E-commerce products
Project Proposal for NLP Course, Winter 2022

S. Rećko WUT 01151399@pw.edu.pl
M. Sperkowski WUT 01151430@pw.edu.pl
P. Tomaszewski WUT 01151442@pw.edu.pl
K. Ułasik WUT 01151444@pw.edu.pl
supervisor: A. Wróblewska WUT anna.wroblewska1@pw.edu.pl

Abstract

This project's goal is to present an approach leveraging machine learning, especially Natural Language Processing (NLP), techniques to establish a robust similarity measure between products in e-commerce. The research aims to differentiate between identical, slightly different, and distinct products through taxonomy enrichment, information extraction from descriptions, and potential attribute generation from item descriptions.

1 Introduction

The overarching goal of this project is to develop an innovative and robust framework leveraging machine learning and NLP techniques to create an automated and comprehensive system capable of measuring the similarity between products across multiple dimensions within e-commerce platforms. This research aims to achieve an advanced level of accuracy in defining and establishing similarity metrics between products based on various attributes, including taxonomy (categories), textual descriptions and titles. To be practical for the end users, calculations must be performed quickly. Therefore, the calculations need to work in real-time, so we set a limit of at least 0.01 seconds (10ms) per pair of products. We consider this a crucial requirement for the project.

The primary scientific objective is to construct a sophisticated similarity measurement that accurately captures the semantics and nuances of product relationships, distinguishing between identical, closely related (such as variations within the same product line differing in specifications like memory sizes), and entirely distinct products. This entails the exploration and development of algorithms that can comprehend and differentiate the diverse relationships among products, accommodating subtle variations in features while still recognizing commonalities.

Furthermore, the project will focus on innovating techniques for automatically extracting key information from item descriptions and titles using deep learning algorithms. These generated attributes will supplement existing data, enriching the product information available on e-commerce platforms.

Ultimately, the scientific aim is to significantly enhance e-commerce search functionalities, facilitate the simple yet efficient introduction of new products into online marketplaces, and provide users with access to a comprehensive range of available price ranges for similar products. This research endeavour will contribute to advancing the capabilities of e-commerce platforms by refining the accuracy and depth of product recommendations, thereby improving the overall user experience in online product search and comparison.

The rise of e-commerce has been an unprecedented change in the sales market. Over the past years, e-commerce has witnessed exponential growth, accelerated further by global shifts in consumer behavior, especially the increased adoption of online shopping. The convenience, accessibility, and wide array of products available online have transformed how people shop. The COVID-19 pandemic further expedited this shift, with many consumers transitioning to online platforms for everyday needs. As e-commerce continues to expand, the need for efficient recommendation systems becomes even more critical, (Zhang et al., 2019). With an ever-growing number of products available, these systems play a pivotal role in helping consumers navigate through the vast array of choices, making their shopping experiences more personalized, streamlined, and enjoyable.

This system harnesses the power of data, algorithms, and user behavior to offer personal-

ized product recommendations, thereby creating a more engaging and tailored shopping experience.

By analyzing user preferences, purchase history, and behavior, an e-commerce recommendation system can offer personalized suggestions. This level of customization enhances user experience, making shopping more efficient and enjoyable (Bobadilla et al., 2013).

Tailored recommendations lead to increased user engagement and extended time spent on the platform, potentially resulting in higher conversion rates and sales.

When users are presented with items that align with their interests, the likelihood of making a purchase increases. This directly impacts the conversion rates and revenue of the e-commerce platform.

Recommendation systems can suggest related or complementary products, effectively enabling upselling and cross-selling, which contribute to increased average order value.

Utilizing machine learning and algorithms, recommendation systems can predict future trends and customer preferences, assisting in inventory management and product development.

When customers find what they are looking for effortlessly, they tend to have a more satisfying shopping experience. This satisfaction leads to customer loyalty and retention.

Encouraging return visits and repeat purchases is crucial for the sustainability of any e-commerce platform. Tailored recommendations play a significant role in achieving this goal.

Building an e-commerce recommendation system involves cutting-edge technologies, fostering advancements in AI and machine learning applications. These advancements have broader implications for various industries beyond e-commerce.

1.1 Research questions

In e-commerce, the sheer volume and diversity of products pose a significant challenge in accurately measuring their similarity across multiple dimensions. The current methodologies for comparing and categorizing products often fall short in capturing the nuanced differences and similarities between items. This leads to sub-optimal search experiences for users and hampers the efficiency of introducing new products into e-commerce platforms. The need for an automated and precise system to measure product similarity, considering varying attributes and features, is crucial for enhancing product search accuracy and user satisfaction.

