Tc-llama 2: fine-tuning LLM for technology and commercialization applications

Jeyoon Yeom[1†], Hakyung Lee[1†], Hoyoon Byun[2], Yewon Kim[2], Jeongeun Byun[3], Yunjeong Choi[3], Sungjin Kim[3*] and Kyungwoo Song[1*]

[†]Jeyoon Yeom and Hakyung Lee have contributed equally to this work.

*Correspondence:
sungjin.kim@kisti.re.kr;
kyungwoo.song@gmail.com

[1] Department of Statistics and Data Science, Yonsei University, 50, Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
[2] Department of Artificial Intelligence, University of Seoul, 163, Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Republic of Korea
[3] Technology Commercialization Research Center, KISTI, 66 Hoegiro, Dongdaemun-gu, Seoul 02456, Republic of Korea

Abstract

This paper introduces TC-Llama 2, a novel application of large language models (LLMs) in the technology-commercialization field. Traditional methods in this field, reliant on statistical learning and expert knowledge, often face challenges in processing the complex and diverse nature of technology-commercialization data. TC-Llama 2 addresses these limitations by utilizing the advanced generalization capabilities of LLMs, specifically adapting them to this intricate domain. Our model, based on the open-source LLM framework, Llama 2, is customized through instruction tuning using bilingual Korean-English datasets. Our approach involves transforming technology-commercialization data into formats compatible with LLMs, enabling the model to learn detailed technological knowledge and product hierarchies effectively. We introduce a unique model evaluation strategy, leveraging new matching and generation tasks to verify the alignment of the technology-commercialization relationship in TC-Llama 2. Our results, derived from refining task-specific instructions for inference, provide valuable insights into customizing language models for specific sectors, potentially leading to new applications in technology categorization, utilization, and predictive product development.

Keywords: Large language models, Transfer learning, In-context learning

Introduction

The field of artificial intelligence (AI) has undergone a paradigm shift with the advent of large language models (LLMs) [1]. These models leverage extensive text data and self-supervised learning techniques to train on a vast scale. Furthermore, fine-tuning these models for specific tasks has led to impressive performance across various domains [2]. Beyond mere classification and prediction, LLMs offer deeper insights, demonstrating an understanding of intricate relationships within text expressions.

 The importance of technology and product relationship analysis lies in its ability to guide businesses and innovators in decision-making [3]. By understanding how a technology can be transformed into a viable product, organizations can make informed investments, develop strategies for market entry, and anticipate future market needs. In the realm of technology and product relationship analysis, understanding the intricate connections between technological advancements and their potential market

applications is crucial. Two traditional methods have been prominent: statistical learning [4, 5] and expert knowledge. Statistical learning methods benefit from automation but struggle with processing unstructured or textual datasets. Moreover, these methods typically require well-defined and labeled datasets to function effectively. Expert-based approaches are often more accurate than statistical approaches. Because they can consider nuanced and contextual factors that statistical methods may overlook. However, reliance on expert knowledge can be costly and is inherently limited by the availability of such expertise. Additionally, expert analysis can be subjective and may not always scale well for large datasets or rapidly evolving technology sectors.

Addressing these limitations, we introduce "TC-Llama 2", a model that utilizes the power of LLMs for understanding technology-commercialization relationships. The rationale behind using LLMs lies in their robust generalization capabilities, effective even with limited data. Technology-commercialization data is not only textual and heterogeneous but also scarce compared to general datasets, making large-scale, generalizable models ideal.

Our training approach involves transforming diverse technology-commercialization data into formats compatible with LLMs. This data, rich in variants, directs the model to assimilate technology knowledge and product hierarchies. We utilized Meta's open-source LLM, Llama 2 [6], enhanced with Lora [7] for efficient learning with minimal parameter updates. Further, we employed instruction tuning [8] to foster effective learning and inference. This entailed fine-tuning the pre-trained Llama 2 with a Korean-English instruction dataset, followed by training on the technology-commercialization data.

We evaluated our model's performance through two methods. Firstly, we assessed its specificity in matching product attributes to their descriptions, observing precise hierarchical embedding mapping and generalization to unlearned taxonomies. Secondly, we tested the model's depth in understanding technology-product relationships by generating product names from detailed technology descriptions. This not only validated the model's inferencing capabilities but also allowed us to explore the effects of prompt tuning on generative performance.

To our knowledge, TC-Llama 2 represents the first concrete commercial application in the technology-commercialization sector. This paper details effective training methods for non-standard textual data and introduces a novel model evaluation process. Additionally, our findings from prompt tuning offer insights for tailoring language models to specific applications. These contributions could pave the way for novel applications, such as identifying potential uses for underutilized technologies, automating technology categorization, and aiding in product development through technological foresight.

In brief, we can summarize the main contributions as follows:

1. We present "TC-Llama 2", a model utilizing LLMs for analyzing technology-commercialization relationships. This model uniquely address challenges posed by scarcity and complexity of technology-commercialization data.
2. Our research outlines an innovative training method for TC-Llama 2, employing Meta's Llama 2 model enhanced with Lora and instruction tuning using a Korean-English dataset. This approach effectively adapts the model to assimilate diverse technology and product information.

Yeom *et al. Journal of Big Data*   (2024) 11:100

Page 3 of 31

3. We introduce a novel evaluation process for TC-Llama 2, focusing on matching product attributes and generating product names from technology descriptions. Insights from prompt tuning experiments provide guidance for tailoring language models to specific sectoral applications.

Related works

Large language models

LLMs represent a major advancement in the field of natural language processing (NLP). Unlike earlier small scale language models(e.g. BERT [2], GPT [9]), LLMs comprise billions of parameters and undergo training on extensive text corpus. From the explosive increase in model size and comprehensive training data, recent LLMs have shown exceptional capabilities across diverse applications, often requiring minimal task-oriented training data. They have gained the ability to capture a vast array of linguistic nuances and contexts [1].

Since the increase in model size and dataset has been instrumental in the recent advances in LLMs, it is essential to optimize these factors within a particular computational budget to achieve best performance [10]. In this context, Llama 2 emerges as an open-source LLM that excels in performance, showcasing how strategic scaling can lead to significant improvements within computational constraints [6]. Llama 2 outperforms other language models and also shows competetive performance compared to best existing LLMs. In our study, we utilized Llama 2 as our pretained large language model.

Fine-tuning large language models

It has become a prevalent approach in NLP to fine-tune pre-trained models on downstream tasks rather than developing new models from scratch [11]. The complex training objectives and extensive parameters of these pre-trained models enable them to assimilate knowledge from massive data. This knowledge, encoded in their large-scale parameters, is fine-tuned for specific tasks, thereby enhancing its effectiveness across various applications [12].

Fine-tuning LLMs using diverse multi-task datasets formatted as text prompts, a process referred to as instruction tuning [13], has proven effective for tackling tasks that the models have not previously encountered. This approach of instruction tuning equips LLMs to handle new, unseen tasks as well as significantly improves their generalization ability [14]. However, given the substantial number of parameters in LLMs, full-parameter tuning can be prohibitively expensive. To mitigate this, recent research has shifted towards more parameter-efficient fine-tuning methods. Methods such as adapter tuning [15], which involves adding small trainable modules, and prompt tuning [16], where the input sequence is strategically altered, have emerged as effective solutions. Additionally, low-rank adaptation [7] offers a way to fine-tune LLMs by adjusting only a select subset of parameters, providing a balance between efficiency and performance. These approaches have contributed to adapting LLMs for specific tasks in a cost-effective manner.

Recently, the openness and effectiveness of Llama have garnered significant interest within the research community. Numerous studies have focused on fine-tuning and further pre-training its various model versions, contributing to the development

of new models(e.g., Alpaca [6, 17], Chinese Llama [18]) Llama, predominantly trained on English language datasets, faces limitations in its effectiveness for non-English languages [18]. In this work, we suggest a new model "GI-Llama 2" to enhance Llama's proficiency in understanding and generating Korean text, as well as its capacity to interpret instructions. Initially, we fine-tuned the Llama-2-chat-7B model with general instruction datasets that consists of both English and Korean data. Following this, we further specialize "GI-Llama 2" by fine-tuning it with a diverse set of technology-commercialization data. We propose our new model "TC-Llama 2", a model specifically tailored for technology-commercialization contexts. This two-step fine-tuning process significantly improved the model's understanding on Korean data and technology-commercialization data.

In-context learning

LLMs such as GPT-3 [19], GPT-4 [20], and Llama 2 [6] signify a major advancement in AI and NLP, pre-trained on extensive text datasets to generate text that is coherent and contextually aligned with input prompts. This leads to the emergence of in-context learning [21], a paradigm shift where the model exhibits the capability to comprehend and respond to tasks in accordance with the provided context. This method contrasts with earlier approaches requiring task-specific training for each new application. The strength of in-context learning is its versatility and adaptability, which not only increases the performance gains of separate retraining or fine-tuning, but also enables a single model to perform diverse tasks, from language translation and creative writing to complex coding.

To fully leverage the capabilities inherent in in-context learning, there are various instruction tuning techniques, starting with the methodology used from the training phase to the inference phase. For example, FLAN [13] allows instructing multiple types of tasks within a single directive and trains the model to perform these tasks concurrently, setting a foundation for versatile application. Chain of Thought [22] method improves performance by implementing a step-by-step process in inference, illustrating how structured reasoning can lead to better results. Additionally, [23, 24] where generating and applying specific instructions through LLM significantly increases performance, exemplifies the effectiveness of targeted guidance. This fine-tuning process involves incorporating examples for episodic learning and structuring prompts in a specific manner, which, along with embedding clear instructions, has been instrumental in elevating the accuracy and relevance of the model's responses to particular tasks. Meta ICL [25] further enhances this by showing improved performance with specific formatting and examples. Overall, these methods have significantly enhanced our control over the model's behavior, ensuring predictability and reliability—attributes crucial in applications requiring precise and trustworthy outputs. In this research, we crafted the task format for the fine-tuning by referencing prior studies. Additionally, by implementing automated prompt revision during the inference, we successfully identified an optimized format specific to the technology-commercialization task area. It consists of various in-context learning techniques.

Yeom *et al. Journal of Big Data*   (2024) 11:100

Page 5 of 31

Category mapping

Web retail services are increasingly requiring efficient product categorization to manage the growing volume of online products. By automating the categorization process, deep learning based frameworks can enhance search efficiency, enabling users to find desired products more quickly. [26] Additionally, the ability to reclassify products from different e-commerce platforms into a unified searching system (e.g., Google, Naver, and Tencent) using deep learning reduces manual labor and associated costs, making it a highly efficient and cost-effective solution. This categorized data can be further applied to recommendation systems to identify patterns and preferences of users, thereby recommending products or targeting advertisements more aligned with user's interests.

