Keyword-based Annotation of Visually-Rich Document Content for Trend and Risk Analysis using Large Language Models

Giuseppe Gallipoli[1], **Simone Papicchio**[1], **Lorenzo Vaiani**[1], **Luca Cagliero**[1], **Arianna Miola**[2,3], **Daniele Borghi**[2]

[1]Politecnico di Torino, Turin, Italy
[2]Intesa Sanpaolo Innovation Center, Turin, Italy
[3]Università degli Studi di Milano-Bicocca, Milan, Italy
{name.surname}@polito.it, {name.surname}@intesasanpaolo.com

Abstract

In the banking and finance sectors, members of the business units focused on Trend and Risk Analysis daily process internal and external visually-rich documents including text, images, and tables. Given a facet (i.e., topic) of interest, they are particularly interested in retrieving the top trending keywords related to it and then use them to annotate the most relevant document elements (e.g., text paragraphs, images or tables). In this paper, we explore the use of both open-source and proprietary Large Language Models to automatically generate lists of facet-relevant keywords, automatically produce free-text descriptions of both keywords and multimedia document content, and then annotate documents by leveraging textual similarity approaches. The preliminary results, achieved on English and Italian documents, show that OpenAI GPT-4 achieves superior performance in keyword description generation and multimedia content annotation, while the open-source Meta AI Llama2 model turns out to be highly competitive in generating additional keywords.

Keywords: Visually-Rich Document Understanding, Trend and Risk analysis, Large Language Models

1. Introduction

Understanding and exploring the content of visually-rich documents such as PDF files and scanned documents is of primary importance for trend and risk analysts of the banking and finance sectors. Since these documents have variable layout and content, with a mixture of text, images, and tables, their deep understanding requires both advanced multimodal learning capabilities.

The goal of this work is to enhance the research and analysis capabilities of a primary Italian financial institution, focusing on emerging trends within both the national and international contexts. Improving these functions is crucial for the strategic positioning of the bank and for providing value-added services to its customers. The partial automation of the research process allows for the inclusion of a greater number of data sources that were previously untapped due to operational limits. Given the relentless flow of information in today's environment, this represents a strategic step towards expanded informational access and a stronger ability to proactively adapt to market evolution.

In this work, we provide bank analysts with a financial document annotator relying on multimodal Large Language Models (LLM). Given a topic of interest (hereafter denoted by *facet*), the LLM produces a list of facet-related keywords as well as the corresponding textual descriptions and high-dimensional vector representations. In parallel, the multimodal document content is split into textual paragraphs, images, and tabular elements and conveniently processed to generate embeddings of the equivalent text versions. Finally, the annotation process is tackled as a keyword retrieval task on the document elements driven by textual semantic similarity. An extensive empirical analysis, supported by a bilingual testing document collection and a team of experts who validated the keyword descriptions, provide an in-depth performance comparison between the open-source Meta AI Llama2 and the proprietary OpenAI GPT-4 models.

2. Problem statement

Given a set of multi-page financial documents $\mathcal{D}$ and a set of facets $\mathcal{F}$ describing the topics of interest, our purpose is threefold:

1. **Keyword generation and description**. Generate for each facet $f_i \in \mathcal{F}$ a set of keywords $k_j \in \mathcal{K}^i$ related to f_i. Next, annotate each keyword k_j with a free-text description *descr*(k_j) summarizing its general meaning.
2. **Captioning of non-textual document elements**. Produce textual descriptions of multimedia document elements $e_l \in \mathcal{E}^m$, where an arbitrary element e_l in a document $d_m \in \mathcal{D}$ can be either an image, a table, or a textual paragraph.

20 May, 2024.

3. **Keyword-based content annotation**. For each element e_l, retrieve the keywords k_j that are most relevant to e_l.

Our goal is to compare the performance of LLMs, in zero-shot or few-shot learning, to address all the above-mentioned tasks. Hereafter, we will consider Llama2 (Touvron et al., 2023) (or its Italian version Camoscio (Santilli and Rodolà, 2023)) as representative open-source model and GPT-4 (OpenAI, 2023) as representative proprietary model.

