URL: Universal Referential Knowledge Linking via Task-instructed Representation Compression

Zhuoqun Li[1,4], **Hongyu Lin**[1], **Tianshu Wang**[1,4], **Boxi Cao**[1,4], **Yaojie Lu**[1], **Weixiang Zhou**[1], **Hao Wang**[3], **Zhenyu Zeng**[3], **Le Sun**[1,2], **Xianpei Han**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]Alibaba Cloud Intelligence Group
[4]University of Chinese Academy of Sciences, Beijing, China
{lizhuoqun2021,hongyu,tianshu2020,boxi2020,luyaojie,weixiang}@iscas.ac.cn
cashenry@126.com zhenyu.zzy@alibaba-inc.com {sunle,xianpei}@iscas.ac.cn

Abstract

Linking a claim to grounded references is a critical ability to fulfill human demands for authentic and reliable information. Current studies are limited to specific tasks like information retrieval or semantic matching, where the claim-reference relationships are unique and fixed, while the referential knowledge linking (RKL) in real-world can be much more diverse and complex. In this paper, we propose universal referential knowledge linking (URL), which aims to resolve diversified referential knowledge linking tasks by one unified model. To this end, we propose a LLM-driven task-instructed representation compression, as well as a multi-view learning approach, in order to effectively adapt the instruction following and semantic understanding abilities of LLMs to referential knowledge linking. Furthermore, we also construct a new benchmark to evaluate ability of models on referential knowledge linking tasks across different scenarios. Experiments demonstrate that universal RKL is challenging for existing approaches, while the proposed framework can effectively resolve the task across various scenarios, and therefore outperforms previous approaches by a large margin.

1 Introduction

Access to reliable, authentic, and well-founded information is a fundamental human necessity (Clifford, 1877; Cole, 2011). In recent years, with the rise of AI-generated content (AIGC), there has been an increasing demand for the verification of information authenticity and the identification of corresponding references (Lewis et al., 2020; Cao et al., 2023; Zhao et al., 2023). Within this context, linking a claim to its grounded references, a task we refer to as ***referential knowledge linking (RKL)***, plays an indispensable role. Given a claim, the objective of referential knowledge linking is to associate it with relevant references within trustworthy information sources, thereby facilitating the verification of related information (Yue et al., 2023; Huang et al., 2023; Gao et al., 2023). RKL is crucial for affirming the accuracy and validity of information, ensuring the reliability and legality of AIGC-produced content, and the efficient organization and management of information.

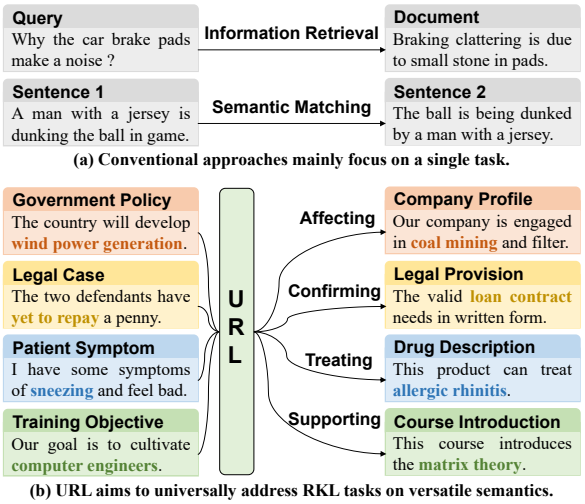

Figure 1: Compared to conventional approaches focus on a single task, URL aims to universally address RKL tasks on versatile semantics with deep knowledge.

Currently, the work on referential knowledge linking is often confined to specific tasks and contexts. As illustrated in Figure 1, information retrieval (Hambarde and Proença, 2023; Muennighoff et al., 2023; Xiao et al., 2023) focuses on locating documents that contain answers to a query, primarily concentrating on the "contains-answer" type of relationship. Semantic matching, on the other hand, aims to identify pairs of text segments that share the same semantics (Cer et al., 2017; Chandrasekaran and Mago, 2021), essentially looking for segments that are "semantically equivalent". However, the scenarios for referential knowledge linking in reality are far more diverse. Depending on the particular claim and reference, the objectives of linking can vary significantly. For instance,

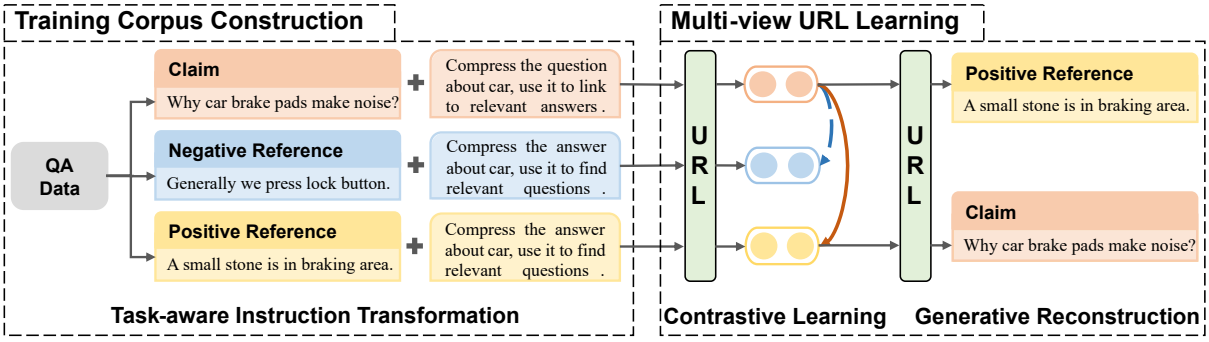

Figure 2: Illustration of training corpus construction and multi-view URL learning. Based on QA data, we set the question as claim and answer as reference, then annotate instructions that describe the field of data and the purpose of representation. For learning, contrastive learning is on embeddings of claims and references, generative reconstruction is to force the model generating positive reference based on claim embedding and vice versa.