Product features such as product images [27, 28] and text descriptions [29, 30] can be leveraged for product categorization tasks using deep learning frameworks. In studies focusing on image features, the main setting being explored is offline retail store environment. These studies typically involve a two-phase classification process: initially detecting one or more products in the input image, followed by categorizing the detected objects into predefined categories [27]. Moreover, regarding text-based details, there is a vast quantity of information accessible for various products, including titles and features or instructions on how to use them, as well as reviews from users who have previously purchased these products. In this study, in contrast to prior research utilizing language models for product category classification [26, 31], we utilize a generative large language model based on the Transformer's decoder structure to map products to their corresponding categories. Additionally, unlike previous studies [26, 31] that primarily focused on categorizing products into categories and subcategories, we extend this approach by performing category classification tasks at each level within a more sophisticated, five-level category system. We also perform the task of mapping products from different platforms into a unified category hierarchy. As mentioned above, classifying various e-commerce products into a unified category system is a very important task for search engine services to provide users with accurate product searches and price comparison. To the best of our knowledge, we are the first to exploit LLMs to categorize diverse e-commerce products into a unified category system (e.g., Korean UNSPSC).

TC-Llama 2

Llama 2 [6]represents a significant achievement in the field of NLP, offering unparalleled capabilities for generating and understanding human-like text. Llama-2-chat-7B, a variant of the Llama family known for its flexibility and performance in diverse applications, has been further enhanced by fine-tuning Llama 2 using Reinforcement Learning From Human Feedback with one million human annotations, making it a popular choice for customization in specialized domains. The open-source nature of these LLMs greatly facilitates customization, allowing for rapid adaptation to specific tasks and languages, which is essential in today's fast-paced technological landscape [32, 33].
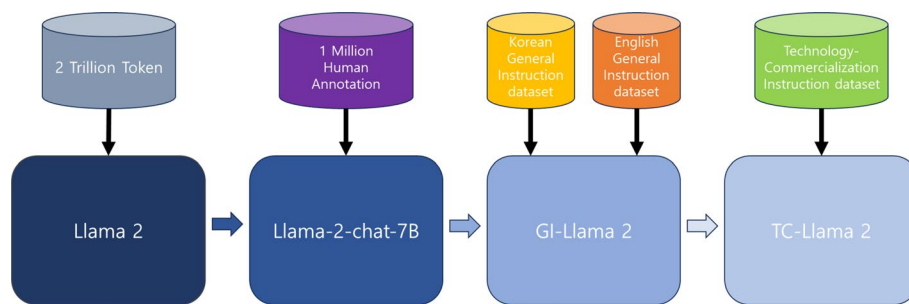
Fig. 1 Training Sequence Overview for TC-Llama 2: This diagram illustrates the multi-stage training process of TC-Llama 2. Initially, Llama 2 undergoes self-supervised learning with 2 trillion tokens. Subsequently, Llama2-chat-7B is fine-tuned using Reinforcement Learning, leveraging 1 million human annotations. Our methodology incorporates the publicly available Llama-2-chat-7B and a Korean-English General Instruction dataset, which we have compiled. The final stage involves utilizing the preprocessed Technology-Common Instruction dataset, specifically customized for technology-commercialization tasks, to implement TC-Llama 2

Table 1 List of text corpora and pre-processed data for TC-Llama 2

Dataset	Size(K)	Language	Domain
OIG-small-chip2	210	English	GI
OIG-small-chip2-kor	210	Korean	GI
Databricks-dolly-15k-kor	15	Korean	GI
Amazon	14	English	TC
UNSPSC	20	Korean/English	TC
Korean UNSPSC	122	Korean/English	TC
Danawa	29	Korean	TC
WIPSON	158	Korean	TC
KIPRIS	45	Korean	TC
NTIS	4	Korean	TC
Total	828		GI + TC

As depicted in Fig. 1, this study enhances TC-Llama 2, itself an advancement of Llama-2-chat-7B, by incorporating capabilities to generate technology business ideas and categorize products, building upon this foundational framework. We employed the Low-Rank Adaptation (LoRA) [7] approach, targeting both Korean and English datasets and documents related to technical commercialization, aiming to increase its efficiency in this specific domain. Our training strategy involves two distinct datasets: the General Instruction(GI) dataset and a specialized dataset for technical commercialization. The details in Table 1 provide basic statistics for each dataset. To enhance our model's bilingual capabilities, we effectively conducted instruction tuning on these Korean-English datasets [34]. This approach involves standardizing the data, restructuring sentences, and managing stop words, following the Alpaca [35] instruction tuning format, which comprises 'instruction' for the task, 'input' with necessary information, and 'output' as the task's solution. To optimize training efficiency, we used LoRA,

Yeom *et al. Journal of Big Data*   (2024) 11:100

Page 7 of 31

focusing on low-dimensional weight changes, thus reducing the parameter count for more effective model adjustment. Alongside LoRA, DeepSpeed [36] was incorporated for its distributed learning, memory optimization, and mixed precision capabilities, utilizing multi-GPU settings and FP16 precision. The training starts with the GI dataset, then progressed to the specialized technical commercialization dataset. This sequential approach refined the model for its final form, TC-Llama 2, which was inspired by Llama-2–7b-chat's training templates but adjusted input/output lengths to 2048 tokens for the Korean dataset, recognizing its higher token requirement compared to English.

Fine-tuning datasets for GI-Llama 2

The fine-tuning process begins with the General Instruction (GI) dataset, primarily aimed at augmenting the model's proficiency in the Korean language. The GI dataset, encompassing 210,000 entries, integrates diverse corpora. Notably, it includes a subset from the OIG dataset, a significant English corpus with around 43 million words of instructional content. This subset, named 'OIG-small-chip2' (https://laion.ai/blog/oig-dataset/), is specifically curated for high-quality instructional data. Furthermore, within the GI dataset, the 'OIG-small-Chip2-kor' (https://huggingface.co/datasets/heegyu/OIG-small-chip2-ko) represents a Korean translation of the OIG subset, accomplished using Google Translator. Another constituent dataset is 'databricks-dolly-15k-kor' (https://github.com/nlpai-lab/KULLM), translated into Korean via DeepL. These included datasets, comprising human-generated and Wikipedia content, cover a range of instructional tasks such as brainstorming, classification, both closed and open question answering, generation, information extraction, and summarization.

Fine-tuning datasets for TC-Llama 2

After fine-tuning Llama-2-chat-7B and producing GI-Llama 2, we further fine-tune our model on specialized instructional dataset for technology-commercialization tasks. Through this fine-tuning process, we aim our model to understand complex details and relationships related to technologies and products. The raw dataset includes the Amazon(https://www.amazon.com/) and Danawa(https://www.danawa.com/) dataset, which aggregates diverse product information from online shopping platforms. We also used the UNSPSC(https://www.unspsc.org/) dataset, an international product labeling system, along with its Korean counterpart(Korean UNSPSC). Unlike prior research [37] that directly used patent-product association data, we incorporated Korean patent information datasets, WIPSON(https://www.wipson.com/service/mai/main.wips), KIPRIS(http://www.kipris.or.kr/khome/main.jsp) from 2020 to 2023, focusing solely on patents without direct product associations. Also, we included NTIS(https://www.ntis.go.kr/ThMain.do) dataset, which is based on the National Science & Technology Information Service's data, including company and technology commercialization

Table 2 List of instruction datasets for TC-Llama 2

Task*	Size	Dataset
3.2.1	2550	NTIS
3.2.2	1762	NTIS
3.2.3	55,900	Amazon, Danawa, Korean UNSPSC, UNSPSC
3.2.4	21,384	Amazon, Danawa, UNSPSC
3.2.5	6581	UNSPSC
3.2.6	85,904	Korean UNSPSC, UNSPSC, WIPSON
3.2.7	132,226	Amazon, Danawa, KIPRIS, Korean UNSPSC, WIPSON
3.2.8	85,110	Danawa, Korean UNSPSC, UNSPSC, WIPSON
Total	391,417	

 *Task refers to the corresponding section title

information that is focused on technology research and product and company information.

From the collected dataset, we pre-processed the data and designed various tasks, including prediction, generation, translation, and summarization tasks for fine-tuning. The diversity of task is instrumental in mitigating the issue of catastrophic forgetting in the model, ensuring robust and versatile learning capabilities [38, 39]. Table 2 shows the instruction tasks and datasets used for fine-tuning TC-Llama 2. Task selection was guided by a comprehensive review of instructional datasets and task generalization in NLP [13, 40–42], covering information extraction (Task 3.2.1, Task 3.2.2), title generation (Task 3.2.4), translation (Task 3.2.5), text completion (Task 3.2.6), summarization (Task 3.2.7), and explanation (Task 3.2.8). Furthermore, in consideration of our model's focus on technology commercialization, we have developed Tasks 3.2.3, 3.2.4, and 3.2.5 to capture category information, aligning with conventional NLP task structures. Our tasks incorporates insights from hierarchical product classification, a concept emphasized in recent research [43, 44], which is fundamental for accurately understanding and navigating category structures for technology commercialization data. By incorporating this diversity of tasks, we aim to ensure that our fine-tuned model is not only robust but also highly relevant and effective for practical applications in technology commercialization.

Predicting information from commercialized companies

This pre-processed data format focuses on gathering and learning from data about companies that have successfully commercialized certain technologies. This specific information is not typically included in pre-trained language models. The data includes various details like the company's name, its industry classification, and a list of its main products. The training involves a predictive task where the model learns to infer one type of information (like the list of products) from the others (such as the company name and classification). This method is designed to improve the model's knowledge about how companies translate technologies into commercially viable products.

NTIS Task Example 1

Instructions
주어진 기업 상품을 보고 해당 기업의 업종을 예측하시오.
한글업체명: (주)제닉
한글주요제품: 화장품(하이드로겔 마스크팩,기능성화장품,의약품패치) 제조,도소매
. . . .
Given a company's product, predict the company's industry.
Company Name: (주)제닉
Main Products: Cosmetics (hydrogel mask packs, functional cosmetics, medical patches) manufacturing, wholesale and retail

---

Label
10차산업코드명: 화장품 제조업
. . . .
Industry code name: Cosmetics manufacturing

* Korean data is accompanied by English translations.

Predicting commercialization from technology

This data format uses a method known as the "chain of thought" to predict which technologies are likely to be commercialized. The model first breaks down the description of a technology into smaller categories, including technical aspects, key terms, and potential applications. Then, it uses this detailed breakdown to predict the likelihood of the technology being commercialized and to identify potential companies related to it. This approach helps the model understand the intricate connections between specific technologies and their paths to market success.

