiveness relationship of various setups of LLMs as expert annotators. The x-axis represents the cost per instance in USD, and the y-axis represents the accuracy in percentage. Note that the x-axis is counterintuitive compared to the usual orientation, with higher costs on the left and lower costs on the right. The upper right corner of the figure indicates better performance, combining lower cost and higher accuracy.

doi:10.6342/NTU202501476

Chapter 7

Conclusion

7.1 Conclusion

In this thesis, we present a comprehensive pilot study on the feasibility of leveraging SOTA LLMs as expert-level annotators. Our investigation spans various methodologies and experimental setups, including single LLMs and multi-agent frameworks, to evaluate their performance across different domains.

Our findings indicate that while LLMs such as GPT-4o, Claude-3-Opus, and Gemini-1.5-Pro exhibit near human expert-level performance on general benchmarks like MMLU, they face significant challenges when applied to domain-specific annotation tasks. The results show that these models struggle to apply intrinsic domain knowledge effectively, which is crucial for expert-level annotations. The study also explores advanced techniques like Chain-of-Thought, self-consistency, and self-refine to enhance the factual knowledge and reasoning capabilities of LLMs. Although these methods yield slight improvements in annotation accuracy, the gains are relatively modest compared to their impact on general domain datasets. This suggests that the inherent knowledge and reasoning capabilities

doi:10.6342/NTU202501476

of current LLMs may not be sufficient for expert-level annotation tasks without further optimization.

Moreover, our exploration of multi-agent frameworks reveals that the majority vote with vanilla setting can offer competitive performance and cost-effectiveness. However, more complex peer-discussion methods do not consistently outperform single LLM setups, indicating that the added complexity does not necessarily translate to better performance in the context of expert-level annotations.

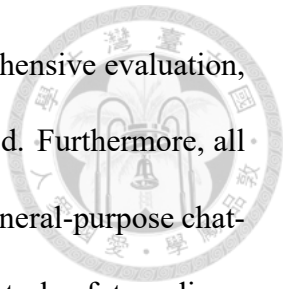
Despite these limitations, the study underscores the potential of LLMs in contributing to human-LLM hybrid annotation schemas. The collective performance of LLMs, coupled with their significantly lower cost, presents a promising avenue for future research and practical applications in data annotation tasks. In conclusion, while SOTA LLMs show promise as cost-effective annotators, their application in expert-level tasks requires further refinement and optimization. Future work should focus on enhancing the domain-specific knowledge and reasoning capabilities of LLMs, as well as exploring more effective multi-agent collaboration strategies to fully realize their potential in expert-level data annotation.

7.2 Limitations & Future Directions

As we aim to provide direct insight and observation on whether top-performing LLMs can perform as expert annotators *out-of-the-box*, we minimize efforts in prompt engineering. Some works have demonstrated that, for specific scenarios, one can achieve sizable improvement through carefully-crafted prompts. Consequently, our results may further benefit from a more exhaustive prompt optimization.

Another potential limitation is that we primarily focus on natural language under-

doi:10.6342/NTU202501476



standing (NLU) tasks with fixed label space. Towards a more comprehensive evaluation, natural language generation (NLG) tasks could be further incorporated. Furthermore, all of our experimental settings involve zero-shot configurations using general-purpose chatbot LLMs. To unveil more of the capabilities of LLMs in annotation tasks, future directions could explore few-shot settings, domain-specific or fine-tuned LLMs tailored to the annotation tasks, retrieval-augmented generation methods, or a promising human-LLM hybrid annotation schema.

doi:10.6342/NTU202501476

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

doi:10.6342/NTU202501476

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*, pages 46–50, Bali, Indonesia. Association for Computational Linguistics.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anais Lhuissier, Yohei Seki, Hanwool Lee, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual ESG impact duration inference. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024*, pages 219–227, Torino, Italia. ELRA and ICCL.

doi:10.6342/NTU202501476

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023c. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023d. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. 2021. Chatgpt goes to law school. *J. Legal Educ.*, 71:387.

Juhwan Choi, Eunju Lee, Kyohoon Jin, and YoungBin Kim. 2024. GPTs are multilingual annotators for sequence generation tasks. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 17–40, St. Julian's, Malta. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings*

doi:10.6342/NTU202501476

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5484–5493, Marseille, France. European Language Resources Association.