given a legal case, one might aim to link it to corresponding legal provisions to find the basis of the verdict. Conversely, given a patient symptom, it could be linked to relevant drugs, treatment plans, or possible cause, depending on the specific need.

Compared to the previous focus on specific contexts such as information retrieval or semantic matching, it is challenging to universally solve referential knowledge linking in diverse real-world scenarios. Firstly, although all under the paradigm of RKL, the target semantics of RKL vary across different contexts, i.e., whether a claim and a reference constitute a linking pair may differ under different RKL contexts. For example, given a legal case and a provision, the decision on whether they should be linked may vary drastically under different relational contexts like "applying for adults" or "applying for minors". Secondly, unlike information retrieval and semantic matching, which often only require superficial semantic understanding (e.g., word matching features and shallow semantic representations), many RKL tasks necessitate deep semantic comprehension and reasoning. Therefore, RKL demands models with robust semantic understanding and, often potentially, reasoning capabilities. Finally, given the typically large size of reference databases, it is impractical to directly perform complex, deep semantic interactions and matching across all claim-reference pairs. Thus, constructing and implementing an efficient, highly adaptable, and semantically capable general RKL model poses a significant challenge.

In this paper, we introduce ***unified referential linking (URL)***, a universal framework for linking claims to references on versatile semantics. Specifically, URL leverages the semantic comprehension and instruction-following capabilities of large language models (LLMs) to facilitate universal referential knowledge linking across diverse contexts. However, applying LLMs to RKL poses several challenges due to differences in training objectives and usage modalities. Firstly, the high computational cost of large language models makes direct pair-wise comparison for claim-reference decision-making highly inefficient. Secondly, since large language models are trained in a language model paradigm, directly employing them for universal referential knowledge linking may lead to mismatches in patterns, thereby significantly impacting the performance of LLMs in general RKL tasks. Addressing these key challenges is crucial for harnessing the capabilities of large language models for referential knowledge linking.

To this end, URL introduces a LLM-driven task-instructed representation compression mechanism, which adaptively integrates task-specific information with claim/reference and converts them into vector representations. This approach allows the representations of claims and references to be efficiently adapted across different task scenarios. Levering these vectorized representations, the linking between claims and references can be conducted by directly calculating the association between their task-aware representations, thereby significantly reducing the need for complex computations with large models and enhancing the efficiency of RKL. To facilitate this, we propose a parameter-efficient, multi-view learning algorithm that enables joint training of the large models in both generation and linking modes under the constraint of only observing compressed representations. As illustrated in Figure 2, this approach can

effectively learn the compressed vector representation by shifting the mode of large language models from context-aware generation to compressed representation-aware generation and linking. Moreover, to train a generalizable and universal RKL representation model effectively, we start with existing question answering (QA) corpus. By transforming QA data from various domains into diverse claim-reference datasets under different linking relationships, we align the model from a language generation mode to a representation compression mode, thereby learning better task-aware RKL representation models.

To validate the effectiveness of universal referential knowledge linking, we introduce a new benchmark—***URLBench***. The primary objective of URLBench is to construct an evaluation set that covers multiple scenarios of referential knowledge linking, thereby assessing the ability of model in generalized referential knowledge linking. Specifically, URLBench encompasses four distinct domains: finance, law, medicine, and education. Within these domains, URLBench evaluates the model capability to link claims and references under various semantic conditions, which primarily includes linking government policy to company profile, legal case to legal provision, patient symptom to drug description, and training objective to course introduction. Experimental results on URLBench demonstrate that URL is an effective approach for achieving universal RKL. Its performance not only surpasses previous models trained on large-scale retrieval and semantic matching datasets, but also significantly outperforms proprietary large-model-based embedding models like OpenAI Text Embedding. This validates the efficacy of the approach proposed in this paper for RKL tasks[1].

The main contributions of this paper are:

- We define the task of universal referential knowledge linking, which extends beyond considering only a few specific relationships to achieve generalized RKL abilities across diverse scenarios.
- We propose task-instructed representation compression, a novel framework adapting LLMs to RKL tasks. Meanwhile, we propose a new method based on multi-view RKL learning, which can effectively finetune LLMs by existing QA data.
- We construct URLBench, a new benchmark across versatile knowledge-rich tasks in various fields, which can effectively evaluate the ability of models in universal referential knowledge linking.