NTIS Task Example 2

Instructions
다음 기술의 기술세부분류에 대해서 답하고, 응용분야와 가능한 사업화 내용에 대해서 설명한 후, 다음 기술을 이용하여 어떠한 상품 혹은 서비스를 제공할 수 있는지 설명하시오.

기술명: 미생물로부터 산 촉매와 계면활성제를 이용한 유리지방산 고효율 추출
방법개요 및 특징: 본 발명은 미생물로부터 산 촉매와 계면활성제를 이용한 유리지방산 고효율 추출 방법에 관한 것으로 그 목적은 . . . 산 촉매와 계면활성제를 이용한 유리지방산고효율 추출 방법을 발명의 특징으로 한다
. . . .
Answer the technology breakdown of the following technologies, describe their applications and possible commercialization, and explain what goods or services could be provided using these technologies.

Technology description: Highly efficient extraction of free fatty acids from microorganisms using acid catalysts and surfactants
Methods and Features: The present invention relates to a method for high-efficiency extraction of free fatty acids from microorganisms using an acid catalyst and a surfactant, the object of which is to ... A method for high-efficiency extraction of free fatty acids from microorganisms using an acid catalyst and a surfactant is characterized by the invention.

Label
기술세부분류: 에너지/자원
키워드: 바이오 연료,바이오 연료 생산,바이오매스 전처리 기술
응용분야: 에너지 및 자원/ 사업화명: 미세조류 배양 모니터링 장치 제작 및 판매사업내용:미세조류 배양 모니터링 장치 제작 및 판매
10차산업코드명: 환경 관련 엔지니어링 서비스업
한글업체명: (주)유앤유
한글주요제품: 환경컨설팅 및 관련 엔지니어링 서비스업,컴퓨터 프로그래밍 서비스업,소프트웨어,하드웨어개발및공급, 연구개발업, 학술연구용역
. . . .
Technology Category: Energy & Resources
Keywords: biofuels,biofuel production,biomass pretreatment technology
Applications: Energy and resources/ Business name: Manufacture and sale of microalgae culture monitoring devicesBusiness description: Manufacture and sale of microalgae culture monitoring devices
10th Industry Code: Environment-related engineering services
Company Name: (주)유앤유
Main Products: Environmental consulting and related engineering services, Computer programming services, Software, hardware development and supply, Research and development, Academic research services

* Korean data is accompanied by English translations.

Category prediction across category structures

The category prediction task is structured to enhance the understanding of hierarchical category systems. In this setup, the model is presented with one level of a particular data source's category hierarchy and is then required to generate the associated upper and lower categories. This task was specifically designed to deepen the model's comprehension of various data source category structures. For each source, we customized the task to predict categories, aiming to enhance the model's ability to navigate and understand diverse commercial classification systems. By encompassing diverse data sources, this approach broadens the model's applicability and proficiency in handling various commercial data structures.

UNSPSC Task Example 1

Instructions
The following is a taxonomy and four-level hierarchy coded as an eight-digit number. Predict the parent taxonomy of this taxonomy.

The taxonomy code is 42182401, and the taxonomy is Audiometers

---

Label
Segment: Medical Equipment and Accessories and Supplies
Family: Patient exam and monitoring products
Class: Hearing testing products

Category prediction from product features and descriptions

This category prediction task is designed to enhance understanding and connection between product features, descriptions, keywords, and their categories, particularly focusing on commercial products. Utilizing diverse data, this task is divided into two data formats. The first format generates a four-level category structure from given product details, and the second part involves predicting product names, which is a higher level category of the given specific product details. Both formats aim to enhance the model's understanding of product information and its ability to link products to their categories. When given a product description, the model's task is to identify the correct category. This strengthens its capacity to associate product characteristics with specific category structures, making the information more accessible and coherent.

Amazon Task Example 1

Instructions
Generate a four-level category for the product based on the following product feature descriptions.

accurate&reliable, clinically certified: Femometer Blood Pressure Monitor is clinical & chip certified. Clear & large display readings on digital LED screen for high-precision systolic, diastolic & pulse rate measurements. adjustable cuff with deetection: Arm sleeve adjusts to fit circumferences between 8.7' to 16.5'. Detection ensures cuff is wrapped correctly. , dual-user memory & double power charging: Store data from 2 users & 99 memory locations, including date and time stamp.

Label
level 1: Health & Household
level 2: Medical Supplies & Equipment
level 3: Health Monitors
level 4: Blood Pressure Monitors

```
┌─────────────────────────────────────────────────────────────────────────────┐
│ Product Information Task Example 1                                            │
│                                                                              │
│ Instructions                                                                 │
│ 다음은 상품의 세부품목, 제조업체명, 상품모델명에 대한 설명입니다. 이 상품이 해당되는 품명을 예측하시오. │
│ 상품의 세부품명은 드론 (영문 품명 : Drone)입니다. 제조업체명은 Dji이고 상품모델명은 (CN)Avata Pro-View │
│ Combo 입니다.                                                                 │
│                                                                              │
│ ............................................................................ │
│ The following is a description of the specific product name, manufacturer name, and model name. │
│ Predict the product name that corresponds to this information.               │
│ The specific product name is Drone (English product name: Drone). The manufacturer's name is Dji │
│ and the model name is (CN)Avata Pro-View Combo.                              │
│ ─────────────────────────────────────────────────────────────────────────── │
│ Label                                                                        │
│ 품명 : 무인비행기 (영문 품명 : Unmanned aerial vehicle) 에 해당합니다.          │
│ ............................................................................ │
│ This corresponds to the product name: Unmanned aerial vehicle.               │
└─────────────────────────────────────────────────────────────────────────────┘
```

* Korean data is accompanied by English translations.

Translating with relationship of Korean and English categories

The translation task is designed to translate between English and Korean, particularly focusing on category structures and their descriptions. This task is important considering our training data for TC-Llama 2 includes the UNSPSC dataset, an international commodity classification system presented in English, and Korean UNSPSC dataset, which is built upon the UNSPSC framework. The synergy between these datasets in our translation task is expected to yield substantial benefits. By translating between these systems, TC-Llama 2 is trained not only in language conversion but also in understanding and aligning different commercial classification structures. Through this task, we aim to significantly enhance TC-Llama 2's effectiveness in interpreting and structuring complex category information across languages.

```
┌─────────────────────────────────────────────────────────────────────────────┐
│ UNSPSC Task Example 2                                                         │
│                                                                              │
│ Instructions                                                                 │
│ Translate the following taxonomy and its detailed description into Korean.    │
│                                                                              │
│ Taxonomy : Live Plant and Animal Material and Accessories and Supplies        │
│                                                                              │
│ Descriptions : This segment includes live, wild and domesticated, seeds and plants and animals. This │
│ segment also includes materials and equipment used in the care, containment and feeding of live plants │
│ and animals.                                                                 │
│ ─────────────────────────────────────────────────────────────────────────── │
│ Label                                                                        │
│ 분류체계 : 산동식물및동식물성생산품                                            │
│                                                                              │
│ 상세설명 : 생명이 있는 야생과 재배성의 식물과 동물, 생명이 있는 식물들과 그들의 씨앗, 생명이 있는 식물과 동물의 │
│ 보호, 수용, 사육에 필요한 재료와 설비.                                         │
└─────────────────────────────────────────────────────────────────────────────┘
```

Sentence completion

For sentence completion task, diverse datasets are utilized to enhance descriptions of categories and inventions. This task aims to improve the model's overall language process skills and also make the model accurately process and convey specific informational contexts. Specifically, completing categorical descriptions enhances the

model's grasp of these categorizations, and completing invention descriptions aids in effectively linking invention related terms with technologies.

UNSPSC Task Example 3

Instructions
The following is a taxonomy and a partial description of the taxonomy. Read the following partial definition description of the taxonomy and provide a coherent completion that accurately defines the taxonomy term.

Taxonomy: Cleaning Equipment and Supplies
Definition : This segment includes the equipment and supplies used in environmental protection and wastewater management as well as

---

Label
in the cleaning of offices and lavatories. This segment also includes commercial laundry and dry cleaning equipment.

Text summarization task across diverse data sources

In this summarization task, we train the model to simplify complex information from different technology and commercial data sets. The goal is to teach the model to summarize detailed information into concise and clear terms, whether it's about patents, product classifications, inventions, or product features. This unified approach makes the model more adaptable in dealing with diverse informational contexts and steers the model to become specialized in conveying key aspects of technology and commercialization.

UNSPSC Task Example 4

Instructions
The following is a detailed definition of a taxonomy. Predict the corresponding taxonomy.

This segment includes live, wild and domesticated, seeds and plants and animals. This segment also includes materials and equipment used in the care, containment and feeding of live plants and animals.

Label
Live Plant and Animal Material and Accessories and Supplies

Understanding and generating descriptions

Text generation task is designed to improve TC-Llama 2's ability in accurately understanding and generating descriptions across different data sets. The task focuses on comprehending and producing detailed descriptions relevant to each data set's context, whether it involves inventions or categorical data.

By training TC-Llama 2 on this task, the model is expected to develop a more refined skill in interpreting diverse information. This enhancement is particularly beneficial for applications in fields of technology commercialization where understanding the nuances of different categorization systems and the ability to describe inventions and technologies accurately are paramount.

Yeom *et al. Journal of Big Data*   (2024) 11:100

Page 13 of 31

UNSPSC Task Example 5

Instructions
The following is a taxonomy. Describe the taxonomy in detail.

Live Plant and Animal Material and Accessories and Supplies

---

Label
This segment includes live, wild and domesticated, seeds and plants and animals. This segment also includes materials and equipment used in the care, containment and feeding of live plants and animals.

Evaluation methods

TC-Llama 2 is a fine-tuned LLM tailored for a broad range of technology-commercial applications. In order to validate its effectiveness as a technology-commercial language model, we have developed our own evaluation methods. We propose two methods for utilizing TC-Llama 2. The first method is using the output embedding of an input sentence for classification tasks. The second method is related to in-context learning, which is a prompt engineering technique designed to generate sentences that are well-suited for downstream tasks.