Guido Giese, Zoltán Nagy, and Linda-Eling Lee. 2021. Deconstructing esg ratings performance: Risk and return for e, s, and g by time horizon, sector, and weighting. *The Journal of Portfolio Management*, 47(3):94–111.

doi:10.6342/NTU202501476

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao

doi:10.6342/NTU202501476

Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3054–3063.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

doi:10.6342/NTU202501476

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ook Lee, Hanseon Joo, Hayoung Choi, and Minjong Cheon. 2022. Proposing an integrated approach to analyzing esg data via machine learning and deep learning algorithms. *Sustainability*, 14(14):8745.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023a. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023b. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

doi:10.6342/NTU202501476

OpenAI. 2023. Gpt-3.5 turbo.

OpenAI. 2024. Hello gpt4-o.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task

market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679, Toronto, Canada. Association for Computational Linguistics.

doi:10.6342/NTU202501476

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

doi:10.6342/NTU202501476

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and

doi:10.6342/NTU202501476

Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

doi:10.6342/NTU202501476

Appendix A — Ensuring DynamicESG Availability

Given that some news articles surpass the input length limitation (512 tokens) of conventional language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), we employ Longformer (Beltagy et al., 2020) to overcome this issue. In our experiment, the Longformer is fine-tuned with the learning rate and weight decay set to 1e-5 and 0.03, respectively.

We present the experimental results in terms of precision, recall, and weighted F1-Score (Weighted F1) in Table A.1. In the task of impact type identification, the models train with the proposed DynamicESG dataset attain near-perfect performance in discerning opportunities and risks from an ESG perspective. Even though the proposed dataset is not extensive, the experimental results suggest its effective use in industrial applications, particularly for the impact type identification task, without significant concerns about the model's performance.

In the task of impact duration inference, the Longformer achieves a Weighted F1-score of 0.728. This underscores the importance of our biweekly meetings with annotators to resolve instances of disagreement. After addressing this issue, the model exhibits good performance when trained with the proposed DynamicESG dataset. These exper-

doi:10.6342/NTU202501476

	Impact Type	Impact Duration	ESG Category
Precision	0.877	0.722	0.671
Recall	0.872	0.772	0.450
Weighted F1	0.865	0.728	0.467

Table A.1: The performance of Longformer on three tasks of DynamicESG.

imental results mitigate concerns about agreement during the annotation process, which merely meets the criteria for fair agreement. In this context, we conduct experiments under three-topic settings for the ESG issue identification task. Given that a news article may include discussions on two or even all topics, we experiment within a multi-label task setting. However, the experimental results indicate that the model still finds it challenging to identify ESG issues, even under the simplified three-topic setting.

doi:10.6342/NTU202501476

Appendix B — The Guideline of the DynamicESG Key Issues

doi:10.6342/NTU202501476

3 Topics	10 Themes	Index	44 Key Issues	Illustration
Environ-mental (E)	Climate Change	E01	Carbon Emissions	Companies are evaluated on the carbon intensity of their operations and their efforts to manage climate-related risks and opportunities.
		E02	Product Carbon Footprint	Companies are evaluated on the carbon intensity of their products and their ability to reduce the carbon footprint in their supply chains or in the use of their products and services.
		E03	Financing Environment Impact	Financial institutions are evaluated on the environmental risks of their lending and underwriting activities and their ability to capitalize on opportunities related to green finance.
		E04	Climate Change Vulnerability	Insurance companies are assessed on the physical risk that climate change may pose to insured assets or individuals.
	Natural Capital	E05	Water Stress	Companies are evaluated on the water intensity of their operations, the water stress in their areas of operation and their efforts to manage water-related risks and opportunities.
		E06	Raw Material Sourcing	Companies are evaluated on the environmental impacts of the raw materials used in their products and their efforts around supply chain traceability and certification
		E07	Biodiversity & Land Use	Companies are evaluated on the potential impact of their operations on biodiversity in their areas of operation and their efforts to manage the environmental impact of their operations.
	Pollution & Waste	E08	Toxic Emissions & Waste	Companies are evaluated on the potential environmental contamination and toxic or carcinogenic emissions arising from their operations and the strength of their environmental management systems.
		E09	Electronic Waste	Companies are evaluated on their production of electronic waste, their potential exposure to e-waste regulations and their efforts around product collection and recycling.
		E10	Packaging Material & Waste	Companies are evaluated on their production of or reliance upon packaging materials, their potential exposure to waste management and packaging regulations and their efforts to reduce the environmental impact of packaging materials.