3. Proposed approach

In the following, we describe the main steps of our method. A sketch of the proposed pipeline is displayed in Figure 1.

3.1. Generation of keywords and keyword descriptions

Given a user-provided facet name f_i, we use the LLM to automatically generate a set of related keywords k_j as well as the corresponding free-text descriptions *descr*(k_j).

We explore the following settings:

- *Zero-Shot learning – Cold Start Setting*: We prompt the LLM with the facet name only, assuming that neither facet-relevant keywords nor examples of textual descriptions are given.

- *Few-Shot learning – Cold Start Setting*: We prompt the LLM with the facet name and h examples chosen randomly from keywords and their corresponding descriptions, previously provided by the domain expert. Here, we assume that some examples of keyword descriptions are already available, but we do not know any facet-related keyword yet, since the selected examples are not necessarily related to the input facet.

- *Few-shot learning – Additional Keyword Recommendation*: We prompt the LLM with the facet name and h examples of facet-related keywords and their corresponding descriptions. Here, we assume that the examples are not chosen randomly but shortlisted by human expert (e.g., by validating a previous output).

In few-shot learning settings, we ensure that the examples of keywords and descriptions provided as input to the LLM do not overlap with the keyword currently being prompted.

The output of this step is then used in the keyword-based content annotation stage.

3.2. Document pre-processing

To process the input PDF documents, we extract the following three main elements: (i) Textual paragraphs (e.g., titles, sections, subsections), (ii) Visual items (e.g., images, sketch of architectures/processes/pipelines, iconography, graphical examples), and (iii) Tables.

Textual paragraphs and tables are extracted from PDF documents using the proprietary Document Intelligence service provided by the Azure AI platform (Azure, 2024). For visual and textual content extraction, we face the following challenges:

- *Slide extraction*: Some input documents consist of slide presentations, which appear to be unsuitable for text and image extraction using standard content extraction tools. To address this issue, we opportunistically generate textual explanations of the slide content using the Multimodal Large Language Model GPT-4 Vision (OpenAI, 2023). Specifically, we train an ad hoc CNN to automatically detect the presence of presentation slides on a PDF document page. If the current page is classified as a *slide*, then the input is processed directly by the Multimodal LLM.

- *Paragraph length*: Some extracted textual elements contain few words (likely due to a misalignment of PDF content). To avoid this issue, we prevent the generation of textual elements consisting of less than 4 words.

- *Redundant table content*: The textual content within table cells sometimes appears incorrectly twice, in separate tabular and textual elements. During table extraction, we early detect possible situations of overlap between the bounding box of the table and the position of the text. The purpose is to disregard duplicated text whenever it is not deemed relevant.

- *Irrelevant images*: The image detector module also recognizes irrelevant visual items such as banners or graphical separators. We define the boundary regions of each document page (e.g., the bottom of the page) and ignore all the images placed in those border regions, as they are unlikely to convey informative content. To prune irrelevant content, we apply the following filters to all visual elements: (1) *Minimum image size*: we drop visual elements containing less than 150 pixels; (2) *Minimum height-width ratio*: we drop visual elements whose absolute ratio is above 500% (i.e., greater than 5:1 or 1:5); (3) *Percentage of pixels of the same color*: we drop visual elements whose percentage of pixels with the same color is above 80%.