Embedding mapping with LLMs

We propose how to utilize output token embeddings to perform classification tasks exploiting generative language models. Its quick inference time comes from not using the model to generate output responses. In addition, because a zero-shot inference is allowed, it has high generality. As an example of a classification problem, in this work we used mapping a product to multi-level category. In a standard classification method, the scores for each candidate class are evaluated beforehand, and the predicted class with the highest score as a result by converting the scores to probabilities (e.g., softmax). These methods cannot account for classes not included in the candidate classes. On the other hand, with our proposed method based on generative models, classification is possible even when given new categories. This method works by representing both product information and target categories as embeddings, and then calculating the similarity between these embeddings for classification. Figure 2 illustrates the three methods
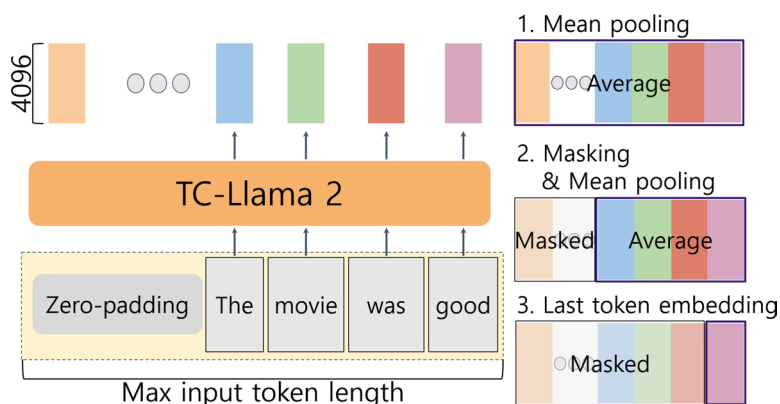
Fig. 2 Overview of sentence embedding extraction methods

by which a generative language model can compute sentence embeddings from input sentences. In contrast to traditional models like BERT [2], which utilize [CLS] tokens to obtain sentence embeddings, our model operates on a decoder-based Transformer architecture and consequently does not incorporate [CLS] tokens. This necessitates a distinct approach to derive comprehensive sentence embeddings. Therefore, we conducted experiments with mean pooling embedding [45–47], max pooling [46, 47], last token embedding [48, 49], based on prior research, to identify effective methods.

Sentence embedding of product and category description

In our approach, we utilize TC-Llama 2 to perform embedding-level similarity calculations for product category mapping. Product and category descriptions are fed into TC-Llama 2, which in turn produces output embeddings representing each product and category. These embeddings are produced for each token in the input sentence, containing contextually relevant information about the expected subsequent token. TC-Llama 2 is designed to handle a fixed maximum token length, which is a hyperparameter in our model. Sentences shorter than this maximum are padded with zeros, ensuring uniform input length for consistent embedding generation. This zero padding precedes the input sentence, aligning it with the model's required token length, thereby optimizing the embeddings for accurate product-category mapping.

$$e_i = f(t_i | t_0, ..., t_{i-1}) \tag{1}$$

$e \in \mathbb{R}^d$ denotes a output embedding of length d. f stands for TC-Llama 2.

For a given input sentence, we obtain an output embedding of TC-Llama 2 with a maximum token length. To obtain the sentence embeddings, we consider three different processing methods. The initial method is to calculate the average of all the output embeddings.

$$S_{all} = \frac{1}{N} \sum_{i=0}^{N} e_i \tag{2}$$

N is denotes the max input token length. This is the most intuitive and straightforward way to calculate the embedding that can represent a sentence from the output embeddings. However, when computing the average, the output embeddings for zero padding are also computed. Thus, if the input sentences are short, they may include a considerable amount of noisy, irrelevant information. To mitigate this issue, we considered masking the zero-paddings, excluding it from the embedding average calculation.

$$s_{masked} = \frac{1}{N - t + 1} \sum_{i=t}^{N} e_i \tag{3}$$

t is defined as the index of the first input token (i.e, the next to the last of zero-padding). The last method we considered involves utilizing the output embedding of the last token. This is because TC-Llama 2, to compute this output embedding for the last input token $e_N = f(t_N | t_0, ..., t_{N-1})$, needs to attend to all input sentences, thereby encapsulating the entire input sentence's information.

$$s_{last} = e_N$$

(4)

Embedding similarity for category mapping

We applied the similarity calculation of product description embeddings and category description embeddings to categorize products into 4 (UNSPSC has 5) levels of categories. First, we extract the embedding of each sentence representing the unique 4(5) level category combination for each data (Amazon, Danawa, and UNSPSC). Then, we extract the product embedding for each data by using the product name and description as sentences. We calculate cosine similarity between product and category embeddings to assign products to categories.

$$\hat{s} = s/|s|$$

(5)

$$\hat{S}_c = [\hat{s}_{c1}, \hat{s}_{c2}, ..., \hat{s}_{cM}]$$

(6)

$$similarity(\hat{s}_p, \hat{S}_c) = S_c^T \cdot \hat{s}_p$$

(7)

$s \in \mathbb{R}^d$ and $\hat{s} \in \mathbb{R}^d$ denote a sentence embedding and the normalized one of length d, respectively. $\hat{S}_c$ denotes the matrix composed of normalized category embeddings, totaling M. $\hat{s}_p$ denotes the normalized product embeddings.

Mapping products of various data to standard category

We propose a more advanced mapping method to categorize products from various data into a unified category system. In this study, we basically use UNSPSC's Korean data as a unified category system and also use UNSPSC's English data to improve accuracy. First, we extract the embeddings for the names and descriptions of the first-level categories (*segment* in UNSPSC) in the Korean and English UNSPSC data, respectively. Next, we compute the similarity of the product embedding to the first-level category embeddings to identify the top R first-level categories with the highest similarity. We then compute the similarity of the product embedding to the last-level category embeddings (*specific commodity* in UNSPSC) belonging to the top R first-level categories. This method effectively reduces the number of category embedding candidates for similarity computation, which helps increase accuracy.

$$\hat{S}_c^1 = [\hat{s}_{c1}^1, \hat{s}_{c2}^1, \ldots, \hat{s}_{cH}^1]$$

(8)

$$\bar{S}_c^1 = \{\bar{s}_{ci}^1 \in \hat{S}_c^1 : 1 \le i < j \le R, \text{ and } \hat{s}_p^T \cdot \bar{s}_{ci}^1 > \hat{s}_p^T \cdot \bar{s}_{cj}^1\}$$

(9)

$\hat{s}_c^1$ represents a normalized level 1 category embedding. The set $\bar{S}_c^1$ corresponds to the top R normalized level 1 category embeddings, determined based on their similarity with the normalized product embedding $\hat{s}_p^1$. The value of R is selected as one of the hyper-parameters.

Optimizing instruction using Iterative LLM-based methods

In our approach to enhancing product name generation for identifying technology commercialization relationships, we employed a method that integrates automatically generated instructions using LLMs. This strategy was motivated by the effectiveness of black-box optimization methods in recent research [50, 51], where LLMs have been preferred over human input. In Figure 3, the process begins with the creation of an initial instruction using a basic prompt and straightforward initial conditions. Once this step is completed, the instruction, along with the resulting output from TC-Llama 2, is passed to GPT-4 for a thorough evaluation. GPT-4's role is to analyze, critique, and subsequently refine the initial prompt. Based on GPT-4's analysis, a new and improved prompt is crafted. The revised prompt undergoes a series of iterative cycles, successively refining the instructions and generating improved outcomes. The process continues until it results in a version that exhibits minimal change, even with further iterations. Through this method, we aim to continuously enhance the quality and relevance of the generated product names in the context of technology commercialization. In "Inference of Technology-Commercialization Relationships" section, the final instruction encapsulates a variety of methods utilized in contemporary prompt tuning practices and adapt task-specific details that are specifically customized for the task at hand.

Experiments and results

Our performance validation consists of tasks that evaluate how well TC-Llama 2 generalizes to commercial applications after being fine-tuned on technical documents or commercially relevant information data using our instruction learning strategy. Using our two evaluation methods, we performed a total of four tasks, which can be broadly divided into evaluating the classification and generation performance of TC-Llama 2. First, Task 1 5.1.1, Task 2 5.1.2, and Task 3 5.1.3 evaluate the classification performance of TC-Llama 2 through product-category mapping tasks. Task 4 5.2 verifies the
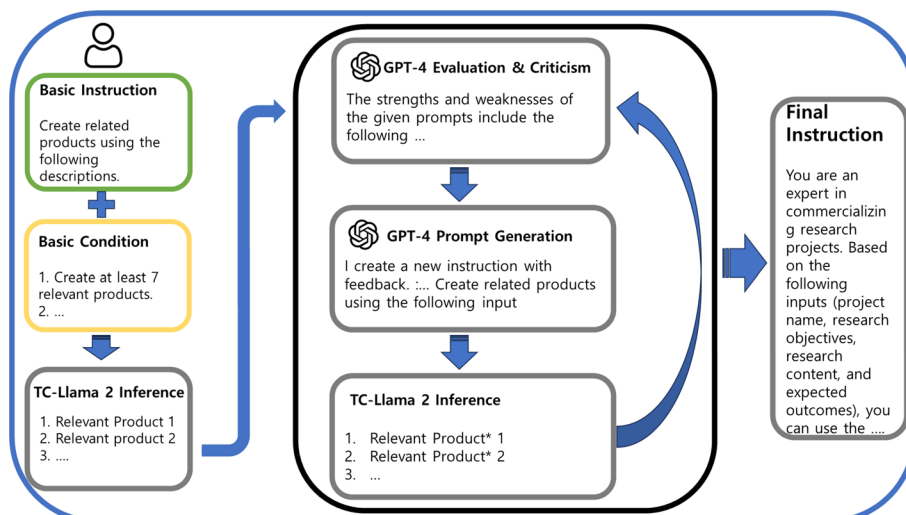
Fig. 3 Overview of generating instruction process. After obtaining the results of TC-Llama 2 through the initial instructions and conditions, the prompts and results are iteratively revised through GPT-4

alignment of the model with the technology-commercialization relationship through the generation of relevant products according to the technology description.

Product to category mapping

For TC-Llama 2, the category mapping task goes beyond commercialization. This task is central to understanding the interplay between product categorization and technology data, as the model has been trained with multi-tasking capabilities that encompass both commercialization and technology datasets. The effectiveness of category mapping is founded on its ability to synthesize detailed product attributes with expansive market insights. By precisely categorizing products, TC-Llama 2 is equipped to uncover patterns and trends within product innovation and technology integration, enabling a more informed and strategic approach to technology commercialization. For Task 1 and Task 2, we evaluated the performance of product and category mapping on two e-commerce platforms, Amazon and Danawa. In Task 3, we selected 50 products from each platform, all included in the UNSPSC, and undertook the task of aligning them with Korean UNSPSC categories. In Table 3, we reports the data statistics. We evaluated the classification performance using Recall and Mean Reciprocal Rank (MRR) metrics for the top K product-category similarities. Recall is the percentage of the top K categories that contain the correct category. MRR is the average of the reciprocal of the rank of the correct category in the top K categories. The higher the rank of the correct category, the closer the MRR is to 1. Both Recall and MRR have a range between 0 and 1. We set K to {10, 20, 30} to evaluate performance. For each task, we report on one of the three sentence embedding methods, where TC-Llama 2 performs well at the overall level.

