

← Go to **ACL ARR 2024 October** homepage (/group?id=aclweb.org/ACL/ARR/2024/October)

Latent Factor Models Meets Instructions: Goal-conditioned Latent Factor Discovery without Task Supervision



*Zhouhang Xie (/profile?id=~Zhouhang_Xie1),
Tushar Khot (/profile?id=~Tushar_Khot1),
Bhavana Dalvi Mishra (/profile?id=~Bhavana_Dalvi_Mishra2),
Harshit Surana (/profile?email=harshits%40allenai.org),
Julian McAuley (/profile?id=~Julian_McAuley1),
Peter Clark (/profile?id=~Peter_Clark1),
Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1)*



📅 15 Oct 2024 (modified: 20 Dec 2024) 📁 ACL ARR 2024 October Submission 👁 October, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers 📄 Revisions (/revisions?id=C3vUF559Mh) © CC BY 4.0
(<https://creativecommons.org/licenses/by/4.0/>)

Abstract:

Instruction-following LLMs have recently allowed systems to discover hidden concepts from a collection of unstructured documents based on a natural language description of the purpose of the discovery (i.e., goal). Still, the quality of the discovered concepts remains mixed, as it depends heavily on LLM's reasoning ability and drops when the data is noisy or beyond LLM's knowledge. We present Instruct-LF, a goal-oriented latent factor discovery system that integrates LLM's instruction-following ability with statistical models to handle large, noisy datasets where LLM reasoning alone falls short. Instruct-LF uses~LLMs to propose fine-grained, goal-related properties from documents, estimates their presence across the dataset, and applies gradient-based optimization to uncover hidden factors, where each factor is represented by a cluster of co-occurring properties. We evaluate latent factors produced by Instruct-LF on movie recommendation, text-world navigation, and legal document categorization tasks. These interpretable representations improve downstream task performance by 5-52% than the best baselines and were preferred 1.8 times as often as the best alternative, on average, in human evaluation.

Paper Type: Long

Research Area: NLP Applications

Research Area Keywords: hierarchical & concept explanations, topic modeling, knowledge discovering

Contribution Types: NLP engineering experiment

Languages Studied: English

Response PDF: [📄 pdf \(/attachment?id=C3vUF559Mh&name=response_PDF\)](#)

Reassignment Request Action Editor: This is not a resubmission

Reassignment Request Reviewers: This is not a resubmission

Software: [📄 tgz \(/attachment?id=C3vUF559Mh&name=software\)](#)

A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: Yes

A2 Elaboration: Appendix section M

B Use Or Create Scientific Artifacts: Yes

B1 Cite Creators Of Artifacts: Yes

B1 Elaboration: We cited the libraries we used in Appendix E,F,G

B2 Discuss The License For Artifacts: Yes

B2 Elaboration: We explicitly states the models/libraries we use (Mistral/Huggingface) are open-source in Appendix E,F,G

B3 Artifact Use Consistent With Intended Use: N/A

B3 Elaboration: This is a research work, we use all open-source models/libraries for research purpose

B4 Data Contains Personally Identifying Info Or Offensive Content: N/A

B4 Elaboration: Non applicable: we use dataset open-sourced by others.

B5 Documentation Of Artifacts: Yes

B5 Elaboration: Appendix E,F,G and other appendix sections

B6 Statistics For Data: Yes

B6 Elaboration: Appendix E,F,G; section 6,7,8

C Computational Experiments: Yes

C1 Model Size And Budget: Yes

C1 Elaboration: Appendix E,F,G

C2 Experimental Setup And Hyperparameters: Yes

C2 Elaboration: Appendix E,F,G

C3 Descriptive Statistics: Yes

C3 Elaboration: section 6,7,8

C4 Parameters For Packages: Yes

C4 Elaboration: Appendix E,F,G

D Human Subjects Including Annotators: Yes

D1 Instructions Given To Participants: Yes

D1 Elaboration: Appendix I

D2 Recruitment And Payment: Yes

D2 Elaboration: Appendix I

D3 Data Consent: Yes

D3 Elaboration: Appendix I

D4 Ethics Review Board Approval: N/A

D4 Elaboration: We use commercial crowd-source platform, and the nature of our task hasn't needed ethical review in prior works.

D5 Characteristics Of Annotators: Yes

D5 Elaboration: Appendix I

E Ai Assistants In Research Or Writing: Yes

E1 Information About Use Of Ai Assistants: Yes

E1 Elaboration: Appendix K

Reviewing No Volunteers Reason: N/A - An author was provided in the previous question.

Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are only willing to serve as regular reviewers.