[1]Our code and data are openly available at https://github.com/Li-Z-Q/URL

2 Related Work

2.1 Semantic Matching

Semantic matching is to measure the semantic similarity between two blocks of text (Chandrasekaran and Mago, 2021). This is a traditional and fundamental task in natural language processing, containing datasets across many domains and language (Agirre et al., 2012, 2013, 2014, 2015, 2016; Marelli et al., 2014; Cer et al., 2017). However, the relationship in this task is single and static, which is merely about shallow semantic of two sentences, judging whether that "have similar meaning".

2.2 Information Retrieval

Information retrieval (IR) is the process of searching and returning relevant documents for a query from a collection (Hambarde and Proença, 2023; Muennighoff et al., 2023). IR contains datasets in various fields and task formats such as question-answer, title-passage, query-document (Thakur et al., 2021; Kwiatkowski et al., 2019; Cohan et al., 2020; Xiao et al., 2023; Thorne et al., 2018). However, the essence of all the IR tasks is identical, which is judging if the claim and reference "contain similar information". In other words, the relationship in IR is still lacked of universality.

2.3 Sentence Embedding

Early embedding methods are based on BERT-style models (Reimers and Gurevych, 2019). Then some works use contrastive learning and get improvement (Gao et al., 2021; Izacard et al., 2021). Recent common methods are using two-stage contrastive learning by large scale corpus (Wang et al., 2022; Xiao et al., 2023; Li et al., 2023b), based on special trained RetroMAE (Xiao et al., 2022). And some works also use task instructions on BERT-style models (Su et al., 2023).

Recently, some works use LLMs to generate sentence embeddings by instruction (Jiang et al., 2023), and contrastive learning (Ma et al., 2023; Zhang et al., 2023). And embeddings generated by LLMs can be used to do retrieval (Wang et al., 2024; Li et al., 2023a), or compression (Gupta et al., 2023; Mu et al., 2023; Chevalier et al., 2023).

The important distinction between this work and previous LLM embeddings is that URL focus on addressing versatile RKL tasks from a unified perspective, which receives less attention before. Additionally, beyond contrastive learning, this work proposes a multi-view learning approach.

3 Universal RKL via Task-instructed Representation Compression

Universally addressing RKL tasks poses significant challenges because of versatile linking relationships and deep semantic of claim-reference pairs. Firstly, the relationship between claims and references is flexible across different contexts, requiring varied semantic abilities for linking. Secondly, the relationship involves deep semantic correlations and even some reasoning, demanding deep knowledge and reasoning capabilities from the model. Moreover, handling large-scale reference datasets directly by complex semantic processing on each claim-reference pair causes large time costs.

In this paper, we propose a novel framework for universal RKL tasks based on LLMs with powerful knowledge, semantic understanding and universality. But applying LLMs presents several challenges due to computational cost and original training objective. Firstly, direct pairwise comparison by LLMs is inefficient, exacerbating efficiency issues with large-scale reference databases. Secondly, since LLMs are trained as language models, directly applying to RKL encounters pattern mismatching. To tackle mentioned challenges, URL employs a task-instructed representation compression mechanism driven by LLMs. The method dynamically incorporates task-specific details with claim/reference data, generating vector representations. To facilitate LLMs for this, this paper introduces a multi-view learning algorithm to enable simultaneous training in both generation and linking modes, on only compressed representations.

Specifically, by LLM-driven task-instructed representation compression, URL combines task information with claims/references and gets task-instructed embeddings, achieving efficient adaptation across different scenarios. In finetuning, multi-view learning injects knowledge of LLMs into representations and aligns representations with claim-reference embedding-similarity paradigms. To construct training data, URL leverages existing question answering data, transforming QA data from different domains into claim-reference data. This data construction method is a convenient way to simulate real versatile RKL tasks.

3.1 Task-instructed Compression for RKL

Applying LLMs to practical RKL tasks faces two major challenges: adaptability and efficiency. In adaptability, linking from claim to reference requires diverse semantic abilities depending on the context or application, necessitating a model capable of adapting to various contexts or tasks. In efficiency, there is large computation cost because of the low operating speed of LLMs and possible large scale of reference databases, requiring some efficient linking better than direct comparison. In this paper, URL addresses mentioned challenges by adding task-aware instructions to claims and references, enabling the model to generate task-instructed representations. By computing vectorized representations similarity, URL achieves efficient and adaptive linking for versatile RKL tasks.

Specifically, task-instructed compression compresses the claim/reference, as well as the instruction of target task to be performed into the same vectorized representation, thereby constructing a task oriented vectorized representation:

$$\mathbf{H} = \text{LLM}(\boldsymbol{x}_s, \boldsymbol{x}_{ins}, \boldsymbol{x}_{suffix})$$
$$\mathbf{e}_s = \text{Pooling}(\mathbf{H}_{suffix})$$

where $\boldsymbol{x}_s$ is the claim or reference, $\boldsymbol{x}_{ins}$ is the task-aware instruction adapted with the claim or reference, $\boldsymbol{x}_{suffix}$ means some suffix tokens, $\mathbf{H}$ is last-layer hidden states of all tokens and $\mathbf{H}_{suffix}$ is that of suffix tokens, $\mathbf{e}_s$ is the output embedding.

In summary, by calculating vector similarity of task-instructed representations of the claim and reference, URL gets the linking score of claim-reference pairs, and address RKL tasks efficiently and universally without complex computation.