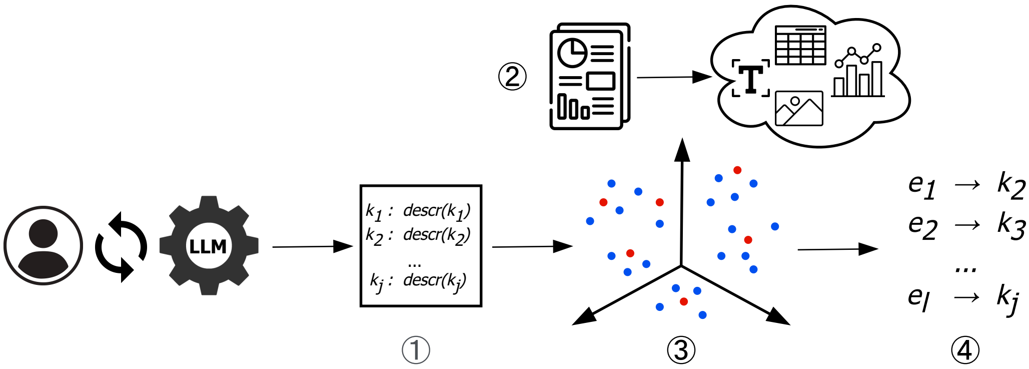

Figure 1: The figure illustrates the main steps of the proposed method: (1) keyword and description generation; (2) document preprocessing; (3) document element and keyword description encoding; and (4) keyword-based content annotation. In step (3), blue and red dots represent the embedding representations of document elements and keyword descriptions, respectively.

3.3. Keyword-based content annotation

For each document element e_l within each multi-page financial document $d_m \in \mathcal{D}$, we retrieve the keywords k_j that are the most relevant to e_l. Specifically, we return a ranked list $k^1_{e_l,d_m}, ..., k^K_{e_l,d_m}$ of the top-K keywords assigned to e_l. Notice that the assigned keywords can arbitrarily refer to any facet and, possibly, the retrieved list can be empty.

Focusing on text-only content, we address the retrieval of keywords relevant to each element in an unsupervised fashion using various textual similarity approaches, including both syntax-oriented and semantic-oriented methods. For each element of the document e_l, we assign the K keywords whose textual descriptions are most similar to e_l according to the following measures:

- Syntactic similarities: (1) **ROUGE-1/2/L F1-Score** (Lin, 2004) measures syntactic overlap in terms of common unigrams, bigrams or longest matching subsequence; (2) The **Levenshtein**, **Jaro**, and **Jaro-Winkler edit distances** measure the number of character-level operations needed to transform one piece of text into another.
- Semantic similarity: **SentenceBERT** (Reimers and Gurevych, 2019) and **proprietary embeddings**, used to compare document elements and keyword descriptions via cosine similarity.

Additionally, we experiment with **prompting GPT-4** with both the document element to be labeled and all possible keywords, asking the model to assign the K most pertinent ones.

Notice that, for the sake of simplicity, in Figure 1 document elements and keyword descriptions are displayed as embedding representations in a latent space. However, we also experiment with the syntactic similarity and prompting approaches discussed above.

4. Experimental evaluation

We run our experiments on a machine equipped with a single NVIDIA® RTX A6000 48GB GPU. We leverage standard Python libraries to calculate syntactic similarity measures, while for semantic similarity we rely on SentenceBERT `paraphrase-MiniLM-L6-v2` model and `text-embedding-ada-002` as proprietary OpenAI model. We employ Llama2-Chat 7B with 16-bit quantization. GPT-4 (`gpt-4-0613`), GPT-4 Vision (`gpt-4-1106-vision-preview`) and `text-embedding-ada-002` have all been accessed through OpenAI API.

Dataset. Business Units provided the following two in-domain datasets: (1) **ICT Risk Analysis**, consisting of 11 documents and annotated with 2 facets and 25 keywords. It contains 991 textual elements, 13 images, and 15 tables. (2) **Trend Analysis**, consisting of 4 documents, annotated with 1 facet and 12 keywords, and including 69 images. Most images are presentation slides, which are handled by the LLM to get the textual reformulation. We also have additional facets and keywords, along with their corresponding descriptions (92 overall), which analysts have not used for element annotation.

Evaluation Metrics. To evaluate the efficacy of element annotation, we employ the following metrics for information retrieval (Manning et al., 2008):

- **Precision at K** (P@K): percentage of returned keywords that occur in the expected keyword list.
- **Recall at K** (R@K): percentage of expected keywords that occur in the returned keyword list.
- **Mean Reciprocal Rank** (MRR): mean of the multiplicative inverse of the rank of the first correctly assigned keyword.

where K is the number of keywords retrieved that are considered. The rank order is based on the similarity score used to retrieve the keywords.