1

Figure B.1: The guideline of ESG key issues in DynamicESG (1/6).

doi:10.6342/NTU202501476

Environ-mental (E)	Env. Opportunity	E11	Opportunities in Renewable Energy	Companies are evaluated on their efforts to develop renewable power generation capacity and/or enable renewable power development through network expansion and "green power" offerings.
		E12	Opportunities in Clean Tech	Companies are evaluated on their clean tech innovation capacity, strategic development initiatives, and revenue generated from clean technologies.
		E13	Opportunities in Green Building	Companies are evaluated based on the resource consumption and carbon intensity of their property assets, their potential exposure to environmental building regulations and their efforts to improve the environmental performance of their real estate assets.
Social (S)	Human Capital	S01	Labor Management	Companies are evaluated on the complexity of their workforce (size, labor intensity, and regions of operation), the relationship between management and labor, the strength of worker protections, and their employee engagement efforts.
		S02	Health & Safety	Companies are evaluated on their management of workplace safety and the workplace safety standards in the industries and regions in which they operate.
		S03	Human Capital Development	Companies are evaluated on their workforce talent requirements and their ability to attract, retain, and develop a highly skilled workforce.
		S04	Supply Chain Labor Standards	Companies are evaluated on the management and transparency of their supply chain and the working standards in the regions in which their suppliers are located.
		S05	Human Rights & Community Relations	The category addresses management of the relationship between businesses and the communities in which they operate, including, but not limited to, management of direct and indirect impacts on core human rights and the treatment of indigenous peoples. More specifically, such management may cover socio-economic community impacts, community engagement, environmental justice, cultivation of local workforces, impact on local businesses, license to operate, and environmental/social impact assessments. The category does not include environmental impacts such as air pollution or waste which, although they may impact the health and safety of members of local communities, are addressed in separate categories.

2

Figure B.2: The guideline of ESG key issues in DynamicESG (2/6).

doi:10.6342/NTU202501476

		S06	Chemical Safety	Companies are evaluated on the possible presence of harmful chemicals in their product portfolio, their potential exposure to strengthening or pending chemical regulations and their efforts to develop less harmful alternatives.
		S07	Privacy & Data Security	Companies are evaluated on the amount of personal data they collect, their exposure to evolving or increasing privacy regulations, their vulnerability to potential data breaches, and their systems for protecting personal data.
		S08	Responsible Investment	Companies are evaluated on their integration of environmental, social and governance considerations in the management of their own assets or the assets they manage on behalf of others.
		S09	Product Safety & Quality	Companies are evaluated on their exposure to possible recalls or product safety concerns, the strength of their supply chain and sourcing systems, their quality management efforts in manufacturing and their responsible marketing practices.
	Product Liability	S10	Consumer Financial Protection	Financial institutions are evaluated on product stewardship and transparency, including efforts to mitigate potential reputational and regulatory risks arising from unethical lending practices or mis-selling financial products to consumers.
Social (S)		S11	Insuring Health & Demographic Risk	Insurance companies are evaluated on their management of emerging social risks and their development of insurance products to address emerging needs that may arise from major public health and demographic trends.
		S12	Selling Practices & Product Labeling	The category addresses social issues that may arise from a failure to manage the transparency, accuracy, and comprehensibility of marketing statements, advertising, and labeling of products and services. It includes, but is not limited to, advertising standards and regulations, ethical and responsible marketing practices, misleading or deceptive labeling, as well as discriminatory or predatory selling and lending practices. This may include deceptive or aggressive selling practices in which incentive structures for employees could encourage the sale of products or services that are not in the best interest of customers or clients.
		S13	Product Design & Lifecycle Management	The category addresses incorporation of environmental, social, and governance (ESG) considerations in characteristics of products and services provided or sold by the company. It includes, but is not limited to, managing the lifecycle impacts of products and services, such as those related to packaging, distribution, use-phase resource intensity, and other environmental and social externalities that may occur during their use-phase or at the end of life. The category captures a company's ability to address customer and societal demand for more sustainable products

Figure B.3: The guideline of ESG key issues in DynamicESG (3/6).