3.2 Multi-view URL Learning

Under the aforementioned URL framework, a core issue is how to effectively learn task-instructed representation for versatile RKL tasks. Given that LLMs are primarily designed for language generation tasks, directly applying representations from LLMs to RKL tasks is challenging. And solely employing claim-reference pairs contrastive learning presents damaging the original knowledge and reasoning capabilities obtained during pre-training. In this paper, we propose a novel multi-view learning approach that facilitates simultaneous gener-

ative reconstruction and contrastive learning processes, enabling large models to swiftly and cost-effectively adapt to URL norms while preserving existing knowledge and reasoning abilities.

Specifically, multi-view URL learning comprises two components: generative reconstruction and contrastive learning. The generative reconstruction employs claim embeddings to generate relevant references and reference embeddings to generate relevant claims, enabling the injection of model knowledge into compressed representations to better address reasoning-style linking tasks in RKL. The second component, contrastive learning, minimizes the distance between related claim and reference representations while maximizing the distance between unrelated ones, enabling model-generated representations to align with the embedding-similarity paradigm.

Generative Reconstruction. By generative loss based on the claim representation with relevant reference, and reference representation with relevant claim, generative reconstruction injects knowledge of the LLM into the compressed representation:

$$\mathcal{L}_1 = \sum_{\mathcal{D}} -\log p(\boldsymbol{x}_{pos}|\mathbf{e}_c, \boldsymbol{x}_p)$$

where $\mathcal{D}$ means training data, $\mathbf{e}_c$ is the claim embedding, $\boldsymbol{x}_{pos}$ is the positive reference, $\boldsymbol{x}_p$ is the prompt used to guide the LLM to generate the relevant reference. On another hand, reconstructing claim based on reference embedding has symmetric formula as above. Through generative reconstruction, the representation contains great knowledge and deep understanding of the input sentence because the LLM can generate relevant contents merely based on the compressed representation rather than original complete sentences.

Contrastive Learning. By minimizing the distance between representations of the relevant claim and reference and maximizing that of the irrelevant claim and reference, contrastive learning directly train the LLM for suiting the claim-reference embedding-similarity pattern for URL:

$$\mathcal{L}_2 = \sum_{\mathcal{D}} -\log \frac{e^{\langle \mathbf{e}_c, \mathbf{e}_{pos} \rangle / \tau}}{e^{\langle \mathbf{e}_c, \mathbf{e}_{pos} \rangle / \tau} + \sum_{\mathbf{e}_{neg}} e^{\langle \mathbf{e}_c, \mathbf{e}_{neg} \rangle / \tau}}$$

where $\mathbf{e}_{pos}$ is embedding of the positive reference, $\mathbf{e}_{neg}$ is for the negative reference, and τ is the temperature parameter. Through contrastive learning, the LLM are trained to efficiently address RKL tasks by embedding-similarity method.

Multi-view Learning. By composing the loss of generative reconstruction and contrastive learning, multi-view URL learning integrates the advantages of two training methods, the LLM can be well aligned for versatile knowledge-rich RKL tasks:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_1 + (1 - \alpha)\mathcal{L}_2$$

where α is a parameter. In actual training, URL uses symmetric bi-direct loss as some recent works (Xiao et al., 2023). Respectively, for generative reconstruction, URL also calculates loss by generating the claim based on embedding of the correct reference, for contrastive learning, URL also calculates loss among embeddings of one reference, the relevant claim, and irrelevant claims.

In summary, multi-view URL learning ensures task-instructed representations encompass knowledge of the LLM, and suit for the claim-reference embedding-similarity pattern. The total training process can effectively align the LLM for universally addressing versatile RKL tasks.

3.3 Constructing URL Training Data via QA Corpus Transformation

In order to support the multi-view URL learning mentioned above, a core issue is to construct training data for versatile RKL tasks. This presents a crucial challenge as existing works primarily focus on information retrieval and semantic matching, both of which entail singular linking relationships, lacking diverse training data for versatile linking tasks. In this paper, we propose constructing a URL training corpus by transforming QA data. By annotating instructions for ordinary question-answer data according to the domain of data, this construction method utilizes versatile domain data to simulate versatile tasks in RKL.

Specifically, for 1000 question-answer pairs selected from mMARCO (Bonifacio et al., 2021), we manually categorize data into 40 domains, annotate instructions for each domain, and set the question as claim and the answer as reference. As illustrated in Figure 2, each data includes a claim, a positive reference, some negative references by random sampling, and two simulating task-aware instructions. In addition, this training dataset has versions for both Chinese and English language.

In summary, the method uses heterogeneous domains to simulate versatile relationships of RKL. This is an efficient alternative solution in the situation absenting real versatile training data for RKL.

Domain	Application	Relationship	Claim				Reference			
			Content	Number	Length		Content	Number	Length	
Finance	Stock Decision	Affecting	Government Policy	1000	157	252	Company Profile	709	1026	922
Law	Legal Judgement	Confirming	Legal Case	1000	236	390	Legal Provision	3627	56	103
Medicine	Medical Prescription	Treating	Patient Symptom	750	80	148	Drug Description	1000	36	70
Education	Course Planning	Supporting	Training Objective	133	90	145	Course Introduction	787	115	153

Table 1: Detailed statistics of the benchmark. Length is the average token number of sentences by LLM tokenizers, the left value is for Chinese version and the right is for English version.