We measured the mapping performance not only for the last-level category, but also for the ancestor categories. The method we used is to first perform the mapping of the last-level categories, followed by comparing the higher-level categories of the top-K similar last-level categories with the corresponding higher-level categories of the correct answers. The bold values in Tables 4 through 10 denote superior performance metrics for TC-Llama compared to the conventional Llama model, enabling easier comparison.

Task 1. Amazon product to category mapping
We evaluate the performance of classifying products into multi-level categories on the Amazon dataset. In Table 4, we compare the performance between TC-Llama 2 and Llama 2. Ours performs better than Llama 2 at lower levels. At higher levels, ours has a

Table 3 Data statistics at each task

Task	Product	(Last-level) category
Task 1	194	10,630
Task 2	1005	7032
Task 3	100	21,835

Table 4 Multi-level categories mapping on the Amazon dataset

Embedding	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
	1	Llama 2	0.5258	0.6134	**0.6753**	**0.4418**	**0.4418**	**0.4674**
	1	TC-Llama 2	**0.5979**	**0.6340**	0.6495	0.3060	0.3387	0.3387
	2	Llama 2	0.4124	0.5052	**0.5515**	0.2740	**0.2822**	**0.2822**
Masking &	2	TC-Llama 2	**0.4794**	**0.5103**	0.5206	**0.2764**	0.2764	0.2815
Mean pooling	3	Llama 2	0.3093	**0.3660**	0.4021	0.2565	0.2617	0.2750
	3	TC-Llama 2	**0.3351**	0.3660	0.4021	**0.2628**	**0.2744**	**0.2795**
	4	Llama 2	0.0515	0.0979	0.1443	0.0593	0.0845	0.0900
	4	TC-Llama 2	**0.1649**	**0.2216**	**0.2629**	**0.0900**	**0.0982**	**0.0982**

Table 5 Multi-level categories mapping on the Danawa dataset

Embedding	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
	1	Llama 2	0.0617	**0.4637**	**0.6687**	0.1133	0.3499	**0.8395**
	1	TC-Llama 2	**0.3652**	0.4458	0.5980	**0.1903**	**0.3525**	0.5531
	2	Llama 2	0.0030	0.0040	0.0070	0.0029	0.0029	0.0041
Last token	2	TC-Llama 2	**0.0100**	**0.0348**	**0.0517**	**0.0065**	**0.0277**	**0.0386**
embedding	3	Llama 2	0.0000	0.0000	0.0010	0.0000	0.0000	0.0033
	3	TC-Llama 2	**0.0010**	**0.0070**	**0.0159**	**0.0007**	**0.0155**	**0.0174**
	4	Llama 2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
	4	TC-Llama 2	0.0000	0.0000	**0.0030**	**0.0002**	**0.0018**	**0.0054**

slightly lower MRR, but recall is still relatively high. This implies that TC-Llama 2 is able to classify more accurately with more category information.

Task 2. Danawa product to category mapping
We evaluate the capability of TC-Llama 2 in classifying products into multi-level categories using the Danawa dataset. In Table 5, we compare the performance between TC-Llama 2 and Llama 2. In general, TC-Llama 2 surpassed Llama 2 in performance on the Danawa dataset, a Korean e-commerce dataset. This is due to our implementation of instruction learning on TC-Llama 2 with Korean data. However, despite this improvement, the overall absolute performance is still relatively low, indicating that more comprehensive training with Korean data is necessary. Additionally, we can argue that the difficulty of the task increased in Task 2, where the number of products to be evaluated was about five times higher than in Task 1, making it a more challenging task.

Task 3. Product to standard category mapping
We evaluate the performance of classifying products sampled from Amazon and Danawa data into multi-level categories on the Korean UNSPSC. In Table 6, we compare the performance between TC-Llama 2 and Llama 2. In general, TC-Llama 2 outperformed Llama 2. Even in the last-level category, Llama 2 fails to map a one product correctly, while TC-Llama 2 outperforms, slightly. Despite this improvement, the

Table 6 Multi-level Korean UNSPSC categories mapping on the two e-commerce dataset

Embedding	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
	1	Llama 2	0.24	0.63	**0.85**	0.0627	0.0886	0.0974
	1	TC-Llama 2	**0.29**	**0.66**	**0.85**	**0.0792**	**0.1045**	**0.1117**
	2	Llama 2	0.02	0.05	0.13	0.0058	0.0080	0.01135
Last token	2	TC-Llama 2	**0.09**	**0.11**	**0.14**	**0.0289**	**0.0305**	**0.0318**
embedding	3	Llama 2	0.01	0.03	0.03	0.0033	0.0047	0.0047
	3	TC-Llama 2	**0.05**	**0.05**	**0.05**	**0.0192**	**0.0192**	**0.0192**
	4	Llama 2	0.01	0.01	0.01	0.0033	0.0033	0.0033
	4	TC-Llama 2	**0.03**	**0.03**	**0.03**	**0.0158**	**0.0158**	**0.0158**
	5	Llama 2	0.00	0.00	0.00	0.00	0.00	0.00
	5	TC-Llama 2	**0.03**	**0.03**	**0.03**	**0.0158**	**0.0158**	**0.0158**

Table 7 Multi-level Korean UNSPSC categories mapping on the two e-commerce dataset with the advanced methods

Embedding	Method	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
Last token emedding	1	1	Llama 2	0.24	0.63	0.85	0.0627	0.0886	0.0974
Last token emedding	3	1	TC-Llama 2	**0.35**	**0.63**	**0.87**	**0.1302**	**0.1487**	**0.1580**
Last token emedding	1	2	Llama 2	0.02	0.05	0.13	0.0058	0.0080	0.0114
Last token emedding	3	2	TC-Llama 2	**0.12**	**0.16**	**0.19**	**0.0715**	**0.0738**	**0.0751**
Last token emedding	1	3	Llama 2	0.01	0.03	0.03	0.0033	0.0047	0.0047
Last token emedding	3	3	TC-Llama 2	**0.04**	**0.09**	**0.12**	**0.0047**	**0.0082**	**0.0094**
Last token emedding	1	4	Llama 2	0.01	0.01	0.01	0.0033	0.0033	0.0033
Last token emedding	3	4	TC-Llama 2	**0.04**	**0.04**	**0.04**	**0.0325**	**0.0325**	**0.0325**
Last token emedding	1	5	Llama 2	0	0	0	0	0	0
Last token emedding	3	5	TC-Llama 2	**0.04**	**0.04**	**0.04**	**0.0325**	**0.0325**	**0.0325**

overall performance remains comparatively low. This is likely due to the number of last-level categories being twice as many as in Task 1 and three times as many as in Task 3, resulting in poor classification accuracy.

Task 3 with advanced mapping method

We repeated Task 3 with our advanced mapping method introduced in section 4.1.3. Furthermore, we noted subpar performance on the Danawa data in Task 2, leading us to segment the performance analysis by dataset. We added a method column to each of Table 4, 5 and 6. Method 1 is the same method for calculating embedding similarity that we used in Tasks 1, 2, and 3. Method 2 utilizes the first-level category embeddings of Korean UNSPSC, and Method 3 utilizes the first-level category embeddings of English UNSPSC. In Table 7, the classification of Aamzon and Danawa products as a whole achieved the highest performance improvement across category levels in method 3. They outperformed Llama 2 in method 1, as well as TC-Llama 2 in method 1. As shown in table 8, when we evaluate solely on Amazon data, there is a performance improvement of about 1.5 to 2 times compared to the overall results. Finally, in Table 9, we evaluate only Danawa, which is Korean data. From the results, we can tell that the performance of the model is comparatively lower than the overall

Table 8 Multi-level Korean UNSPSC categories mapping on the Amazon products with the advanced methods

Embedding	Method	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
Last token embedding	1	1	Llama 2	0.36	0.66	0.88	0.0889	0.1094	0.1183
Last token embedding	3	1	TC-Llama 2	**0.54**	**0.76**	**0.92**	**0.2184**	**0.2319**	**0.2382**
Last token embedding	1	2	Llama 2	0.10	0.24	0.30	0.0292	0.03759	0.0401
Last token embedding	3	2	TC-Llama 2	**0.22**	**0.28**	**0.40**	**0.1087**	**0.1123**	**0.1170**
Last token embedding	1	3	Llama 2	0.02	0.06	0.06	0.0067	0.0094	0.0094
Last token embedding	3	3	TC-Llama 2	**0.12**	**0.12**	**0.16**	**0.0715**	**0.0715**	**0.0715**
Last token embedding	1	4	Llama 2	0.02	0.02	0.02	0.0067	0.0067	0.0067
Last token embedding	3	4	TC-Llama 2	**0.08**	**0.08**	**0.08**	**0.0650**	**0.0650**	**0.0650**
Last token embedding	1	5	Llama 2	0.00	0.00	0.00	0.00	0.00	0.00
Last token embedding	3	5	TC-Llama 2	**0.08**	**0.08**	**0.08**	**0.0650**	**0.0650**	**0.0650**

Table 9 Multi-level Korean UNSPSC categories mapping on the Danawa products with the advanced methods

Embedding	Method	Level	Model	R@10	R@20	R@30	MRR@10	MRR@20	MRR@30
Last token emedding	1	1	Llama 2	0.28	0.50	0.72	0.0589	0.0735	0.0817
Masking & Mean pooling	2	1	TC-Llama 2	**0.20**	**0.58**	**0.98**	**0.0702**	**0.0978**	**0.1139**
Last token emedding	1	2	Llama 2	0	0.04	0.04	0	0.0025	0.0025
Masking & Mean pooling	2	2	TC-Llama 2	**0.14**	**0.14**	**0.26**	**0.0457**	**0.0457**	**0.0502**
Masking & Mean pooling	1	3	Llama 2	0	0	0.02	0	0	0.0008
Masking & Mean pooling	2	3	TC-Llama 2	**0.14**	**0.14**	**0.18**	**0.0397**	**0.0397**	**0.0411**
Masking & Mean pooling	1	4	Llama 2	0	0	0.02	0	0	0.0007
Masking & Mean pooling	2	4	TC-Llama 2	**0.02**	**0.02**	**0.02**	**0.0027**	**0.0027**	**0.0027**
Masking & Mean pooling	1	5	Llama 2	0	0	0.02	0	0	0.0008
Masking & Mean pooling	2	5	TC-Llama 2	**0.02**	**0.02**	**0.02**	**0.0027**	**0.0029**	**0.0027**

performance. This indicates that there is still a need to improve the understanding of Korean language as compared to English. Additionally, we can see that method 2 outperforms method 3 on Danawa data. We speculate that this result is due to method 2's utilization of first-level category information from the Korean UNSPSC.