doi:10.6342/NTU202501476

Social (S)	Product Liability			and services as well as to meet evolving environmental and social regulation. It does not address direct environmental or social impacts of the company's operations, nor does it address health and safety risks to consumers from product use, which are covered in other categories.
		S14	Supply Chain Management	The category addresses management of environmental, social, and governance (ESG) risks within a company's supply chain. It addresses issues associated with environmental and social externalities created by suppliers through their operational activities. Such issues include, but are not limited to, environmental responsibility, human rights, labor practices, and ethics and corruption. Management may involve screening, selection, monitoring, and engagement with suppliers on their environmental and social impacts. The category does not address the impacts of external factors – such as climate change and other environmental and social factors – on suppliers' operations and/or on the availability and pricing of key resources, which is covered in a separate category.
	Stakeholder Opposition	S15	Controversial Sourcing	Companies are evaluated on their dependence on and purchasing volume of raw materials procured from conflict areas and their efforts around traceability and certification.
		S16	Community Relations	Companies are evaluated on their management of local community relations, policies on conflict and human rights, and efforts to distribute benefits to local communities.
	Social Opportunity	S17	Access to Communication	Companies are evaluated on their efforts to expand connectivity and access to information in developing countries and historically underserved markets (e.g., rural, elderly).
		S18	Access to Finance	Companies are evaluated on their efforts to expand financial services to historically underserved markets, including small business lending and the development of innovative distribution channels.
		S19	Access to Health Care	Companies are evaluated on their efforts to expand health care products and services to developing countries and underserved markets (e.g., low regional physician concentration), including equitable pricing mechanisms, patents, capacity advancement and product donations.
		S20	Opportunities in Nutrition & Health	Companies are evaluated on the nutritional content of their food products and their efforts to introduce products with an improved nutritional or health profile.

4

Figure B.4: The guideline of ESG key issues in DynamicESG (4/6).

doi:10.6342/NTU202501476

Gover-nance (G)	Corporate Governance	G01	Board	Companies are evaluated on the effectiveness of their board in overseeing management and corporate strategy, protecting investor value, and representing the interests of shareholders. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G02	Pay	Companies are evaluated on the alignment between their pay and other incentive practices and corporate strategy. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G03	Ownership & Control	Companies are evaluated on their equity ownership structure and its potential impact on shareholder rights and the interests of other investors. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G04	Accounting	Companies are evaluated on the transparency, independence and effectiveness of their audit and financial reporting practices. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G05	Critical Incident Risk Management	The category addresses the company's use of management systems and scenario planning to identify, understand, and prevent or minimize the occurrence of low-probability, high-impact accidents, and emergencies with significant potential environmental and social externalities. It relates to the culture of safety at a company, its relevant safety management systems and technological controls, the potential human, environmental, and social implications of such events occurring, and the long-term effects to an0 organization, its workers, and society should these events occur.
		G06	Systemic Risk Management	The category addresses the company's contributions to, or management of systemic risks resulting from large-scale weakening or collapse of systems upon which the economy and society depend. This includes financial systems, natural resource systems, and technological systems. It addresses the mechanisms a company has in place to reduce its contributions to systemic risks and to improve safeguards that may mitigate the impacts of systemic failure. For financial institutions, the category also captures the company's ability to absorb shocks arising from financial and economic stress and meet stricter regulatory requirements related to the complexity and interconnectedness of companies in the industry.
		G07	Management of the Legal & Regulatory Environment	The category addresses a company's approach to engaging with regulators in cases where conflicting corporate and public interests may have the potential for long-term adverse direct or indirect environmental and social impacts. The category addresses a company's level of reliance upon regulatory policy or monetary incentives (such as subsidies and taxes), actions to influence industry policy (such as through lobbying), overall reliance on a favorable regulatory environment for business competitiveness, and ability to comply with relevant regulations. It may relate to the alignment of management and investor views of regulatory engagement and compliance at large.

5

Figure B.5: The guideline of ESG key issues in DynamicESG (5/6).