Research Questions:

- Is there a measure that can incorporate the taxonomy of products in a similarity measurement and if so, can it be used as a loss for fine-tuning transformer models?
- Can we leverage the LLMs complexity to augment the data and achieve better representation of the products?
- Can real-time performance of the pair similiarity measurement be achieved using transformer models? (Reaching 10ms per pair comparison)

1.2 Report structure

In Section 2 we describe the scientific literature related to our project. We focus on the state-of-the-art in NLP and text embeddings, ecommerce recommendation systems, similarity measures and Large Language Models. Furthermore the closely related works are described in detail, together with the datasets used in research. In Section 3 the research methodology adopted in this project is described, especially the techniques and tools for both research and result analysis. In Section 4 the experiments on the WDC datasets are described, including the exploratory data analysis, data augmentation using LLMs and finetuning of the transformer model. In Section 5 a discussion of our results in comparison to the literature is made. In Section 6 the project conclusions are written, summing up what has been done and proposing future works for this project and the domain.

2 Related works

Machine learning approaches for extracting meaningful attributes from product descriptions include NLP techniques, word embeddings, and Named Entity Recognition (NER). These attributes can be integrated into the similarity measurement process by converting them into feature vectors for products. Utilizing metrics such as cosine similarity or Euclidean distance on these vectors allows for an effective quantification of product similarity, enhancing the overall recommendation system.

Siamese Neural Networks (Koch et al., 2015), Graph Neural Networks (Scarselli et al., 2008) (GNN), and Transformer models (Vaswani et al.,

2017) like BERT (Devlin et al., 2019) excel in capturing semantics and nuanced relationships between products. Siamese networks are adept at understanding subtle differences, GNNs model complex dependencies, and Transformers provide contextualized embeddings, collectively offering a robust framework for differentiating between identical, slightly different, and entirely distinct items.

The integration of generated product attributes into existing e-commerce platforms can be achieved by developing an attribute-based search functionality, enhancing recommendation systems, and optimizing the introduction of new products. By leveraging these attributes, platforms can offer more personalized search options, improve recommendation accuracy, and streamline the process of introducing new products efficiently into the market, ultimately enhancing the overall user experience.

To reach minimal pair comparison time while utilizing State-of-the-Art (SOTA) models with complex architectures, various strategies can be employed. Techniques like model quantization (Polino et al., 2018) and pruning (Liu et al., 2018) help reduce the computational load, while hardware acceleration using specialized processors like GPUs or TPUs speeds up inference. Additionally, caching and batch processing can be implemented to precompute certain calculations and perform parallelized comparisons, ensuring efficient and real-time processing without compromising the sophistication of the underlying models. Distilling the knowledge of a bigger model to a smaller one, by enforcing similar outputs on the training data, is a another method commonly used for creating smaller networks (Gou et al., 2021).

In recent years, Large Language Models (LLMs) have emerged as transformative tools across various domains, showcasing their remarkable versatility and utility. These models, such as GPT-3 (Brown et al., 2020), have demonstrated an unprecedented ability to understand and generate human-like text, enabling advancements in natural language processing, text generation, and information retrieval. In the realm of artificial intelligence, LLMs have been pivotal in enhancing chatbot capabilities, language translation, and content creation. Moreover, they have proven instrumental in automating mundane tasks, facilitating more efficient data analysis, and even contributing to the development of novel applications in healthcare (Thirunavukarasu et al., 2023), finance (Wu et al., 2023), and education (Kasneci et al., 2023). The extensive capabilities of Large Language Models (Wei et al., 2022) underscore their potential to revolutionize how we interact with technology, opening up new frontiers for innovation and problem-solving across diverse fields.

3 Approach & research methodology

In this section, we will first provide an overview of the dataset we will be using, followed by a description of methodologies related to our project. Last subsection contains explanation of our chosen approach.

3.1 Dataset

To create our solution, we have selected the Web Data Commons - Training Dataset and Gold Standard for Large-Scale Product Matching as the primary dataset for our project. Several key motivations drive the dataset selection. In recent years, entity resolution has shifted towards deep learning-based matching methods, necessitating large training data. Traditional benchmark datasets often prove inadequate for evaluating these methods due to their limited size and source diversity. The ”Web Data Commons” dataset addresses these challenges by offering a substantial volume of data, including 16 million English-language offers, sourced from 79 thousand websites. This diversity and scale make it an ideal choice for assessing deep learning-based matches and improving their evaluation and comparison. The dataset includes categorization using distant supervision from Amazon product data. Lexica containing terms and their TF-IDF scores for 26 product categories (look Figure 1) were created using publicly available Amazon product reviews and metadata. Each offer in the dataset is assigned the product category whose terms maximize the sum of overlapping TF-IDF scores. In cases with minimal overlap, the offer is categorized as ”not found”. We exclusively utilized the ”Gold Standard” for the training of our product matching method. The gold standard, derived from the English product data corpus, comprises a set of 1,100 pairs of offers from each of the four product categories: Computers Accessories, Camera & Photo, Watches, and Shoes. For each product, the gold standard includes two matching pairs of offers (positives) and five or six non-matching pairs of