4 URLBench: Benchmarking Universal Referential Knowledge Linking

Currently, RKL benchmarks primarily focus on two specific tasks: semantic matching and information retrieval. The former emphasizes "having similar semantics", while the latter focuses on "containing similar information". However, RKL tasks are versatile and with challenging knowledge-rich relationships in real-world applications, there is a lack of a benchmark that covers versatile RKL tasks and reflects complex deep linking relationships, posing a significant obstacle to evaluating unified models. In this paper, we propose a new benchmark that encompasses various RKL tasks across multiple domains and real applications. The benchmark can greatly reflect the diversity and knowledge-rich challenge of RKL tasks.

Specifically, by focusing on four real-world applications including stock decision, legal judgement, medical prescription and course planning, URLBench collects data from web or existing datasets to create four evaluation tasks: linking government policies and company profiles, legal cases and legal provisions, patient symptoms and drug descriptions, and training objectives and course introductions. Focusing on stock decision, URLBench contains linking from government policies to companies that can be affected, constructed by transforming an existing task (Wang et al., 2023) to policy-company linking format. Focusing on legal judgement, URLBench contains linking from legal cases to provisions that can confirm the case, constructed by transforming and extending an existing classification task (Xiao et al., 2021) referring to web[2]. Focusing on medical prescription, URLBench contains linking from patient symptoms to drugs that can treat the symptom, constructed by transforming and extending an existing classification task (He et al., 2022) referring to web[3]. Focusing on course planning, URLBench contains linking from training objectives to courses that can support the objective. This task is constructed based on a university enrollment handbook[4].

Clear statistics are shown in Table 1, the scale of the benchmark can ensure experiments quick and effective, and the length of datas is suitable for most models. Note that these datasets are originally mainly in Chinese language, we do translation by gpt-3.5-turbo[5] to get English version. We then do some filter and checking to ensure the quality.

5 Experiments

5.1 Experimental Settings

We choose LLAMA-2-7B-Chat (Touvron et al., 2023) as the base model for English tasks, and Baichuan-2-7B-Chat (Yang et al., 2023) for Chinese tasks. We finetune LLMs by the 1000 training data by LoRA method (Hu et al., 2022) in multi-view learning, setting lora_r as 8, lora_alpha as 16 and lora_dropout as 0.05. Respectively, we set lora_target as "W_pack" for Baichuan model and "q_proj v_proj" for LLAMA model. Thanks to these settings, we finetune 7B-size LLMs on one A-100 80G GPU. The total finetuning process has 62 training steps and needs about 20 minutes.

We choose NDCG (Wang et al., 2013) and MAP (Carterette and Voorhees, 2011) as metrics in evaluating. And we utilize the Python interface of the official TREC evaluation tool (Gysel and de Rijke, 2018) to ensure the reliability of results.

5.2 Baselines

Referring to MTEB leader board[6], we select some top-ranking and common used models as baseline, including a traditional sparse retriever BM25 (Robertson et al., 2009), four powerful closed source API, text-embedding-ada-002 and text-embedding-3-large[7], baichuan-text-