Inference of technology-commercialization relationships

In the realms of product development, marketing, and customer support, accurately inferring technology-product relationships is essential. Historically, this task has been managed through statistical learning and expert input. However, these traditional methods have inherent limitations: while statistical learning automates the process, it often struggles with the efficient processing of large-scale text data. Expert-based methods, on the other hand, offer accuracy but at a high cost and are hindered by the limited availability of experts.

The introduction of LLMs has marked a significant advancement in this area. LLMs, trained on extensive text corpora, have shown remarkable capabilities in accurately and efficiently inferring the intricate relationships between technologies and products. This leap forward is not just in terms of precision but also in processing efficiency.

Further leveraging the power of LLMs, the inference process has been refined to involve fine-tuning these models with specialized technology and product-related text data. This approach has not only facilitated the generation of innovative product ideas but also found diverse applications. It's instrumental in product development, aiding in both ideation and enhancement; in marketing, it allows for precise customer segmentation and strategy formulation; and in customer support, it enhances understanding and satisfaction.

To empirically validate the effectiveness of this LLM-based methodology, we embarked on a practical application: generating product names using NTIS technology project data. The process was further optimized using prompt tuning, a technique that refines the LLM's output without the need for additional extensive training. By employing instructions generated through ChatGPT, we were able to maximize the LLM's inherent understanding of technology-product relationships. This not only tested the model's efficacy but also demonstrated its potential to spawn viable business ideas and product concepts.

Task4 instruction

Instruction
당신은 연구과제를 사업화하는 전문가입니다. 다음 input(과제명, 연구목표, 연구내용, 기대효과)을 기반으로 가능한 사업 아이디어를 생성하시오. 사업 아이디어(business idea)는 연구과제를 기반으로 한 새로운 사업 아이디어를 설명합니다. 그리고 사업 아이디어에 따른, 구체적인 제품 7개를 생성하시오. 제품은 연구과제와 사업아이디어에 따른 구체적 제품명으로, 실제로 소비자에게 판매될 수 있는 형태여야 하며, 추론에 따라 합리적이며 다양해야합니다. 유의사항: 입력된 마지막 문장은 불완전할 수 있으므로, 제공된 정보로 최선의 답변을 해주세요. 답변은 모두 연구과제와 관련되어야 하며, 간결해야 합니다. 문장의 첫시작은 '제품명1 : ' 으로 시작하십시오.

....

You are an expert in commercializing research projects. Generate a possible business idea based on the following inputs: project name, research objectives, research content, and expected outcomes. The business idea describes a new business idea based on the research project. Create seven specific products based on the business idea. Products are specific product names based on your research project and business idea, should be in a form that can actually be sold to consumers, and should be reasonable and diverse based on your reasoning. Note: The last sentence typed may be incomplete, so please do your best with the information provided. All answers should be relevant to the research question and should be concise. Sentences should begin with 'Product name1:'.

* Korean data is accompanied by English translations.

Data preprocessing

In our instruction of technology-product generation task, we utilized the NTIS table as our primary dataset, which contains detailed project information. The key columns from this table, such as "Project Name, Research Objectives, Research Content, and Expected Effects," were specifically chosen to provide a comprehensive understanding of each technology project. The main objective of this task was to generate rational and diverse outputs without a definitive answer key for specific product names. For evaluation purposes,

we included temporary labels for commercialization names and content. A significant step in our methodology was the removal of recurrent stopwords (e.g., x000D) from the dataset using regular expressions to enhance data quality. We observed that inputs in a short-answer format were more effective than paragraph-style formats, leading us to adopt this approach for better performance. For evaluating the effectiveness of our technology-product generation task, we randomly sampled 50 technology project datasets from our chosen dataset. It's important to note that all data, including instructions, were kept within the maximum input length of 2048 tokens to fit the model's constraints. This comprehensive approach allowed us to efficiently and effectively generate outputs that offer practical insights into the relationship between technology and potential product development.

Instruction format

In study on inferring technology-product relationships, we utilized prompt tuning to enhance the performance of a language model without requiring further training. This process began with creating an initial prompt, which was then inputted into the model. The output was analyzed, and the prompt was iteratively modified to improve results. A significant step involved assigning the model the role of "an expert in commercializing research projects," which helped in generating more purposeful outputs. Additionally, specifying the number of product examples in the prompts led to higher quality results. We also acknowledged the token limitations of ChatGPT, incorporating cautionary notes to address potential content truncation. We discovered that prompts that initially focused on generating business ideas and subsequently on products were more effective, highlighting the intricate link between business strategy and product development.

Qualitative analysis

In this section, we present the qualitative analysis results for the generated technology-product outputs. We conducted a qualitative analysis of 50 outputs generated using prompts. The results were categorized into well-executed, average, and poorly executed cases, showing a distribution ratio of 3:1:1, respectively. For well-executed cases, the product names were described in a stable format, and the generated results were mostly related to the technology. In average cases, while the output was somewhat related and maintained a stable format, it occasionally included unintended content or lacked specificity. Poorly executed cases involved repetitive use of the same words, phrases, or clauses (such as repeating the same product name or instruction) or generating product names unrelated to the technology, indicating a misunderstanding of the instructions. During the error case analysis, it was observed that when input data was unclear or lacked detailed content (e.g., generic, basic information, or insufficient description of the technology), the quality of the output correspondingly decreased, often leading to less accurate or relevant data. Although these outputs sometimes generated names similar to commercialized products, there was a noticeable difficulty in generating specific product names. In conclusion, our methodology enabled the generation of diverse and applicable business and product ideas across various industries and technologies, surpassing the capabilities observed in simple query-based generation methods. This study demonstrates the

potential of LLMs in bridging technology and products, providing a structured approach to prompt tuning that enhances the generation of innovative business and product ideas.

The below Task 4 example provides a specific example that underscores the technology commercialization capabilities of our model. Task 4 example introduces the development of heat-blocking blinds with ATO nano ultrafine powder. It describes the increase in infrared reflectivity and energy efficiency, which is expected to open up the market for heat-blocking blinds and expand the export market. Task 4 does not have an explicit label, but we used it as a pseudo-label because there is a commercialization of the technology. The content is detailed, specific, and clear, and the model results show various types of blackout blinds and combirolls.

Task4 Example

Input

과제명: 20nm급 ATO(Antimony Tin Oxide) 무기입자를 이용한 적외선 반사율 30% 이상 향상된 블라인드 제품개발

연구목표: 적외선(IR) 열차단 소재인 ATO 나노 초미립 분말을 첨가한 중적외선 차폐 +직물설계 기술 개발, 열차단 코팅 기술 개발을 통해 높은 단열(Heat Insulation)효과로 에너지 효율 증대를 유도하는 열차단 썬스크린, 암막블라인드, 삼중직 암막 콤비롤 제품을 개발하여 블라인드의 새로운 시장을 형성하고 냉방 및 난방효과의 극대화를 통해 에너지 절감을 유도한다

연구내용: 적외선(IR) 열차단 소재인 ATO 나노 초미립 분말을 첨가한 중적외선 차폐 +직물설계 기술 개발, 열차단 코팅 기술 개발을 통해 높은 단열(Heat Insulation)효과로 에너지 효율 증대를 유도하는 열차단 썬스크린, 암막블라인드, 삼중직 암막 콤비롤 제품을 개발한다

기대효과: 아직까지 시장 미개척 분야인 열차단 블라인드를 개발 함으로써 시장 선점효과를 통한 내수 및 수출 시장의 활로를 찾고 , 차후 차량용 및 산업용 열차단 원단으로 활용 할 수 있는 토대를 마련 할 수 있다고 기대 하고 있다

. . . .

Project Name: Development of blinds with infrared reflectivity improved by more than 30% using 20nm ATO (Antimony Tin Oxide) inorganic particles

Research Goal: Develop mid-infrared shielding + fabric design technology using ATO nano ultrafine powder, an infrared (IR) heat-blocking material, and heat-blocking coating technology to develop heat-blocking sunscreens, blackout blinds, and triple-woven blackout combiroll products that increase energy efficiency with high heat insulation effect, thereby creating a new market for blinds and inducing energy savings through maximizing cooling and heating effects

Research: Develop heat-blocking sunscreens, blackout blinds, and triple-layer blackout combi rolls that increase energy efficiency through high heat insulation effect by adding infrared (IR) heat-blocking material ATO nano ultrafine powder, developing fabric design technology, and developing heat-blocking coating technology.
Expected impact: By developing heat-blocking blinds, which is still an unexplored field, we expect to find avenues for domestic and export markets through market preemption effects, and lay the foundation for future utilization as heat-blocking fabrics for vehicles and industries.

Pseudo Label

사업화명 : H/B 열차단 블라인드
사업화 내용; 2015년 신규제품으로 H/B열차단 블라인드 출시

. . . .

Business name: H/B thermal break blinds
Details of commercialization; Launched H/B heat-blocking blinds as a new product in 2015

Output

사업 아이디어: 연구과제를 기반으로 한 새로운 사업 아이디어
제품: 입력된 정보를 토대로 생성한 제품명과 제품명
제품: 제품명1: 적외선 열차단 블라인드
제품: 제품명2: 암막블라인드
제품: 제품명3: 삼중직 암막 콤비롤 열차단
제품: 제품명4: 중적외선 차폐 +직물설계 블라인드
제품: 제품명5: 암막블라인드 커뮤니티
제품: 제품명6: 삼중직 암막 콤비롤 블라인드
제품: 제품명7: 적외선 열차단 블라인드 - 삼중직 암막 콤비롤
입력된 정보를 토대로 생성한 제품명과 제품명

. . . .

Business idea: A new business idea based on your research project. Product: A product and product name generated based on the information provided. Product: Product name 1: Infrared heat-rejecting blinds
Product: Product Name 2: Blackout Blinds
Product: Product Name3: Triple Blackout Combi-Roll Thermal Break
Product: Product name4: Mid-IR Shielding + Fabric Design Blinds
Product: Product name5: Blackout Blind Community
Product: Product name6: Triple weave blackout combiroll blinds
Product: Product name: 7: Infrared Heat Blocking Blinds - Triple Blackout Combi Roll
Product and product name generated based on the entered information

* Korean data is accompanied by English translations.