Preprint: no

Preprint Status: We are considering releasing a non-anonymous preprint in the next two months (i.e., during the reviewing process).

Preferred Venue: NAACL

Consent To Share Data: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: <https://github.com/acl-org/acl-style-files>) (<https://github.com/acl-org/acl-style-files>).

Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree

Submission Number: 1340

Discussion (/forum?id=C3vUF559Mh#discussion)

Filter by reply type... ▼

Filter by author... ▼

Search keywords...

Sort: Newest First

☰

☰

☰

-

=

☰

🔗

👁

Everyone

Submission1340...

Submission1340 Area...

Submission1340 Authors

15 / 16 replies shown

Submission1340...

Program Chairs

Submission1340...

Submission1340...

Submission1340...

Submission1340...

Submission1340...

Submission1340...

✕

Add:

Author-Editor Confidential Comment**Withdrawal**

Meta Review of Submission1340 by Area Chair ANqz

Meta Review by Area Chair ANqz 📅 06 Dec 2024, 08:29 (modified: 20 Dec 2024, 12:12)

👁 Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

📄 Revisions (/revisions?id=91aalfWgNt)

Metareview:

This paper introduces Instruct-LF, a goal-oriented latent factor discovery framework that combines the instruction-following capabilities of LLMs with latent factor models to process and interpret large-scale noisy datasets. The core idea is to utilize LLMs for generating goal-specific natural language property descriptions and then leverage clustering methods to discover hidden factors in the data, making the approach both interpretable and scalable.

Summary Of Reasons To Publish:

The strengths identified by all reviewers:

The integration of information-theoretic measures with structured latent factor models for clustering properties is unique in the LLM domain. The paper includes evaluations across three diverse tasks (movie recommendation, next-action prediction, document categorization) and supplements these with human evaluations, strengthening the validity of the results. The experiments are robust, with strong results compared to baselines and human evaluations reinforcing the efficacy of the method.

Summary Of Suggested Revisions:

Reviewer 5kqV raised the need to justify the choice of the base dense retriever model and discuss why alternatives like DPR or Contriever were not used.

Reviewer 5kqV recommended comparing the inference latency of the proposed method with previous work to highlight its advantages.

Reviewer pQCV suggested demonstrating how the method can reduce reliance on large LLMs and potentially use smaller models, and testing on more open-sourced LLMs to observe potential performance variance

Reviewer qKUw recommended adding a subsection on latent factor discovery and describing evaluation metrics like H@1 and H@5 in more detail.

Reviewer qKUw suggested exploring the connection between latent factor discovery and input data to assess its dependency on user goals.

Additionally, there are some writing suggestions for the authors to consider. For example:

Reviewer 5kqV suggested including a figure to visualize the data-property matrix and improving the details in Figure 2 for better clarity.

Reviewer pQCV suggested exploring related work in the concept bottleneck literature to enhance the interpretability layer and align with existing research.

Reviewer qKUw raised the need to (1) clarify methodology and notation, including terms like TC, Z, and C, as well as the loss function in Section 3. (2) specify the dataset used for training the retriever model.

Overall Assessment: 4 = There are minor points that may be revised

Best Paper Ae: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No


Author Identity Guess: 1 = I do not have even an educated guess about author identity.

Reported Issues: No

Add: **Author-Editor Confidential Comment**

A note re. review revision period and a gentle reminder

Official Comment

by Authors ( harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=~Julian_McAuley1), Peter Clark (/profile?id=~Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

 26 Nov 2024, 21:23 (modified: 02 Jan 2025, 08:07)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

 Revisions (/revisions?id=LpYrg2DXwN)

Comment:

Dear Reviewers,


As the discussion period where authors can post responses comes close to its end, we wanted to thank you again for your time and effort in providing constructive feedback for our work. Meanwhile, we wanted to send a gentle reminder regarding our responses - please do not hesitate to contact us if you have any additional concerns and/or questions. We look forward to further discussions with you.

Best Wishes, Authors

Add: **Author-Editor Confidential Comment**

General response from authors

Official Comment

by Authors ( harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=~Julian_McAuley1), Peter Clark (/profile?id=~Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

 25 Nov 2024, 09:23 (modified: 02 Jan 2025, 08:07)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

 Revisions (/revisions?id=96N9h7aOU7)

Comment:

Dear reviewers, we appreciate your time and effort in providing valuable feedback for our work. We are encouraged to find that you find our comprehensive automated/human evaluation a strength (5kqv, pQCV, qKUw, 43cV), our contribution well-justified (43cV), our framework clearly differentiates itself from prior works/novel (pQCV, qKUw), our method effective (5kqv) with good results compared to baselines (qKUw), and our details for reproducibility helpful (5kqv). We will also open-source our code upon acceptance.