doi:10.6342/NTU202501476

Gover-nance (G)	Corporate Governance	G08	Business Model Resilience	The category addresses an industry's capacity to manage risks and opportunities associated with incorporating social, environmental, and political transitions into long-term business model planning. This includes responsiveness to the transition to a low-carbon and climate-constrained economy, as well as growth and creation of new markets among unserved and underserved socio-economic populations. The category highlights industries in which evolving environmental and social realities may challenge companies to fundamentally adapt or may put their business models at risk.
	Corporate Behavior	G09	Business Ethics	Companies are evaluated on their oversight and management of business ethics issues such as fraud, executive misconduct, corrupt practices, money laundering, or anti-trust violations. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G10	Tax Transparency	Companies are evaluated on their estimated corporate tax gap (i.e., gap between estimated effective tax rate and estimated corporate income tax rate), revenue reporting transparency and their involvement in tax-related controversies. This is a Key Issue in the Governance Pillar and is relevant for all companies.
		G11	Competitive Behavior	The category covers social issues associated with existence of monopolies, which may include, but are not limited to, excessive prices, poor quality of service, and inefficiencies. It addresses a company's management of legal and social expectation around monopolistic and anti-competitive practices, including issues related to bargaining power, collusion, price fixing or manipulation, and protection of patents and intellectual property (IP).
(E)	Not related to Company	E00	Related to Environmental	Not related to company but related to Environmental topic.
(S)		S00	Related to Social	Not related to company but related to Social topic.
(G)		G00	Related to Governance	Not related to company but related to Governance topic.
NN			None of the above	None of the above category. Not related to ESG topic.

Figure B.6: The guideline of ESG key issues in DynamicESG (6/6).

doi:10.6342/NTU202501476

Appendix C — Annotation Guidelines of the Six Existing Datasets

doi:10.6342/NTU202501476

Relation Extraction (RE) is the task of extracting relationships between entities in a sentence.

You will be given a sentence that contains two entities: entity 1 and entity 2.
Entity 1 is enclosed in double asterisks (i.e., **entity**) and entity 2 is enclosed in double underscores (i.e., __entity__).
Each entity has its own entity type specified in square brackets before the entity (e.g., [PERSON]**entity 1**).
The definition of the entity types are as follows:
- PERSON: People, including fictional.
- ORG: Companies, agencies, institutions, etc.
- UNIV: Universities, colleges, etc.
- GOV_AGY: Government agencies and departments.
- DATE: Absolute or relative dates or periods.
- GPE: Countries, cities, states.
- MONEY: Monetary values, including unit.
- TITLE: Positions or titles, including military.

Please annotate the relation between entity 1 and entity 2 described in the given sentence according to the following label descriptions.
Note that the relation is directional, meaning that the order of entity 1 and entity 2 matters.
Note that you can only select the most appropriate label that is consist of the given type of entities.
If you think there is no relation or other relation between entity 1 and entity 2, please select the label 0.

- 0: **entity 1** has no relation or other relation to __entity 2__
- 1: [PERSON]**entity 1** has/had the job title of [TITLE]__entity 2__
- 2: [PERSON]**entity 1** is/was an employee of [ORG]__entity 2__
- 3: [PERSON]**entity 1** is/was a member of [ORG]__entity 2__
- 4: [PERSON]**entity 1** is/was a founder of [ORG]__entity 2__
- 5: [PERSON]**entity 1** is/was a employee of [UNIV]__entity 2__
- 6: [PERSON]**entity 1** is/was a member of [UNIV]__entity 2__
- 7: [PERSON]**entity 1** has/had attended [UNIV]__entity 2__
- 8: [PERSON]**entity 1** is/was a member of [GOV_AGY]__entity 2__
- 9: [ORG]**entity 1** is/was formed on [DATE]__entity 2__
- 10: [ORG]**entity 1** is/was acquired on [DATE]__entity 2__
- 11: [ORG]**entity 1** is/was headquartered in [GPE]__entity 2__
- 12: [ORG]**entity 1** has/had operations in [GPE]__entity 2__
- 13: [ORG]**entity 1** is/was formed in [GPE]__entity 2__
- 14: [ORG]**entity 1** has/had shares of [ORG]__entity 2__
- 15: [ORG]**entity 1** is/was a subsidiary of [ORG]__entity 2__
- 16: [ORG]**entity 1** is/was acquired by [ORG]__entity 2__
- 17: [ORG]**entity 1** has/had a agreement with [ORG]__entity 2__
- 18: [ORG]**entity 1** has/had a revenue of [MONEY]__entity 2__
- 19: [ORG]**entity 1** has/had a profit of [MONEY]__entity 2__
- 20: [ORG]**entity 1** has/had a loss of [MONEY]__entity 2__
- 21: [ORG]**entity 1** has/had a cost of [MONEY]__entity 2__

Figure C.7: The annotation guideline of REFinD dataset.

doi:10.6342/NTU202501476

Hawkish-Dovish classification is to classify the sentiment about the future monetary policy stance into Dovish, Hawkish, or Neutral.
In general:
- 0: Dovish sentences were any sentence that indicates future monetary policy easing.
- 1: Hawkish sentences were any sentence that would indicate a future monetary policy tightening.
- 2: Neutral sentences were those with mixed sentiment, indicating no change in the monetary policy, or those that were not directly related to monetary policy stance.