To assess keyword and description generation, we compare the produced and expected outcomes using the following established metrics for evaluating sequence-to-sequence models, i.e., ROUGE-1/2/L (R1/2/L) F1 score (Lin, 2004) for syntactic similarity and BERTScore (BS) F1-score (Zhang et al., 2020) for semantic similarity.

Prompt description. We present in the following some examples of prompts provided to the LLMs to perform keyword and description generation tasks. Prompts were selected according to preliminary experiments and their format may vary depending on the LLM under consideration.

Keyword generation: *The [K] most relevant keywords for the [FACET] domain are:*
where we replace [K] and [FACET] with the desired number of keywords and the facet name of interest, respectively.

Description generation: *Explain in a few lines the word between the quotation marks: “‘[KEYWORD]’”*
where we replace [KEYWORD] with the keyword for which to generate a description.

When conducting experiments in the Italian language, we use the corresponding Italian translations as prompts.

4.1. Results on content annotation

Similarity measure	ICT Risk Analysis	Trend Analysis
R1	0.458	0.300
R2	0.367	0.279
RL	0.472	0.258
Levenshtein	0.347	0.247
Jaro	0.483	0.249
Jaro-Winkler	0.483	0.249
SentenceBERT	0.658	0.430
embedding-ada-002	**0.779**	**0.610**
GPT-4	0.729	0.500

Table 1: Mean Reciprocal Ranks.

Textual semantic similarity based on contextual embeddings and LLM prompting achieve very promising results (MMR above 0.7) and outperform both syntactic measures and edit distances (see Table 1). System's precision decreases while increasing the number K of retrieved keywords whereas its recall shows an opposite trend (see Figure 2). Similarity based on OpenAI embedding performs best, e.g., for $K = 3$, P@K $>$ 40% and R@K $>$ 50% on both ICT Risk and Trend.

4.2. Results on keyword and description generation

Tables 2 and 4 summarize system performance on keyword description and keyword generation tasks, respectively. Due to space constraints, we report here only the outcomes on a single dataset, i.e., ICT Risk, for both languages.

- *Proprietary vs. open-source LLM*: Proprietary GPT-4 performance is superior to that of open-source (Llama2/Camoscio) on keyword description generation for both tested languages (e.g., +33% ROUGE-1 on Italian documents). Conversely, open-source LLMs are highly competitive on keyword generation, likely because training examples of smaller models are more focused on specific domains, such as finance. This trend is confirmed by the results on Italian documents (not shown here due to space constraints).
- *Italian vs. English*: Both LLMs perform better on English than Italian text. The gap in performance is more evident for the open-source LLMs, e.g., ROUGE-1 for description generation 0.31 Italian vs. 0.39 English.
- *In-context learning*: Prompting LLMs with few training examples (from 3 to 5) turns out to be beneficial for both keyword generation and description generation. Few-shot learning has shown to be more beneficial for open-source LLMs because of their lower pre-trained model complexity.

4.3. Human evaluation

Each generated description, for both Italian and English languages, was annotated by five domain experts using a 5-point Likert scale based on five criteria (Iskender et al., 2021): (1) **Usefulness** (effectiveness in conveying key information); (2) **Coherence** (logical and semantic coherence); (3) **Non-Redundancy** (conciseness); (4) **Grammaticality** (linguistic correctness); (5) **Overall Quality** (holistic evaluation of the generated description).

Results (see Table 3) are satisfactory and coherent with quantitative outcomes (see Section 4.2). The perceived quality of Italian-written descriptions is lower than that of English ones, likely due to the more limited capabilities of LLMs on languages other than English.

4.4. Qualitative examples

To better illustrate the proposed approach, we provide examples of outputs of the different steps of our method.

Considering the ICT Risk Analysis dataset, one of the keywords associated with the *cyber risk* facet is *third-party risk*.