While it was mentioned that our work is well-written and well-explained (pQCV), two of the reviewers proposed some presentation suggestions, mainly re. section 3 (qkUw, 5kqV), which we will improve by incorporating a few corresponding updates (details in individual responses). Re. the embedding model "all-MiniLM-L6-v2" (qkUw and 5kqV) - we initially chose this model since it is widely used (95M downloads on Huggingface), light-weight, and has solid performance. Our goal here is to verify that Instruct-LF can perform well with a general text embedding model. To further verify this, we tested Contriever on Inspired with GPT-3.5-Turbo-generated properties, and the trend is consistent - see individual responses for additional details.

Add: **Author-Editor Confidential Comment**

Official Review of Submission1340 by Reviewer 5kqV

Official Review by Reviewer 5kqV 📅 23 Nov 2024, 18:46 (modified: 20 Dec 2024, 12:12)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 5kqV, Commitment Readers

📄 Revisions (/revisions?id=oNIiUqMDlh)

Paper Summary:

This paper proposes a goal-oriented latent factor discovery system that combines the instruction following ability of LLMs with statistical models to handle large-scale noisy datasets. The proposed method, Instruct-LF, uses LLMs to generate document properties that are related to the specific goals, where co-occurring properties can be clustered to represent hidden factors. It trains a dual-embedding model to perform data-property link prediction. In the latent factor discovery step, Instruct-LF adopts Linear Corex so that the learned latent variables can also be clustered for correlated properties. Evaluations show that Instruct-LF out-performs baselines on movie recommendation, next-action prediction, and fine grained document categorization tasks.

Summary Of Strengths:

1. The proposed method showcases the effectiveness of combining LLM reasoning with statistical algorithms and can be potentially applied in more tasks.
2. The evaluation is comprehensive and includes three different tasks. Further, the authors include human annotators to compare the quality of the different methods.
3. The authors provide implementation details and prompts used for different tasks, which is useful for reproducibility of the work and future studies.

Summary Of Weaknesses:

1. In Section 3, the data-property matrix is a bit hard to understand at the first glance and it would be helpful to include a figure to visualize the concept. The Figure 2 is not illustrative and needs more details and descriptions.
2. The selection of the base dense retriever model is not well justified as there are numerous well-known dense retriever models like DPR and Contriever. It is unclear whether a base model matters since the authors are performing task-specific fine-tuning.
3. In line 265, the authors mention a drawback of previous work. It would be helpful to understand the superiority of the proposed method compared to previous work if the inference latency is compared.

Comments Suggestions And Typos:

Please address the weakness above.

The attached software cannot be properly opened.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 3.5

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Best Paper: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Official Comment by Authors

Official Comment

by Authors (harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=-Julian_McAuley1), Peter Clark (/profile?id=-Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=-Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

25 Nov 2024, 09:25 (modified: 02 Jan 2025, 08:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 5kqV, Commitment Readers

Revisions (/revisions?id=XfAYUIUQkc)

Comment:

Thank you for your valuable feedback on our work. With regards to the concerns you have:

- [1] Re. sec 3: Thanks for mentioning this - we will add a figure where there's a matrix illustration (like the one in Fig. 2-(2)), with a concrete example for a data point and a property. Re. fig 2 - we will fit some concrete examples under captions, and \Cref each its subcomponent to its corresponding subsection, so readers can easily pair the sub-figures with corresponding sub-sections
- [2] Re encoder model: Thanks for the suggestion. We chose this model since it is a lightweight, widely used (95M downloads on Huggingface Transformers library, and a default model for Sentence Transformers library's code example) embedder model. The goal here is to show that our framework can perform well with a "reasonable" text embedding model.
 - When we try the Contriever model, the $H@{1, 5, 20}$ is 3.3/15.8/25.0 on Inpsired (with GPT-3.5-Turbo generated properties). The performance is slightly higher than that from all-minilm-l6-v2, and the trend where Instruct-LF has better performance is consistent.
- [3] Re L265 and efficiency of our framework: Thanks for the suggestion. Due to the space limit, we included this detailed comparison in Appendix K - we will add a pointer at line 265 to make this information more accessible to future readers.

Please let us know if there are any further questions/concerns you have, we look forward to discussing with you.