[2]http://www.law-lib.com/

[3]https://zyp.yilianmeiti.com

[4]https://dean.pku.edu.cn/web/download.php

[5]https://platform.openai.com/docs/guides/text-generation

[6]https://huggingface.co/spaces/mteb/leaderboard

[7]https://platform.openai.com/docs/guides/embeddings

Method	Policy-Company				Case-Provision				Symptom-Drug				Objective-Course			
	NDCG		MAP		NDCG		MAP		NDCG		MAP		NDCG		MAP	
	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20	@10	@20
English Evaluation																
BM25	36.3	39.3	27.3	28.7	2.9	3.8	1.5	1.9	4.6	6.4	2.7	3.2	22.0	24.9	11.1	13.1
Voyage-lite-02-instruct	43.9	47.5	34.0	35.9	12.8	14.7	6.7	7.8	10.7	12.8	7.0	7.6	38.0	41.0	22.6	25.6
Text-embedding-ada-002	46.7	50.3	36.7	38.7	9.3	11.9	6.5	7.9	11.1	13.5	7.3	8.1	38.7	42.5	23.7	27.4
Text-embedding-3-large	**49.3**	**53.0**	**39.3**	**41.3**	20.1	21.0	14.2	14.4	10.2	12.7	6.1	6.8	44.3	48.0	28.0	32.2
E5-7B (Wang et al., 2024)	41.8	45.7	32.5	34.6	18.5	20.2	12.7	13.7	11.9	15.1	7.2	8.1	39.1	42.7	23.2	26.7
UAE (Li and Li, 2023)	46.6	50.2	36.4	38.3	5.9	7.5	3.1	4.0	8.4	11.0	5.2	6.0	42.1	46.2	26.8	30.6
BGE (Xiao et al., 2023)	43.6	47.7	33.8	35.8	9.5	11.2	5.8	6.6	8.2	10.0	5.1	5.6	39.3	43.2	24.2	27.7
GTE (Li et al., 2023b)	46.4	50.0	36.4	38.2	14.7	16.6	10.2	11.4	9.4	12.2	6.0	6.8	42.1	45.9	26.2	30.0
E5 (Wang et al., 2022)	38.8	42.6	29.6	31.4	7.4	8.5	5.2	5.9	6.4	8.7	4.0	4.5	34.7	38.9	21.0	24.3
INSTRUCTOR (Su et al., 2023)	45.5	49.6	35.6	37.8	17.5	19.1	12.2	12.6	10.0	13.0	6.1	7.0	37.3	42.2	22.2	26.2
URL (ours)	47.1	51.0	37.5	39.7	**29.3**	**30.4**	**20.2**	**21.3**	**13.8**	**16.1**	**9.7**	**10.4**	**48.2**	**53.4**	**31.1**	**37.4**
Chinese Evaluation																
BM25	23.7	26.3	17.4	18.4	10.2	11.1	8.0	8.3	5.5	7.1	3.4	3.9	18.7	20.9	10.4	11.7
Baichuan-text-embedding	26.6	30.5	18.8	20.4	25.0	26.9	16.9	17.5	10.3	12.6	6.7	7.3	16.0	17.5	7.2	8.2
Text-embedding-ada-002	47.3	51.3	37.2	39.4	17.5	19.9	10.5	11.2	7.0	9.1	4.3	4.8	34.7	38.4	20.7	23.8
Text-embedding-3-large	45.8	49.6	35.7	37.7	24.2	26.6	16.1	16.9	10.2	13.3	6.0	6.9	42.8	46.4	27.1	31.1
E5-7B (Wang et al., 2024)	37.1	41.2	27.3	29.3	16.7	18.6	10.3	10.7	10.6	13.8	6.8	7.7	40.3	43.6	25.0	28.3
GTE (Li et al., 2023b)	32.9	36.7	24.5	26.2	7.7	9.1	4.3	4.6	10.2	13.6	6.3	7.2	40.3	43.6	25.2	28.5
BGE (Xiao et al., 2023)	38.7	42.8	29.2	31.2	24.3	26.5	16.4	17.1	11.6	14.1	7.1	7.8	38.3	41.2	22.5	25.6
URL (ours)	**49.4**	**53.0**	**39.3**	**41.5**	**31.0**	**33.1**	**21.1**	**21.9**	**14.7**	**17.2**	**9.8**	**10.5**	**50.5**	**55.1**	**32.9**	**38.9**

Table 2: Main experiments in URLBench. Best/second-best performing score in each column is highlighted with bold/underline. It shows that URL is with better performance and universality than baselines.

Method	Policy-C.	Case-P.	Symptom-D.	Objective-C.
URL (ours)	**47.1**	**29.3**	**13.8**	**48.2**
Further Finetuning BERT-style Models				
BGE (Xiao et al., 2023)	36.7↓	7.8↓	7.8↓	40.9↑
GTE (Li et al., 2023b)	35.6↓	12.1↓	7.1↓	42.6↑
Using Task-aware Instructions on API				
Text-embedding-ada-002	36.9↓	7.8↓	10.1↓	35.6↓
Text-embedding-3-large	35.0↓	19.6↓	11.4↑	42.4↓

Table 3: NDCG@10 scores for further finetuning BERT-style models by our training data and using task-aware instructions on powerful API. It shows some fluctuations, yet still significantly under-performs URL.

embedding[8], and voyage-lite-02-instruct[9], some open source embedding models trained by large scale retrieval datas, uae-large-v1 (Li and Li, 2023), gte-large and gte-large-zh (Li et al., 2023b), bge-large-en-v1.5 and bge-large-zh-v1.5 (Xiao et al., 2023), e5-large-v2 (Wang et al., 2022), instructor-xl (Su et al., 2023). Additionally, we also evaluate e5-mistral-7b-instruct (Wang et al., 2024), which is a LLM embedder trained by contrastive learning and expensive high-quality multi-language datas.

5.3 Overall Results

In order to validate the performance of URL on versatile RKL tasks and illustrate the necessity of URL, experiments are conducted based on URL-Bench from three perspectives, including comparing performance of URL and baselines, further finetuning BERT-style models by our 1000 training data, and using task-aware instructions on API. Experiments show that URL is universal and accurate, simple finetuning on BERT-style models and direct adding instructions on API are not good solutions.

Firstly, by comparing the performance of open source embedding models, powerful closed-source API, and URL on the benchmark, as shown in Table 2, results demonstrate that although some baseline models achieve similar performance as URL on the certain dataset, none of them can consistently perform well across all the versatile tasks. The conclusion is that URL can perform high performance and great universality for RKL tasks.

Secondly, by observing performance of original BERT-style models trained further on our training data, as shown in Table 3, it shows some performance fluctuations, but BERT-style models are surely not enough to address universal RKL as URL. Conclusion is that BERT-style models cannot achieve same level of universal high performance as URL with only 1000 training data samples.

Finally, by observing the performance of API with task-aware instructions, as shown in Table 3, scores of both text-embedding-ada-002 and text-embedding-3-large become worse in most tasks when directly using task-aware instructions on it. The conclusion is that directly using API cannot achieve good and universal performance.