Reference description: *It refers to the potential risks or threats to an organization arising from relationships with third parties, such as suppliers, business*

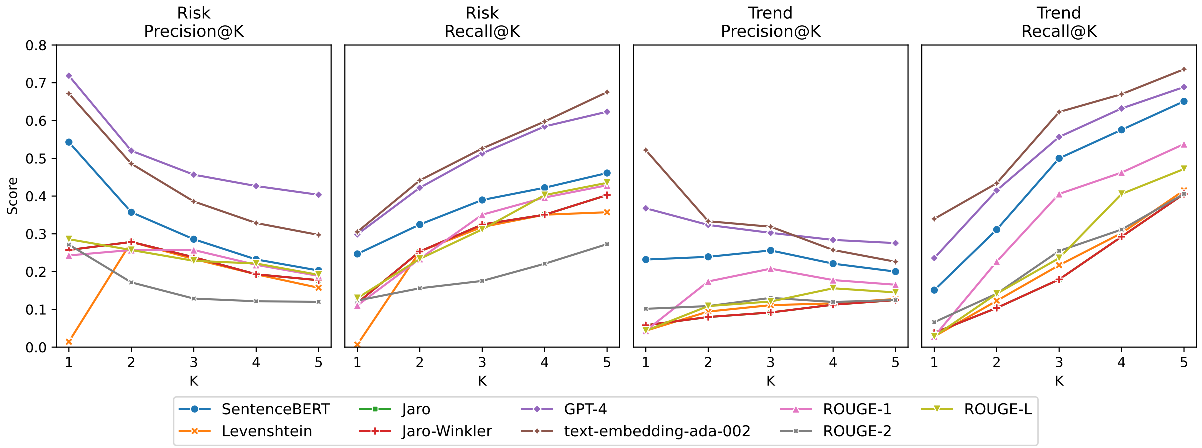

Figure 2: Precision@K and Recall@K values of different similarity measures on the ICT Risk Analysis (left) and Trend Analysis (right) datasets. English language.

	K = unspecified		$K = 3$		$K = 5$		$K = 10$		$K = 20$	
	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2
RL	0.051	**0.066**	**0.070**	0.062	0.058	**0.062**	0.057	**0.065**	0.058	**0.060**
BS	**0.860**	0.859	0.862	**0.864**	0.861	**0.863**	**0.865**	0.860	**0.861**	0.857
P@K	0.771	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.867	**0.944**	0.833	**0.894**
R@K	**0.447**	0.296	**0.133**	**0.133**	**0.221**	**0.221**	0.375	**0.420**	0.721	**0.783**

Table 4: Evaluation of keyword generation for varying K. ICT Risk Analysis dataset. English language.

	Italian		English	
	Camoscio	GPT-4	Llama2	GPT-4
R1	0.310	**0.413**	0.394	**0.437**
R2	0.082	**0.169**	0.131	**0.150**
RL	0.208	**0.279**	0.254	**0.284**
BS	0.719	**0.773**	0.760	**0.902**

Table 2: Evaluation of keyword description generation performance. ICT Risk Analysis dataset.

	Italian	English
Usefulness	**4.38**±1.60	4.33±1.78
Coherence	4.52±0.93	**4.62**±1.15
Non-Redundancy	4.38±1.12	**4.52**±1.11
Grammaticality	4.60±0.73	**4.81**±1.20
Overall Quality	4.33±1.37	**4.34**±1.66

Table 3: Human evaluation of keyword descriptions. ICT Risk Analysis dataset.

partners, or external contractors. These risks [...]

Generated description: *It is the risk that arises from the use of third-party vendors, suppliers, or partners that provide goods or services to an organization. Third-party risk can include a wide range of* [...]

Document element: *The image presents* [...] *in the context of retail banking leaders.* [...] *security providers aim to protect company, payment, card, and consumer data* [...] *the importance of various data privacy and security measures and where they stand in terms of industry focus and market trends.*

Target keywords: *third-party risk*, *regulation*
Assigned keywords: *third-party risk*, *regulation*, *compliance*

5. Conclusions

We presented an automatic pipeline for annotating visually-rich financial documents for Trend and Risk analysis in banking and finance sectors. The main takeaways can be summarized as: (1) *Semantic similarity*: proprietary embeddings outperform open-source solutions for both Italian and English text; (2) *Keyword generation*: open-source LLMs perform as good as or even better than GPT-4 in zero-shot and few-shot learning settings on the tested documents, likely due to a higher in-domain specialization; (3) *Description generation*: GPT-4 performs best, while open-source LLMs perform reasonably well. Human feedback is in line with quantitative results based on established performance metrics.