You will be given a sentence that falls into one of the following eight categories enclosed in square brackets. Please annotate the sentiment of the sentence according to the following detailed label descriptions.
Note that you can only select one label that is most appropriate.

Detailed label descriptions:
[Economic Status: A sentence pertaining to the state of the economy, relating to unemployment and inflation.]
- 0: when inflation decreases, when unemployment increases, when economic growth is projected as low.
- 1: when inflation increases, when unemployment decreases when economic growth is projected high when economic output is higher than potential supply/actual output when economic slack falls.
- 2: when unemployment rate or growth is unchanged, maintained, or sustained.
[Dollar Value Change: A sentence pertaining to changes such as appreciation or depreciation of value of the United States Dollar on the Foreign Exchange Market.]
- 0: when the dollar appreciates.
- 1: when the dollar depreciates.
- 2: N/A
[Energy/House Prices: A sentence pertaining to changes in prices of real estate, energy commodities, or energy sector as a whole.]
- 0: when oil/energy prices decrease, when house prices decrease.
- 1: when oil/energy prices increase, when house prices increase.
- 2: N/A
[Foreign Nations: A sentence pertaining to trade relations between the United States and a foreign country. If not discussing United States we label neutral.]
- 0: when the US trade deficit decreases.
- 1: when the US trade deficit increases.
- 2: when relating to a foreign nation's economic or trade policy.
[Fed Expectations/Actions/Assets: A sentence that discusses changes in the Fed yields, bond value, reserves, or any other financial asset value.]
- 0: Fed expects subpar inflation, Fed expecting disinflation, narrowing spreads of treasury bonds, decreases in treasury security yields, and reduction of bank reserves.
- 1: Fed expects high inflation, widening spreads of treasury bonds, increase in treasury security yields, increase in TIPS value, increase bank reserves.
- 2: N/A
[Money Supply: A sentence that overtly discusses impact to the money supply or changes in demand.]
- 0: money supply is low, M2 increases, increased demand for loans.
- 1: money supply is high, increased demand for goods, low demand for loans.
- 2: N/A
[Key Words/Phrases: A sentence that contains key word or phrase that would classify it squarely into one of the three label classes, based upon its frequent usage and meaning among particular label classes.]
- 0: when the stance is "accommodative", indicating a focus on "maximum employment" and "price stability".
- 1: indicating a focus on "price stability" and "sustained growth".
- 2: use of phrases "mixed", "moderate", "reaffirmed".
[Labor: A sentence that relates to changes in labor productivity.]
- 0: when productivity increases.
- 1: when productivity decreases.
- 2: N/A

Figure C.8: The annotation guideline of FOMC dataset.

doi:10.6342/NTU202501476

A/P Relation classification is to classify the relation between Assessment and Plan Subsection in daily progress notes into DIRECT, INDIRECT, NEITHER, or NOT RELEVANT.

You will be given a pair of passages, Assessment and Plan Subsection, from daily progress notes.
Assessment describes the patient and establishes the main symptoms or problems for their encounter.
Plan Subsection addresses each differential diagnosis/problem with an action plan or treatment plan for the day.

Please annotate the relation between Assessment and Plan Subsection in the given pair according to the following label descriptions.
Note that you can only select one label that is most appropriate.

Label descriptions:
- 0: DIRECT. Assessment section includes a primary diagnosis/problem and it is mentioned in the Plan subsection, or Progress note includes a primary diagnosis/problem for hospitalization and it is mentioned in the Plan subsection, or Plan subsection contains a problem/diagnosis related to the primary signs/symptoms in the Assessment section.
- 1: INDIRECT. Plan subsection contains complications/subsequent events or organ failure related to the primary diagnosis/problem from the Assessment section, or Plan subsection contains other listed diagnoses/problems from the overall Progress Note or in the Assessment section that are not part of the primary diagnosis/problem, or Plan subsection contains a diagnosis/problem that is not previously mentioned but closely related (i.e., same organ system) to the primary diagnoses/problems mentioned in the overall Progress Note or Assessment section.
- 2: NEITHER. None of the criteria for Directly Related or Indirectly Related are met but a diagnosis/problem or other signs/symptoms are mentioned.
- 3: NOT RELEVANT. Plan subsection does not include a diagnosis/problems OR signs/symptoms.