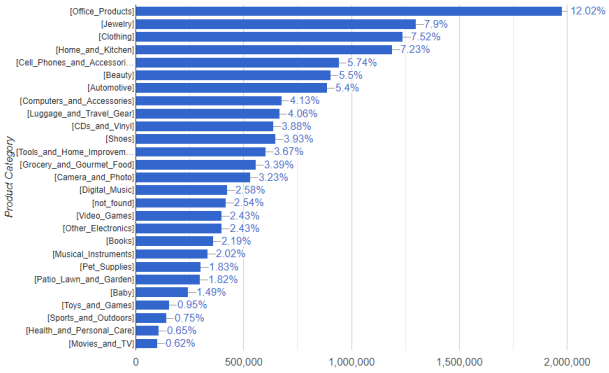

Figure 1: Distribution of offer entities per category in the English Training Set.

offers (negatives).

Additionally, we wanted to test or solution on Polish dataset for product matching created by the authors of (Michał Mozdzonek, 2022). The dataset was derived from popular Polish stores, involving data collection, subsequent cleaning, and transformation into a tabular format compatible with the WDC dataset. However due to the computational and time limitation we were unable to utilize this dataset.

3.2 Related Methodology

Our solution is based on the pipeline presented in (Tracz et al., 2020). In this article, the authors propose the usage of a transformer architecture, specifically a fine-tuned version of BERT, to embed and then compare a pair of products. As the pre-embedding product representation, a concatenation of the title, attribute values, and units was used. For the fine-tuning process, they used a triplet loss objective with the cosine distance. Additionally, various batch construction strategies for selecting the triplets used in training were analysed.

A similar methodology has been described in (Peeters et al., 2020), where a cross-encoder structure was used instead of a bi-encoder. Additionally, textual representation has been replaced by a concatenation of product brand, title, truncated description and truncated specification table.

An alternative approach was presented in (Peeters and Bizer, 2023), where the similarity score was generated with an LLM under proper prompt construction, however, due to performance limitations present in our problem this solution was disregarded.

Authors of the (Michał Mozdzonek, 2022) article focused not only on the Product matching problem but also on the idea of using transfer learning for data in different languages. Data preparation for the model involved selecting the title column and concatenating it with token markers. This resulted in a single input string for the model, which was then tokenized using a pre-trained model-specific tokenizer.

For the experimental part of the work, by using HuggingFace Transformers library, two types of pre-trained models were used: mBERT and XLM-RoBERTa. The models were pre-trained on Wikipedia articles in about 100 languages. The authors run the models on both WDC and Polish datasets. The F1 score was used as a metric to compare the models.

The mBERT (Devlin et al., 2019) and XLM-RoBERT (Conneau et al., 2020) models consistently outperformed other models. Notably, mBERT excelled, particularly in smaller-sized datasets (small, medium), a trend also observed in the Polish dataset. In the ”Shoes” category, results were slightly lower, with the mBERT model. However. XLM-RoBERT performed exceptionally well in ”large” datasets.

These results demonstrate that multilingual models effectively address the product matching problem, often yielding comparable or superior results to prior studies.

3.3 Choosen Approach

Our main aim was to test a method for product similarity matching using reliable language models. An important part of the project is the limitation of the execution time, which is set to 10 ms. Because of that, in our approach we decided to test out BERT (Bidirectional Encoder Representations from Transformers), DistilBERT (Sanh et al., 2019) as well as RoBERTa (Michał Mozdzonek, 2022). Despite having 40% fewer parameters than the original model, DistilBERT achieves competitive results performance with the benefit of running 60% faster. RoBERTa improves upon BERT by removing the next sentence prediction objective, utilizing dynamic masking during pre-training, and employing larger mini-batches, leading to enhanced performance in various natural language processing tasks. Additionally, as it is not stated otherwise in the project description, we assumed that the imposed time limitation doesn't include feature extraction. Therefore we were able

to utilize Large Language Models (LLMs) as feature extractor. This was important for us because LLMs demonstrated remarkable capabilities in understanding context, capturing complex patterns, and nuances in language.