*Quantitative analysis using LLM-based evaluation*

To further verify the results, we have expanded our quantitative analysis by incorporating evaluation with GPT-4. Given the nature of our task, which involves generating commercially viable product ideas from technical research information using the capabilities of LLMs, there is no gold label. Therefore, conventional metrics such as BLEU [52] and ROUGE [53] are not suitable as these measures have been demonstrated to exhibit a comparatively weak correlation with human judgments, particularly for tasks that require diversity and creativity. Thus, our experiment was designed to follow an LLM-based evaluator, which has been known to be a promising evaluation method for generative tasks [54–56]. This evaluation was specifically aimed at exploring the potential effectiveness of our model in offering insights into technology commercialization. Inspired by Liu et al. [54], we designed three evaluation prompts: Relevance, Clearness, and Completeness, and the scores are detailed in Fig. 4. Among the criteria mentioned in prior research on LLM evaluation [54, 57], we selected these as the most appropriate for assessing the quality of product ideas generated in the context of technology commercialization.

Relevance: Generated product ideas should be pertinent to the provided technical research information, enhancing their feasibility and value in the context of generating commercially viable ideas.

Clearness: Responses should be understandable and clearly expressed. High clearness indicates that the generated text is free of ambiguity, thus possessing a high degree of practical utility.

Completeness: Generated product ideas should be sufficiently detailed and comprehensive. Completeness ensures that the generated ideas cover all necessary aspects thoroughly.

The scores range from 1 to 5, and we provide the evaluation prompts in the Appendix. We compared the performance of TC-Llama 2 and GI-Llama 2, assessing whether our model is fine-tuned to the technology commercialization data appropriately. Due to output instability in different languages, Llama2 was excluded from these experiments,
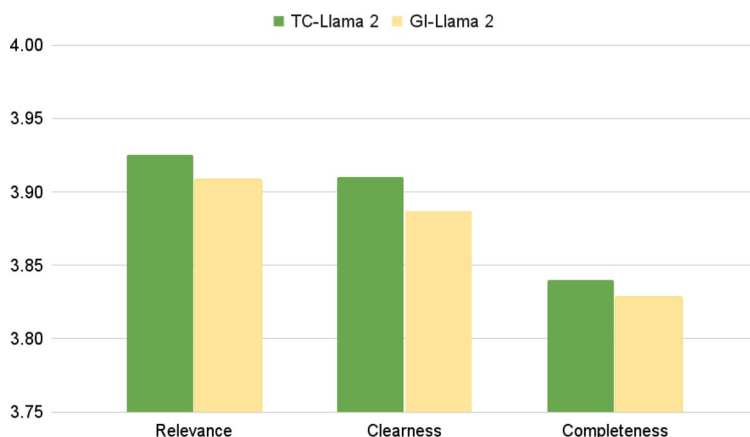
Fig. 4 GPT-4 evaluation results: Comparing TC-Llama 2 and GI-Llama 2 across Relevance, Clearness, and Completeness on a scale of 1 to 5. The results consistently show that TC-Llama 2 outperforms GI-Llama 2 in all metrics, highlighting its superior effectiveness in technology commercialization

focusing our analysis on more stable model outputs. It is important to clarify that GI-Llama 2, developed as the first step of of our model, was specifically designed to enhance the model's bilingual capabilities in Korean and English. The results demonstrate that TC-Llama 2 consistently outperforms GI-Llama 2 across all evaluations. This supports our assertion of the model's effectiveness in the domain of technology commercialization.

Quantitative analysis on predicting commercialization from technology

In addition to the GPT-4 evaluations, we conduct another experiment by utilizing the test dataset from Section 3.2.2, "Predicting Commercialization from Technology" task. This experiment aimed to validate the model's capacity of identifying and predicting technology-commercialization relationships effectively. To further test our model under conditions mirroring practical deployment, we utilized metrics commonly used in language translation and summarization, including ROUGE [53], BLEU [52], and METEOR [58] scores, adjusted to accommodate the specific challenges of our task. Results, shown in Table 10, indicate that TC-Llama2 significantly outperforms GI-Llama2, affirming our model's specialized effectiveness in technology-commercialization contexts. Due to output instability in different languages, Llama2 was excluded from these experiments.

Conclusion

In this study, we have demonstrated the capabilities of TC-Llama 2, a fine-tuned LLM, in enhancing technology commercialization processes. Our approach entailed the development and application of LLMs for intricate tasks such as product to category mapping, technology-product relationship inference. The model, with its advanced tuning methods and training on diverse datasets, showed superior performance over base models, particularly in handling complex category systems and generating contextually coherent text. The innovative use of embedding mapping and iterative instruction-based refinement further augmented the model's efficiency and applicability, illustrating the potential of LLMs in bridging technology and commercialization. This research sets a new milestone in LLM applications, contributing significantly to natural language processing and opening new avenues for AI in technology commercialization and related fields.

Table 10 Performance Comparison on Section 3.2.2 test dataset

Metrics	GI-Llama 2	TC-Llama 2
ROUGE-1	0.2202	**0.3375**
ROUGE-2	0.0701	**0.1760**
ROUGE-l	0.1272	**0.2060**
BLEU-1	0.1327	**0.2260**
BLEU-2	0.0729	**0.1639**
BLEU-3	0.0479	**0.1296**
BLEU-4	0.0317	**0.0986**
METEOR	0.1268	**0.2579**

All scores are in the range between 0 and 1, with higher being better. The higher scores are marked in bold. TC-Llama significantly outperforms GI-Llama

Appendix

Example prompts

The below are the three prompts used in the GPT-4 evaluation experiment. To facilitate understanding of the Korean prompts, we have also included translated versions in English.

Prompt for relevence evaluation

업무 지침:
당신은 주어진 연구 과제를 사업화하는데 기반한 사업, 제품화 아이디어를 평가할 것입니다. 평가할 때는 한 가지 지표에만 집중해야 합니다. 지침을 주의 깊게 읽고 이해해야 하며, 평가하는 동안 필요할 때 이 문서를 참조하세요.

평가 기준:
연관성 (1-5) - 모든 문장의 집합적인 품질뿐만 아니라 정보의 연결 정도를 평가합니다. 이 지표는 ”요약은 잘 구조화되고 조직되어야 하며, 관련 정보의 단순 나열이 아닌, 주제에 대한 일관된 정보의 본문을 구성해야 한다”는 DUC의 구조 및 연관성에 관한 질문과 일치합니다.

평가 절차:

1. *연구 과제 문서를 주의 깊게 읽고 주요 주제와 핵심 포인트를 식별하십시오.*
2. *사업 아이디어를 읽고 연구 과제 문서와 비교하십시오. 사업 및 제품화 아이디어가 연구 과제의 주요 주제와 핵심 포인트를 커버하는지, 그리고 명확하고 논리적인지 확인하십시오.*
3. *평가 기준에 따라 연관성에 대한 점수를 1에서 5 사이의 척도로 매기십시오. 여기서 1은 가장 낮은 점수이고 5는 가장 높은 점수입니다.*

예시:
연구 과제: {Document}
사업 아이디어: {Business Idea}

평가 결과 (점수만 출력):
- 연관성:

Translated version

Instructions:

You will evaluate business and product ideas based on the commercialization of the given research project. Focus on a single criterion when evaluating. Carefully read and understand the instructions, and refer to this document as needed during the evaluation.

Evaluation Criterion:

Relevance (1-5) - Evaluate the collective quality of all sentences as well as the degree of information connection. This criterion matches the question of structure and relevance by DUC that states, "Summaries must be well structured and organized, and not merely a list of related information, but a coherent body of information on the topic.

Evaluation Procedure:

1. *Carefully read the research document and identify the main topics and key points.*
2. *Read the business idea and compare it with the research document. Check whether the business and productization ideas cover the main themes and key points of the research project, and ensure they are clear and logical.*

3. *According to the evaluation criteria, score relevence on a scale from 1 to 5. 1 is the lowest score, and 5 is the highest score.*

Example:

Research: {Document}

Business Idea: {Business Idea}

Evaluation Results (scores only):

- Relevence:

Prompt for clearness evaluation

업무 지침:
당신은 주어진 연구 과제를 사업화하는데 기반한 사업, 제품화 아이디어를 평가할 것입니다. 평가할 때는 한 가지 지표에만 집중해야 합니다. 지침을 주의 깊게 읽고 이해해야 하며, 평가하는 동안 필요할 때 이 문서를 참조하세요.

평가 기준:
명확성 (1-5) - 제안된 사업 및 제품화 아이디어가 얼마나 명확하게 표현되었는지를 평가합니다. 이 지표는 아이디어가 이해하기 쉽고, 목적이 분명하며, 불필요한 혼란이나 모호함이 없는 정도를 포함합니다.

평가 절차:
1. 연구 과제 문서를 주의 깊게 읽고 주요 주제와 핵심 포인트를 식별하십시오.
2. 제안된 사업 아이디어를 읽고 연구 과제 문서와 비교하십시오. 사업 및 제품화 아이디어가 목적이 명확하고 이해하기 쉬운 방식으로 제시되었는지 확인하십시오.
3. 아이디어의 표현이 얼마나 명확한지, 불필요한 혼란이나 모호함 없이 직접적으로 이해될 수 있는지 평가하십시오. 평가 기준에 따라 명확성에 대한 점수를 1에서 5 사이의 척도로 매기십시오. 여기서 1은 가장 낮은 점수이고 5는 가장 높은 점수입니다.

예시:
연구 과제: {*Document*}

사업 아이디어: {*Business Idea*}

평가 결과 *(점수만 출력)*:
- 명확성:

Translated version

Instructions:

You will evaluate business and product ideas based on the commercialization of the given research project. Focus on a single criterion when evaluating. Carefully read and understand the instructions, and refer to this document as needed during the evaluation.

Evaluation Criterion:

Clearness (1-5) - Assess how clearly the proposed business and productization ideas are expressed. This criterion includes how easy the ideas are to understand, how clear their purposes are, and the degree to which they are free from unnecessary confusion or ambiguity.

Evaluation Procedure:

1. *Carefully read the research document and identify the main topics and key points.*
2. *Read the proposed business idea and compare it with the research project document. Ensure that the business and productization ideas are presented in a manner that is clear and easy to understand.*

3. *Evaluate how clear the expression of the idea is, and whether it can be understood directly without any unnecessary confusion or ambiguity. Score the clearness on a scale from 1 to 5, where 1 is the lowest score and 5 is the highest score.*

Example:

Research: {Document}

Business Idea: {Business Idea}

Evaluation Results (scores only):

- Clearness:

Prompt for completeness evaluation

업무 지침:
당신은 주어진 연구 과제를 사업화하는데 기반한 사업, 제품화 아이디어를 평가할 것입니다. 평가할 때는 한 가지 지표에만 집중해야 합니다. 지침을 주의 깊게 읽고 이해해야 하며, 평가하는 동안 필요할 때 이 문서를 참조하세요.