[8]https://platform.baichuan-ai.com/docs/text-Embedding
[9]https://docs.voyageai.com/embeddings/

Method	Components				Policy-Company		Case-Provision		Symptom-Drug		Objective-Course	
	Task-aware	Instruction	Generative	Contrastive	NDCG	MAP	NDCG	MAP	NDCG	MAP	NDCG	MAP
URL (ours)	✓	✓	✓	✓	47.1	37.5	29.3	**20.2**	**13.8**	**9.7**	**48.2**	**31.1**
Ablating Task-aware Instruction												
w/o Instruction	✗	✗	✓	✓	37.7	28.4	23.1	15.3	5.7	3.1	24.6	13.0
w/o Task-aware Instruction	✗	✓	✓	✓	**48.4**	**38.8**	**29.7**	19.7	11.7	8.1	40.0	24.5
Ablating Multi-view Learning												
w/o Learning	✓	✓	✗	✗	19.3	13.1	4.5	2.9	5.3	3.6	28.1	16.4
w/o Generative Reconstruction	✓	✓	✗	✓	46.1	36.6	28.2	19.7	12.7	8.9	46.4	29.1

Table 4: NDCG@10 and MAP@10 results of ablations. It shows that task-aware instructions and multi-view learning are great helpful for addressing RKL tasks universally and effectively.

In summary, through the aforementioned three perspectives of experimentation, the results show that URL is a reliable unified solution, and the framework is valuable and necessary for the development of RKL area because existing BERT-style models and powerful closed API cannot achieve performance close to URL through simple further training or task instructions.

5.4 Ablations for Task-aware Instruction and Multi-view Learning

Note that task-aware instruction and multi-view learning are two important modules. To validate their impact for URL framework, we conduct two types of ablation experiments about instructions and learning, results are shown in Table 4.

Firstly, by observing performance when no instructions are used and when task-aware instructions are not employed, with only a fixed instruction, it shows that using a single instruction is better than using none at all, but clearly not as universal as task-aware instructions. Conclusion is instructions significantly affect performance, and task-aware instruction is with important effect in URL.

Secondly, by comparing performance of LLM without training and LLM trained only via contrastive learning, it shows that raw LLMs perform badly, and multi-view learning shows better performance beyond contrastive learning. The conclusion is that original LLMs need training to apply to RKL tasks, and multi-view learning is a more effective training approach than contrastive learning.

In summary, task-aware instructions and multi-view learning show significant effect in URL framework, these two components are helpful to align LLMs with better ability for versatile RKL tasks.

5.5 Comparing URL and Conventional Models on Some Cases

To further investigate why URL performs better than conventional models, we analyze some claim-reference pairs that URL deals right but BERT-style models do not. And we find that URL is better mainly because of greater ability for knowledge-rich and versatile cases.

	Claim	Knowledgeable Reference
Policy-Company	Establish a number of **Internet of Things** technology laboratories ...	Company mainly engaged in the research of **MEMS sensors** ...
Objective-Course	This major focuses on systematic mastery of **robot engineering** ...	This course is to introduce the **feedback control** systems ...

Table 5: BERT-style models fail on these cases because linking is versatile and knowledgeable, rather than with fixed and simple semantic similarity. URL can handle correctly due to powerful knowledge and adaptability.

As shown in Table 5, the course about "feedback control" can support objective about "robot engineering", the company about "MEMS sensors" is affected by policy about "Internet of Things". These versatile cases are with deep knowledge rather than merely fixed superficial semantic correlation, thus BERT-style models are easy to fail. But URL can handle correctly due to powerful adaptability and knowledgeable semantic understanding ability, further improving multi-view learning is an effective method to align LLMs for RKL.

6 Conclusion

In this paper, we first define versatile referential knowledge linking tasks from a unified perspective. To address versatile RKL tasks efficiently and universally, we propose task-instructed representation compression framework and multi-view learning by transformed existing QA data. Furthermore, we introduce a new benchmark across versatile knowledge-rich tasks in various fields. According to the evaluation, URL can address versatile RKL tasks effectively and universally, perform better than powerful OpenAI Text Embedding and common-used open source models. This paper can promote intelligence to generate more authoritative and precise content in the era of LLMs.

7 Limitations

In this paper, we primarily investigate the scenario of fine-tuning LLMs with a small amount of data under parameter efficient conditions. We find that the aforementioned small-scale fine-tuning and limited data are sufficient to effectively transfer LLMs for universally addressing referential knowledge linking tasks. What remains unexplored in this paper is whether introducing larger-scale data and finetuning more parameters could achieve an even better model for universal referential knowledge linking tasks. This remains as our future work.

8 Ethical Considerations

Our study is with slim chance for ethical risks. On the one hand, we use LLMs for embedding generators rather than content generation. On the other hand, our training and evaluating data are mainly transformed from public safe source. The only possible risk is that we use help of gpt-3.5-turbo when constructing the benchmark. For this, we conduct manual review to ensure that the content generated by the API does not pose any moral hazard.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *ArXiv preprint*, abs/2108.13897.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *ArXiv preprint*, abs/2303.04226.

Ben Carterette and Ellen M. Voorhees. 2011. *Overview of Information Retrieval Evaluation*, pages 69–85. Springer Berlin Heidelberg, Berlin, Heidelberg.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846, Singapore. Association for Computational Linguistics.

William K Clifford. 1877. The ethics of belief. *First Published*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Charles Cole. 2011. A theory of information need for information retrieval that connects information to

knowledge. *Journal of the American Society for Information Science and Technology*, 62(7):1216–1231.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv preprint*, abs/2312.10997.