In the first step we designed a method to evaluate solution that we were working on. We wanted this metric to be able to compare the similarity between many products at the same time and output single number which would measure how good the model is in ranking products on multiple levels. The input vector for the metric consists of cosine similarities between embeddings of products pairs from the transformer models. The resulting metric is a sum of Kendall Tau Distance (KDT) (Cicirello, 2019) and Mean Square Error (MSE). The former is responsible for keeping similarities of products in order and the latter forces similarities to be equaly spaced.

KDT measures the dissimilarity between two rankings by counting the number of pairwise disagreements in the ordering of elements. It considers the number of concordant and discordant pairs (pairs of elements that are in the same or opposite order in the two rankings) to quantify the similarity or dissimilarity between the rankings. In our case, we ranked cosinus similarities of product pairs.

MSE, in our case, measures how much vector of ordered cosine similarities between products pairs are different from the vector of values from 1 to 0 with equal difference between them.

For example: lets say that a, b, c, d are embeddings of products from transformer model and we know that product b is the most similar to product a and product d is the least similar to product a. Therefore we want the cosine similarity to be in the same order:

$$cos(a, a) > cos(a, b) > cos(a, c) > cos(a, d)$$

This is assured by KDT. However using only that, vector of cosine similarities like:

$$[1, 0.99, 0.11, 0.10]$$

would result in perfect score because the values are in order. MSE combats this by forcing a differences of subsequent values to be equal and close as much as possible to:

$$[1, 0.66, 0.33, 0]$$

Next step after creating a metric suitable for evaluation was transforming original dataset to meet our needs. Namely, we extracted representation of each product using LLM and based on positive pairs, negative pairs and other information about them, we created a evaluation data set in which each observation is list of products arranged in order of similarity.

After many iteration we came up with the following prompt: *Given a product title and description, generate a meaningful text representation that captures the essence of the product for effective similarity search. Consider relevant features, attributes, and contextual information to ensure the generated representation reflects the product's unique characteristics, allowing for accurate comparisons in a similarity search algorithm. Do not answer, just create a representation. TEXT TO REPRESENT:* $\{$*title+description*$\}$. Using it we can generated augumented text representations of products that will be used as input for transformer models.

Each record in evaluation dataset consists of representations of 5 products ordered like this $[anchor, anchor', positive, negative, category]$, where:

1. *anchor* is the main product representation that is compared to other products in a record (equivalent of product a in earlier example).
2. *anchor'* is representation of the same product as *anchor* but generated independently from the first one resulting in a different string for the same product
3. *positive* is a text representation of the product from the positive pair from original dataset.
4. *negative* is a text representation of the product from the negative pair from original dataset.
5. *category* is a text representation of a product from different category than original product.

Subsequent products are less and less similar to the $anchor$ thanks to which the created dataset is suitable to be used to evaluate the performance of models using the previously described metric.

To train the BERT-like models we utilized Triple Loss that is expressed like this:

$$L(a, b, c) = \max(||a - b||_2 - ||a - c||_2 + \alpha, 0)$$

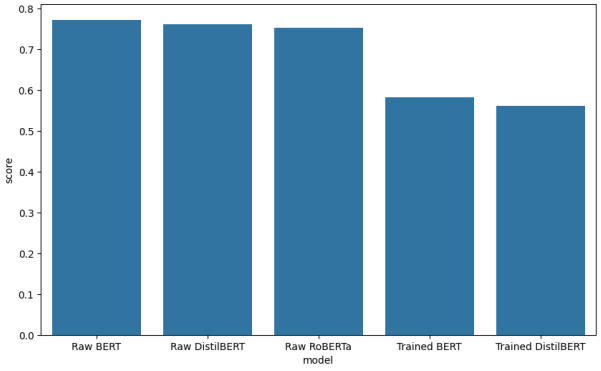

Figure 2: Metric score for pre-trainen models before and after our training

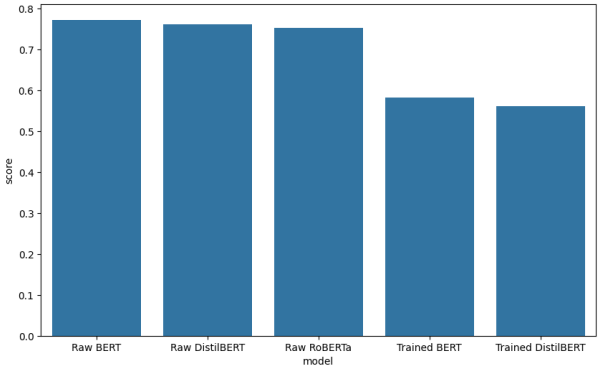

Figure 3: Inference time for choosen models

, where a is embedding of an anchor, b is embedding of product from positive pair and c is an embedding of a product from negative pair. Hyperarameter α is minimal sufficient difference of distances between them.