Add: **Author-Editor Confidential Comment**



Official Review of Submission1340 by Reviewer pQCV

Official Review by Reviewer pQCV 📅 22 Nov 2024, 10:32 (modified: 20 Dec 2024, 12:12)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pQCV, Commitment Readers

📄 Revisions (/revisions?id=UMIHtq5ug4)

Paper Summary:

The paper presents a novel way of discovering relevant topics or concepts in a learned latent space. The idea is to take advantage of the instruction-following abilities of LLMs in conjunction with existing statistical gradient-based methods (latent factoring models) to discover hidden factors from the given dataset. Previous works on mining insights from data using LLMs have heavily relied on their reasoning capabilities. This works aims to ease that dependency by combining LLM with gradient-based methods. Intuitively, this removes a lot of subjectivity coming from frameworks that only rely on LLMs.

Broadly divided into two components, the proposed technique starts with generating goal-related Natural Language property descriptions of the given text which is then used to create a data-property matrix that is learned to obtain compatibility scores between LLM-generated properties and the input data. The next step aims to form clusters of the properties by their correlations to obtain higher-level latent patterns. The authors here use a state-of-the-art latent structure learning model to achieve this. The authors also claim to obtain a sense of interpretability since each property is mapped with a natural language description.

The authors evaluate their models by testing their technique on downstream tasks and study the usefulness of the discovered latent properties. Other than showing the quantitative efficacy of their work as compared to prior work, the authors also perform human evaluation that it is better to learn a dual-embedding model to map data to property instead of prompting LLMs to do the same. Instead, LLMs are much better at simply generating relevant properties for the input.

Summary Of Strengths:

1. The paper is well-written with each component of the proposed technique being well-explained.
2. The paper clearly differentiates their method with previous work.
3. The use of gradient-based methods to cluster and learning of data-property matrix removes a lot of subjectivity from the process that may arise in simple prompt-based LLM methods to do the same.
4. Apart from sufficient use of downstream tasks to evaluate, the authors also report human evaluations and present a strong argument against the use of LLM to generate data-property associations.

Summary Of Weaknesses:

1. While the authors argue that prior LLM-based methods rely heavily on the reasoning capabilities of LLMs, the authors seem to use LLMs that are known to be very performant in general. Intuitively, reducing dependence on the LLM should also enable the use of possibly smaller LMs as well.
2. Following up on the previous point, I felt that there weren't enough open-sourced LLMs tested where there is a chance of observing high variance in results.

Comments Suggestions And Typos:

I felt that the work was somewhat related to the **concept bottleneck** literature which aims to induce an interpretability layer by defining a set of description properties. I didn't see any mention of the literature in the paper. Though it's not super important, the authors may benefit from further studying this literature to push their research. Looking forward to hear from the authors.

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 3.5

Best Paper: No

Limitations And Societal Impact:

I believe the limitations of their work have been sufficiently discussed.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Official Comment by Authors

Official Comment

by Authors (👁️ [harshits@allenai.org \(/profile?id=harshits@allenai.org\)](#), [Julian McAuley \(/profile?id=-Julian_McAuley1\)](#), [Peter Clark \(/profile?id=-Peter_Clark1\)](#), [Bodhisattwa Prasad Majumder \(/profile?id=-Bodhisattwa_Prasad_Majumder1\)](#), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

📅 26 Nov 2024, 21:24 (modified: 02 Jan 2025, 08:07)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pQCV, Commitment Readers

📄 Revisions (/revisions?id=2V8FWh55UJ)

Comment:

We are glad that we can address your concerns. Thank you again for providing valuable feedback on our work.

Add: **Author-Editor Confidential Comment**



Official Comment by Authors

Official Comment

by Authors (👁️ [harshits@allenai.org \(/profile?id=harshits@allenai.org\)](#), [Julian McAuley \(/profile?id=-Julian_McAuley1\)](#), [Peter Clark \(/profile?id=-Peter_Clark1\)](#), [Bodhisattwa Prasad Majumder \(/profile?id=-Bodhisattwa_Prasad_Majumder1\)](#), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

📅 25 Nov 2024, 09:29 (modified: 02 Jan 2025, 08:07)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pQCV, Commitment Readers

📄 Revisions (/revisions?id=f2TJ5luh2Y)

Comment:

Thank you for your valuable comments on our work. We address your concerns and suggestions as follows:

Weakness section: LLM baselines:

- We wanted to clarify that when we say prior work (TopicGPT from Pham et al.) relies heavily on the capability of LLMs, this is to say that their framework cannot function with models weaker than proprietary models from OpenAI. As discussed in the original TopicGPT paper by its authors (“topic generation is too complex for all LLMs we tried other than GPT-4”): TopicGPT requires models to handle long and complex instructions in the input, which is very demanding on the backbone model used.
- To this end, Instruct-LF is the first work in this direction that can work with an open-source (7b) LLM, which is not as performant as proprietary models from OpenAI.