Figure C.9: The annotation guideline of AP-Relation dataset.

You will be given one paper abstract comprising several segments.
Each segment is a short text describing a specific aspect of the paper, including background, purpose, method, finding/contribution, or other.

Please annotate the aspects of each segment according to the following label descriptions.
Note that you can only select one label that is most appropriate for each segment. The total number of labels must be equal to the number of segments in the abstract.

Label descriptions:
- 0: Background. “Background” text segments answer one or more of these questions: Why is this problem important?, What relevant works have been created before?, What is still missing in the previous works?, What are the high-level research questions?, How might this help other research or researchers?
- 1: Purpose. “Purpose” text segments answer one or more of these questions: What specific things do the researchers want to do?, What specific knowledge do the researchers want to gain?, What specific hypothesis do the researchers want to test?
- 2: Method. “Method” text segments answer one or more of these questions: How did the researchers do the work or find what they sought?, What are the procedures and steps of the research?
- 3: Finding/Contribution. “Finding/Contribution” text segments answer one or more of these questions: What did the researchers find out?, Did the proposed methods work? Did the thing behave as the researchers expected?
- 4: Other. Text segments that do not fit into any of the four categories above. Text segments that are not part of the article. Text segments that are not in English. Text segments that contain only reference marks (e.g., “[1,2,3,4,5]”) or dates (e.g., “April 20, 2008”). Captions for figures and tables (e.g. “Figure 1: Experimental Result of ...”). Formatting errors. Text segments the annotator does not know or is not sure about.

Figure C.10: The annotation guideline of CODA-19 dataset.

doi:10.6342/NTU202501476

You will be given a clause from a legal contract. Please annotate the category of the given clause according to the following label descriptions.
Note that you can only select one label for each segment that is most appropriate.

Label descriptions:
- 0: Most Favored Nation. This clause provides that if a third party gets better terms on the licensing or sale of technology/goods/services described in the contract, the buyer of such technology/goods/services under the contract shall be entitled to those better terms.
- 1: Non-Compete. This clause imposes a restriction on the ability of a Party to compete with the other party or operate in a certain geography or business or technology sector.
- 2: Exclusivity. This clause provides for an exclusive dealing commitment between the parties of a contract. This clause also includes: a commitment by a party to procure all “requirements” from the other party of certain technology, goods, or services; or a prohibition against licensing or selling technology, goods or services to third parties, or a prohibition on collaborating or working with other parties.
- 3: No-Solicit of Customers. This clause restricts a party from soliciting, contacting or doing business with the other party’s customers, vendors or partners.
- 4: Competitive Restriction Exception. This clause states the exception(s) to one of the following three labels: Exclusivity, Non-Compete, or No-Solicit of Customers.
- 5: No-Solicit of Employees. A No-Solicit of Employee clause prohibits a party from soliciting or hiring the other party’s employees or consultants for itself or for a third party, during the contract or after the contract ends (or both).
- 6: Non-Disparagement. This clause requires a party not to disparage or defame the other party’s goodwill, reputation or image.
- 7: Termination for Convenience. This clause allows a party to terminate a contract without cause or penalty. It allows a party to unilaterally terminate a contract by giving notice and oftentimes after a waiting period expires.
- 8: Right of First Refusal, Offer or Negotiation (Rofr/Rofo/Rofn). This clause grants one party a right of first refusal, right of first offer or right of first negotiation to purchase, license, market, or distribute equity interest, technology, assets, products or services.
- 9: Change of Control. This clause requires consent or notice of the other party if a party undergoes a change of control, such as a merger, stock sale, transfer of all or substantially all of its assets or business (collectively, “CIC”).
- 10: Anti-Assignment. This clause requires a party to seek consent or notice if the contract is assigned, transferred or sublicensed to a third party, in whole or in part.
- 11: Revenue/Profit Sharing. This clause requires one party to share revenue or profit with the other party for any technology, goods, or services.
- 12: Price Restriction. This clause restricts the ability of a party to raise or reduce prices of technology, goods, or services provided.
- 13: Minimum Commitment. This clause requires a minimum order size or minimum amount or units per-time period that one party must buy from the counterparty under the contract.
- 14: Volume Restriction. This clause charges a fee or requires consent if one party’s use of the product/services exceeds a certain threshold.
- 15: IP Ownership Assignment. This clause provides that intellectual property created by one party becomes the property of the other party, either per the terms of the contract or upon the occurrence of certain events.