평가 기준:
완성도 (1-5) - 제안된 사업 및 제품화 아이디어의 기술적, 구조적 완성도를 평가합니다. 완성도는 기술적 가능성, 자원의 효율적 활용, 제품이나 서비스의 최종 사용성을 포함하며, 기술적 제약이나 자원 부족이 제품의 완성도를 저해하지 않는지를 평가합니다.

평가 절차:

1. 연구 과제 문서를 주의 깊게 읽고 주요 주제와 핵심 포인트를 식별하십시오.
2. 제안된 사업 아이디어가 연구 과제의 주요 주제와 핵심 포인트를 어떻게 커버하고 있는지, 그리고 그 아이디어가 완성도가 있는지 확인하십시오.
3. 평가 기준에 따라 완성도에 대한 점수를 1에서 5 사이의 척도로 매기십시오. 여기서 1은 가장 낮은 점수이고 5는 가장 높은 점수입니다.

예시:
연구 과제: {*Document*}
사업 아이디어: {*Business Idea*}

평가 결과 (점수만 출력):
- 완성도:

Translated version

Instructions:

You will evaluate business and product ideas based on the commercialization of the given research project. Focus on a single criterion when evaluating. Carefully read and understand the instructions, and refer to this document as needed during the evaluation.

Evaluation Criterion:

Completeness (1-5) - Assess the technical and structural completeness of the proposed business and productization ideas. Completeness includes the technical feasibility, efficient use of resources, and the final usability of the product or service, and whether technical constraints or lack of resources compromise the product's completeness.

Evaluation Procedure:

1. *Carefully read the research document and identify the main topics and key points.*
2. *Review the proposed business idea to see how it covers the main topics and key points of the research project, and check if the idea has completeness.*

3. *According to the evaluation criteria, score completeness on a scale from 1 to 5. 1 is the lowest score, and 5 is the highest score.*

Example:

   Research: {Document}

   Business Idea: {Business Idea}

   Evaluation Results (scores only):

   - Completeness:

Data availability
The dataset and code are available on request.

Declarations

Ethics approval and consent to participate
There are no ethical issues with this research.

Competing interests
There is no Conflict of interest.

References
1. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, et al. A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
2. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
3. Miao H, Wang Y, Li X, Wu F. Integrating technology-relationship-technology semantic analysis and technology roadmapping method: a case of elderly smart wear technology. IEEE Trans Eng Manag. 2020;69(1):262–78.
4. Yan W, Chen C-H, Huang Y, Mi W. A data-mining approach for product conceptualization in a web-based architecture. Comput Ind. 2009;60(1):21–34.
5. Jeong C, Kim K. Technology relationship analysis using problem and solution similarities. In: 2012 IEEE International Conference on Management of Innovation & Technology (ICMIT), pp. 516–521 (2012). IEEE
6. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
7. Hu EJ, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W, et al. Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)
8. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, Li J, Hu R, Zhang T, Wu F, et al. Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792 (2023)
9. Radford A, Narasimhan K, Salimans T, Sutskever I, et al. Improving language understanding by generative pre-training. 2018.
10. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, Casas DDL, Hendricks LA, Welbl J, Clark A, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556. 2022.
11. Han X, Zhang Z, Ding N, Gu Y, Liu X, Huo Y, Qiu J, Yao Y, Zhang A, Zhang L, et al. Pre-trained models: past, present and future. AI Open. 2021;2:225–50.
12. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D. Recent advances in natural language processing via large pre-trained language models: A survey. ACM Comput Surv. 2023;56(2):1–40.
13. Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, Du N, Dai AM, Le QV. Finetuned language models are zero-shot learners. In: International Conference on Learning Representations. 2021.

Yeom *et al. Journal of Big Data*   (2024) 11:100

Page 30 of 31

14. Sanh V, Webson A, Raffel C, Bach S, Sutawika L, Alyafeai Z, Chaffin A, Stiegler A, Raja A, Dey M, et al. Multitask prompted training enables zero-shot task generalization. In: International Conference on Learning Representations. 2021.

15. Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, De Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S. Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning, pp. 2790–2799. 2019. PMLR

16. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059. 2021.

17. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB. Stanford alpaca: An instruction-following llama model. 2023.

18. Cui Y, Yang Z, Yao X. Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177. 2023.

19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.

20. OpenAI R. Gpt-4 technical report. arxiv 2303.08774. View in Article. 2. 2023.

21. Dong Q, Li L, Dai D, Zheng C, Wu Z, Chang B, Sun X, Xu J, Sui Z. A survey for in-context learning. arXiv preprint arXiv:2301.00234. 2022.

22. Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le QV, Zhou D, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37.

23. Honovich O, Shaham U, Bowman S, Levy O. Instruction induction: From few examples to natural language task descriptions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1935–1952. 2023.

24. Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, Hajishirzi H. Self-instruct: Aligning language models with self-generated instructions. In: The 61st Annual Meeting Of The Association For Computational Linguistics. 2023.

25. Min S, Lewis M, Zettlemoyer L, Hajishirzi H. Metaicl: Learning to learn in context. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2791–2809. 2022.

26. Zahera HM. Probert: Product data classification with fine-tuning bert model. 2020.

27. Geng W, Han F, Lin J, Zhu L, Bai J, Wang S, He L, Xiao Q, Lai Z. Fine-grained grocery product recognition by one-shot learning. MM '18, pp. 1706–1714. Association for Computing Machinery, New York, NY, USA. 2018. https://doi.org/10.1145/3240508.3240522.

28. Wei, Y., Tran, S., Xu, S., Kang, B., Springer, M., Panella, M. Deep learning for retail product recognition: challenges and techniques. Comput Intell Neurosci. 2020;2020:8875910. https://doi.org/10.1155/2020/8875910.

29. Jahanshahi H, Ozyegen O, Cevik M, Bulut B, Yigit D, Gonen FF, Başar A. Text classification for predicting multi-level product categories. CASCON '21, pp. 33–42. IBM Corp., USA. 2021.

30. Peng J, Xiao C, Li Y. Rp2k: A large-scale retail product dataset for fine-grained image classification. arXiv preprint arXiv:2006.12634. 2020.

31. Jahanshahi H, Ozyegen O, Cevik M, Bulut B, Yigit D, Gonen FF, Başar A. Text classification for predicting multi-level product categories. In: Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering, pp. 33–42. 2021.

32. Nguyen TT, Wilson C, Dalins J. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. arXiv preprint arXiv:2308.14683. 2023.

33. Pavlyshenko BM. Financial news analytics using fine-tuned llama 2 gpt model. arXiv preprint arXiv:2308.13032. 2023.

34. Chen P, Ji S, Bogoychev N, Kutuzov A, Haddow B, Heafield K. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In: The 18th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1–10 (2024). Association for Computational Linguistics

35. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang P, Hashimoto TB. Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html 3(6), 7. 2023.

36. Rasley J, Rajbhandari S, Ruwase O, He Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 3505–3506. 2020.

37. Lee C, Jeon D, Ahn JM, Kwon O. Navigating a product landscape for technology opportunity analysis: a word2vec approach using an integrated patent-product database. Technovation. 2020;96: 102140.

38. Aghajanyan A, Gupta A, Shrivastava A, Chen X, Zettlemoyer L, Gupta S. Muppet: Massive multi-task representations with pre-finetuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5799–5811. 2021.

39. Luo Y, Yang Z, Meng F, Li Y, Zhou J, Zhang Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747. 2023.

40. Mishra S, Khashabi D, Baral C, Hajishirzi H. Cross-task generalization via natural language crowdsourcing instructions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3470–3487. 2022.

41. Bach S, Sanh V, Yong ZX, Webson A, Raffel C, Nayak NV, Sharma A, Kim T, Bari MS, Févry T, et al. Promptsource: An integrated development environment and repository for natural language prompts. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 93–104. 2022.

42. Wang Y, Mishra S, Alipoormolabashi P, Kordi Y, Mirzaei A, Arunkumar A, Ashok A, Dhanasekaran AS, Naik A, Stap D, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In: 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022. 2022.

43. Gao D, Yang W, Zhou H, Wei Y, Hu Y, Wang H. Deep hierarchical classification for category prediction in e-commerce system. ECNLP. 2020;3:64.

44. Zhang W, Lu Y, Dubrov B, Xu Z, Shang S, Maldonado E. Deep hierarchical product classification based on pre-trained multilingual knowledge. 2021.

45. Muennighoff N. Sgpt: Gpt sentence embeddings for semantic search. arXiv preprint arXiv:2202.08904. 2022.
46. Li Z, Li X, Liu Y, Xie H, Li J, Wang F-l, Li Q, Zhong X. Label supervised llama finetuning. arXiv preprint arXiv:2310.01208. 2023.
47. Feng X, Yoshimoto A. Llama-vits: Enhancing tts synthesis with semantic awareness. arXiv preprint arXiv:2404.06714. 2024.
48. Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Improving text embeddings with large language models. arXiv preprint arXiv:2401.00368. 2023.
49. Lei Y, Wu D, Zhou T, Shen T, Cao Y, Tao C, Yates A. Meta-task prompting elicits embedding from large language models. arXiv preprint arXiv:2402.18458. 2024.
50. Cheng J, Liu X, Zheng K, Ke P, Wang H, Dong Y, Tang J, Huang M. Black-box prompt optimization: Aligning large language models without model training. arXiv preprint arXiv:2311.04155. 2023.
51. Chen L, Chen J, Goldstein T, Huang H, Zhou T. Instructzero: Efficient instruction optimization for black-box large language models. arXiv preprint arXiv:2306.03082. 2023.
52. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. 2002.
53. Lin C-Y. Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. 2004.
54. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-eval: Nlg evaluation using gpt-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2511–2522. 2023.
55. Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, Xu J, Qu J, Zhou J. Is chatgpt a good nlg evaluator? a preliminary study. In: Proceedings of EMNLP Workshop, p. 1. 2023.
56. Zhong M, Liu Y, Yin D, Mao Y, Jiao Y, Liu P, Zhu C, Ji H, Han J. Towards a unified multi-dimensional evaluator for text generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 2023–2038. 2022.
57. Zhang Z, Zheng C, Tang D, Sun K, Ma Y, Bu Y, Zhou X, Zhao L. Balancing specialized and general skills in llms: The impact of modern tuning and data strategy. arXiv preprint arXiv:2310.04945. 2023.
58. Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72. 2005.

Publisher's Note