Shivanshu Gupta, Clemens Rosenbaum, and Ethan R Elenberg. 2023. Gistscore: Learning better representations for in-context example selection with gist bottlenecks. *ArXiv preprint*, abs/2311.09606.

Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An extremely fast python interface to trec_eval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 873–876. ACM.

Kailash A. Hambarde and Hugo Proença. 2023. Information retrieval: Recent advances and beyond. *IEEE Access*, 11:76581–76604.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. DialMed: A dataset for dialogue-based medication recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 721–733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv preprint*, abs/2311.05232.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *ArXiv preprint*, abs/2112.09118.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *ArXiv preprint*, abs/2307.16645.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models a better foundation for dense retrieval. *ArXiv preprint*, abs/2312.15503.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *ArXiv preprint*, abs/2309.12871.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *ArXiv preprint*, abs/2308.03281.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *ArXiv preprint*, abs/2310.08319.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jesse Mu, Xiang Lisa Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens. *ArXiv preprint*, abs/2304.08467.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron Courville. 2023. Language model alignment with elastic reset. *ArXiv preprint*, abs/2312.07551.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Jinyuan Wang, Hai Zhao, Zhong Wang, Zeyang Zhu, Jinhao Xie, Yong Yu, Yongjian Fei, Yue Huang, and Dawei Cheng. 2023. Csprd: A financial policy retrieval dataset for chinese stock market. *ArXiv preprint*, abs/2309.04389.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv preprint*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. *ArXiv preprint*, abs/2401.00368.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG type ranking measures. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 25–54. JMLR.org.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *ArXiv preprint*, abs/2309.07597.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *ArXiv preprint*, abs/2309.10305.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *ArXiv preprint*, abs/2309.11325.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. Language models are universal embedders. *ArXiv preprint*, abs/2310.08232.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

A Detailed Parameters in Multi-view URL Learning

Here are some supplementary explanation of the finetuning process for Section 5.1. We finetune LLMs by the 1000 datas for 2 epochs, setting max_length for both question and answer as 256, batch_size as 2, gradient_accumulation_steps as 16, learning_rate as 1e-4 and warmup_ratio as 0.08. For formulas in Section 3.2, we set α as 0.2, τ as 0.8, and prompt $\boldsymbol{x}_p$ as follows:

Generating reference base on claim embedding: According to the above text and instructions, these common questions in the Car domain can retrieve the following explanations:

Generating claim base on reference embedding: According to the above text and instructions, these explanations in the Car domain can be matched with the following common questions:

B Detailed Processing in Constructing URLBench

Here are some detailed processing for Section 4.

For policy-company task, existing dataset is a retrieval task about finding related policies based on a given company profile. Original setting can not directly reflect applications in stock decision. In this paper, we directly do some transformation, getting task that linking to related companies based on a given government policy.

For case-provision task, existing dataset is a classification task, legal case is as text and legal provision is as label. The number of label in original dataset is limited, we manually extend more provisions based on legal web.

For symptom-drug task, existing dataset is a classification task, patient-doctor dialogue is as text and drug name is as label. Firstly, we transform the dialogue to symptom description by the patient because doctor can already talk about the drug. Secondly, we add drug description based on medical web for each drug because drug name contains too little information, is not enough to support linking. Finally, we add more drugs according to the medical web.

For objective-course task, we collect dataset from a university enrollment handbook. Firstly, original file is in PDF format, we transform it to JSON by python tools. Secondly, there are some courses with similar introduction, i.e. Matrix Theory A and Matrix Theory B. We manually recognize these course and only keep one.

Case	Challenging	Easy
URL Success and BGE (Xiao et al., 2023) Fail	75%	25%
URL Success and Only Contrastive Learning Fail	82%	18%

Table 6: Two categories of cases that URL can handle correctly but other methods can not, which are claim-reference pairs requiring challenging knowledge or easy knowledge. In these cases, the proportion of challenging data is far greater than shallow datas.

C Task-aware instructions in Evaluating

Task-aware instructions for the four evaluation tasks are shown as follows:

Task-aware instructions for claim: Above text is description of a {government policy / legal case / patient symptom / training objective}. Based on your own knowledge, compress it into an embedding that can be used to search for relevant {company profile / legal provision / drug description / course introduction}. The embedding is:

Task-aware instructions for reference: Above text is a {company profile / legal provision / drug description / course introduction}. Based on your own knowledge, compress it into an embedding that can be used to match relevant {government policy / legal case / patient symptom / training objective}. The embedding is:

D Statistics for Cases URL Can Handle but Others Cannot

In order to further investigate why URL outperforms BERT-style models and the role of generative reconstruction, as shown in Table 6, we analyze some claim-reference pairs that our model deal right but others do not. With help of gpt-3.5-turbo[5], we annotate each claim-reference pair as data requiring challenging knowledge or easy knowledge, it shows that URL is better mainly because of stronger capabilities in datas requiring challenging knowledge.

In summary, URL effectively harnesses knowledge and reasoning capabilities of LLMs, and the generative reconstruction greatly keep the embedding with more knowledge when aligning the LLM to RKL tasks, thereby to a certain extent reducing the "alignment tax" (Noukhovitch et al., 2023).