Figure C.11: The annotation guideline of CUAD dataset (1/2).

doi:10.6342/NTU202501476

- 16: Joint IP Ownership. This clause provides for joint or shared ownership of intellectual property between the parties to the contract.
- 17: License Grant. This clause authorizes a party to use intellectual property or intangibles of the other party. It can be an authorization to use or to reproduce, distribute, manufacture, etc. certain content, technology, or other items that are protected by intellectual property rights. This clause is very common, and is considered one of the "factual" clauses. The purpose of this label is to help human reviewers to understand what IP is licensed under a contract and what restrictions are imposed on the license, including restrictions on duration, territory and purpose of use.
- 18: Non-Transferable License. This clause prohibits one party to transfer, assign or sublicense IP in the contract.
- 19: Affiliate IP License-Licensor. This clause contains a license grant by affiliates of the licensor or that includes intellectual property of affiliates of the licensor.
- 20: Affiliate IP License-Licensee. This clause contains a license grant to a licensee (incl. sublicensor) and the affiliates of such licensee/sublicensor.
- 21: Unlimited/All-You-Can-Eat License. This clause contains a provision granting one party an "enterprise," "all you can eat" or unlimited usage license.
- 22: Irrevocable or Perpetual License. This clause contains an irrevocable and/or perpetual license of IP. An irrevocable license is a perpetual license that cannot be cut short or terminated. A perpetual license, on the other hand, may not be irrevocable. Namely, a perpetual license can be terminated upon specified events such as material breach. Many license grant clauses use "irrevocable" and "perpetual" in the same sentence. The intent of some contracts may be to use the two terms interchangeably. As a result, for the purpose of CUAD, you should label the two types of licenses under the same label.
- 23: Source Code Escrow. This clause requires one party to deposit its source code into escrow with a third party or into a deposit account with the other party, which can be released to the other party upon the occurrence of certain events (bankruptcy, insolvency, etc.).
- 24: Post-Termination Services. This clause imposes obligations on a party after the termination or expiration of a contract, including any post-termination transition, payment, transfer of IP, wind-down, last-buy, or similar commitments.
- 25: Audit Rights. This clause grants one party the right to audit the books, records, or physical locations of the other party to ensure compliance with the terms of a contract.
- 26: Uncapped Liability. This clause leaves a party's liability uncapped upon the breach of its obligation in the contract. This also includes uncap liability for a particular type of breach such as IP infringement or breach of confidentiality obligation.
- 27: Cap On Liability. This clause includes a cap on liability upon the breach of a party's obligation. This includes time limitation for the counterparty to bring claims or maximum amount for recovery.
- 28: Liquidated Damages. This clause is an agreement to pay a party a pre-determined amount of damages if the other party breaches the contract. For the purpose of CUAD, this clause also includes an early termination fee.
- 29: Insurance. This clause requires a party to maintain insurance for the benefit of the other party.
- 30: Covenant not to Sue. This clause restricts a party from contesting the validity of the other party's ownership of intellectual property or otherwise bringing a claim against the other party that goes beyond the scope of standard Limitation on Liability clauses.
- 31: Third Party Beneficiary. This clause provides that a non-contracting party is a beneficiary to some or all of the clauses in the contract and therefore can enforce its rights against a contracting party.

Figure C.12: The annotation guideline of CUAD dataset (2/2).

doi:10.6342/NTU202501476

You will be given a one-paragraph excerpt of a legal decision. Please annotate the category of the given excerpt according to the following label descriptions.
Note that you can only select one label that is most appropriate for the excerpt.

Label descriptions:
- 0: Facts. A section of the decision that recounts the historical events and interactions between the parties that gave rise to the dispute.
- 1: Procedural History. A section of the decision that describes the parties' prior legal filings and prior court decisions that led up to the issue to be resolved by the decision.
- 2: Issue. A section of the decision that describes a legal or factual issue to be considered by the court.
- 3: Rule. A section of the decision that states a legal rule relevant to resolution of the case.
- 4: Analysis. A section of the decision that evaluates an issue before the court by applying governing legal principles to the facts of the case
- 5: Conclusion. A section of the decision that articulates the court's conclusion regarding a question presented to it.
- 6: Decree. A section of the decision that announces and effectuates the court's resolution of the parties' dispute, for example, granting or denying a party's motion or affirming, vacating, reversing, or remanding a lower court's decision.

Figure C.13: The annotation guideline of FoDS dataset.

doi:10.6342/NTU202501476