As future work, we will explore the integration of the proposed method in a Retrieval Augmented Generation system and address the task of zero-shot document classification using the additional keywords that have not been used for annotation yet. Moreover, we plan to assess the capabilities of other Multimodal LLMs (e.g., LLaVA (Liu et al., 2023)) to generate textual descriptions of multimedia document elements.

Limitations

Text-only processing. In this work, we focus on textual content, whether the content is originally text or is converted from a visual format. Consequently, we have not embedded visual and tabular content directly. We deemed such variant as a potentially valuable extension of the present work as enables the adoption of state-of-the-art multimodal learning techniques.

Limited robustness to document layout variety. Despite our efforts, the substantial variety in document structures, both within and across different domains, may introduce inconsistencies in document pre-processing and content extraction. This could potentially lead to suboptimal results in the subsequent content annotation phase. We intend to refine the document pre-processing and content extraction phase in alignment with the availability of new state-of-the-art document layout understanding models.

Limited scope of Multimodal LLM reasoning. When textual content cannot be successfully extracted, in particular from presentation slides, we rely on the GPT-4 Vision model to generate textual explanations. Although the LLM may disregard potential useful content and/or introduce inaccuracies in the generated textual explanation, also based on a manual inspection of a sample of outputs, we are confident that this approach is sufficiently satisfactory. However, we plan to conduct a more in-depth analysis to assess the LLM capabilities in providing textual explanations of visually-rich domain-specific content.

Ethical Considerations

The use of Large Language Models in critical sectors like banking and finance offers significant advantages, including improved efficiency, automation, and enhanced data analysis capabilities. These models can optimize processes, improve customer interactions, and contribute to informed decision-making. However, it is crucial to acknowledge that deploying LLMs in these domains also presents potential risks and undesired outcomes.

The complexity of banking and financial systems, coupled with the intricate nature of language understanding, may result in unintended consequences, such as biased generations or misinterpretation of information. However, in our specific use case, we view the LLM as an assistant to domain experts who remain fully responsible for the process, supervising the system outputs and possibly refining them through a human-in-the-loop approach. We believe that vigilant oversight, continuous refinement, and ethical considerations are essential to fully exploit the potential of LLMs while minimizing any adverse impacts on critical sectors such as the ones of our work.

We also acknowledge that the use of proprietary models may hinder transparency. However, the active involvement of domain experts who supervise the process is expected to alleviate this issue. Additionally, we sought to address this concern by experimenting also with open-source models.

Data and Code Availability Statement

Documents cannot be disclosed due to confidentiality and copyright issues. Code could be available upon request to the authors.

Conflicts of Interest

We have no Conflicts of Interest to declare.

Credits to financial institutions

Intesa Sanpaolo is a leading banking group in the Eurozone, and the most important one in Italy. Intesa Sanpaolo Innovation Center is part of ISP group, and its mission is exploring business models of the future to discover new assets and skills that support the long-term competitiveness of ISP group and its customers. ISP has established the Innovation Center Labs to respond to the complex needs of the bank and the market, determined by the evolution of market trends and exponential growth technologies.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments.

They would also like to thank Giorgio Bella, Anna Polise, Stefania Vigna, Laura Li Puma, Chiara Napione, Giovanni Troiano, Patrizio Paolo Dionisi, Maura Bertaglia, Carla Monferrato, and Simone Scarsi for their useful comments. The authors would like to thank Luigi Ruggerone for supporting the researcher team.

This work was partially carried out within the FAIR (Future Artificial Intelligence Research) and received funding from the European Union NextGenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000013).

The work by Giuseppe Gallipoli was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

6. Bibliographical References

Microsoft Azure. 2024. Azure AI Document Intelligence.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an Italian instruction-tuned LLaMA.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.