4 Experiments and Results

In our experiments we took into account performance of pre-trained BERT, DistilBERT and RoBERTa models before and after performing our training procedure. We evaluated models using our custom metric described earlier, as well as execution time for inference of each model.

5 Discussion on your results

6 Conclusions and future work

Works okay, needs refinement and further experiments.

In future developments of our ecommerce recommendation systems project, there are several key areas to explore. Firstly, expanding our methodology to include a broader range of datasets is crucial for assessing the generalizability of our product similarity matching system across different domains. Testing the solution on a bigger and cleaner dataset could provide the necessary amount data for the models, as scaling typically improves the results of neural networks.. Additionally, testing our recommendation system on advanced Large Language Models (LLMs) beyond the current state-of-the-art models like GPT-3 or BERT will ensure adaptability to evolving NLP technologies. Experimenting with a variety of prompts for feature extraction and evaluating different system prompts during the recommendation process will contribute to optimizing the system's performance. Furthermore, conducting comparative analyses without the reliance on Large Language Models will provide insights into the specific impact of these models on the recommendation system. Exploring alternative methods for description extraction, including domain-specific techniques or specialized models, is essential for diversifying the approach. Finally, testing the performance and success of the recommendation system in a real ecommerce environment, potentially through collaboration with industry partners or deployment in controlled settings, will validate its practical applicability and reveal opportunities for refinement and enhancement. Continuing research in these areas will contribute to the ongoing evolution of ecommerce recommendation systems, ensuring improved user experiences and advancements in the field.

References

[Bobadilla et al.2013] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems*, 46:109–132.

[Brown et al.2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Cicirello2019] Vincent A Cicirello. 2019. Kendall tau sequence distance: Extending kendall tau from ranks to sequences. *arXiv preprint arXiv:1905.02752*.

[Conneau et al.2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave,

Task (Total time est.)	Main Contributor	Secondary Contributor
Project Proposal (x h)	Everyone	-
Exploratory Data Analysis (x h)	*Kinga Ułasik*	*Szymon Rećko*
Proof Of Concept ()	*Mateusz Sperkowski*	*Patryk Tomaszewski*
First Milestone Presentation & Raport ()	Everyone	-
Reviews ()	Everyone	-
Generating the additional descriptions ()	*Szymon Rećko*	*Mateusz Sperkowski*
Tuning the models ()	*Patryk Tomaszewski*	*Kinga Ułasik*
Second Milestone Presentation & Raport ()	Everyone	-

Table 1: Work contribution and estimated total time for each task. While two main contributors are listed, all tasks were done with the support and help of everyone in the team when necessary.

Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

[Devlin et al.2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

[Gou et al.2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

[Kasneci et al.2023] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

[Koch et al.2015] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

[Liu et al.2018] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. 2018. Rethinking the value of network pruning. In *International Conference on Learning Representations*.

[Michał Mozdzonek2022] Sergiy Tkachuk Szymon Łukasik Michał Mozdzonek, Anna Wróblewska. 2022. Multilangual transformers for product matching – experiments and a new benchmark in polish.

[Peeters and Bizer2023] Ralph Peeters and Christian Bizer. 2023. Using chatgpt for entity matching. *arXiv preprint arXiv:2305.03423*.

[Peeters et al.2020] Ralph Peeters, Christian Bizer, and Goran Glavaš. 2020. Intermediate training of bert for product matching. *small*, 745(722):2–112.

[Polino et al.2018] Antonio Polino, Razvan Pascanu, and Dan-Adrian Alistarh. 2018. Model compression via distillation and quantization. In *6th International Conference on Learning Representations*.

[Sanh et al.2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

[Scarselli et al.2008] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

[Thirunavukarasu et al.2023] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

[Tracz et al.2020] Janusz Tracz, Piotr Iwo Wójcik, Kalina Jasinska-Kobus, Riccardo Belluzzo, Robert Mroczkowski, and Ireneusz Gawlik. 2020. Bert-based similarity learning for product matching. In *Proceedings of Workshop on Natural Language Processing in E-Commerce*, pages 66–75.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[Wei et al.2022] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

[Wu et al.2023] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

[Zhang et al.2019] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38.