- In fact, we chose the Mixtral-7b model because this is the model discussed in TopicGPT, where it was mentioned that this model fails to produce good properties in TopicGPT framework.
- Meanwhile, we added results using Llama-8b, Llama-3b, and Llama-1b. Due to the time limit during rebuttal (since we'd need to host a local model and thus are constrained by our resources), we report performance on Inspired dataset. The H@{1,5,20} are as follows (and pasted Mixtral performance for reference):
 - llama-8b: 3.8; 12.5; 26.4 (comparable to other models)
 - llama-3b: 4.3; 11.5; 25.9 (comparable to other models)
 - llama-1b: 4.3; 12.5; 18.75 (slightly lower performance to other models, particularly on H@20)
 - Mixtral-7b: 4.8; 11.5; 24.0

From the above result, it can be observed that our framework can indeed maintain good performance with other model architectures and smaller models.

Comments/suggestions on Concept Bottleneck Models:

- Thanks for the suggestion. We share your feelings about this being relevant. While we initially included the discussion about concept bottleneck models in the appendix (L1039), we will add a short discussion on these in the related work section (in particular, the key differences section, and point to the appendix for more discussion).

Add: **Author-Editor Confidential Comment**



Official Comment by Reviewer pQCV

Official Comment by Reviewer pQCV 📅 26 Nov 2024, 01:53 (modified: 02 Jan 2025, 08:07)
👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pQCV, Commitment Readers
📄 Revisions (/revisions?id=GikoSMIVBA)

Comment:

I thank the authors for addressing the limitations pointed out regarding LLM baselines. I am increasing the soundness score from 3.5 to 4.

Add: **Author-Editor Confidential Comment**



Official Review of Submission1340 by Reviewer qKUw

Official Review by Reviewer qKUw 📅 21 Nov 2024, 23:58 (modified: 20 Dec 2024, 12:12)
👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qKUw, Commitment Readers
📄 Revisions (/revisions?id=irsnJraCqy)

Paper Summary:

The paper brings latent factor models to the area of LLMs, which is slightly weak on theory, but very strong on experiments. The authors adopt a method proposed in Steeg et al into the LLM space, by first proposing a input-property matrix that can then be used for clustering of properties. The clustering using information theoretic measures is quite unique. So the idea is to use 3 steps:

1. Get goals from the user
2. Using these goals get relevant properties from an LLM (authors to clarify Q1, they mention properties are derived from goals in some places, and derived from one data point in other places)

3. Obtain a subset of these properties using an information-theoretic clustering approach. Call these properties Z .
4. Let Z be the clustered properties, and X be the inputs. The authors use a sentence embedding model obtain a Z - X (property-input) map through a dot product (similar to cosine score)
5. Such a matrix is then used in downstream tasks.

Due to strong experimental results, but weak writing (especially Section 3), I would score it as a weak accept. Happy to increase my score given an improvement of Section 3 (suggested edits/rewrites below).

Summary Of Strengths:

- Strong experiments! Really appreciate that the authors have taken up human evaluations to strengthen their work. Also good evaluations and results compared with baseline methods.
- The idea of using an LLM to propose properties for topics is not new. But I have not seen clustering of topics, by merging information-theoretic measures and structured latent factor models in the area of LLMs.

Summary Of Weaknesses:

- More studies on link prediction may be in order. What are the other models tried? Any ablations on effectiveness of "s/all-MiniLM-L6-v2"?
- While several ablations on property proposals have been made ablations on property scores from different parameterized models (different encoder models) may be useful.
- The authors claim their work combines LLM reasoning with gradient based methods. But I didn't understand the paragraph from 331 to 346. What is C , what is Z , what is TC ? why are these not defined before use?
- Given the goal in line 343, what is the actual loss function? is it sum over all i , $TC(Z | C_i)$? What about the other expression? This paragraph and subsection is quite opaque.
- How does loss function in line 343 compare with a PCA type approach on the embeddings of C ? Is that worth looking at? What about an ablation where an LLM looks at all properties and provides a subset of clusters on its own?
- Can latent factor discovery be linked to input data X ? Currently as I read it, it is independent, and only dependent on user goals.
- Please consider changing the title. What does this have to do with instructions? What about: "Goal-conditioned Latent Factor Discovery without Task Supervision using Latent Factor Models"?

Comments Suggestions And Typos:

Suggestions:

- A subsection on Latent Factor (the classical algorithm) somewhere in the paper may be due.
- The term "goal-oriented data transformation" is not sufficiently explanatory or helpful. Why not simply say that you learn property scores for the input sentences?
- Similar with data property link prediction. Calling input sequences as data points does not contribute to the understanding of the work, especially when used before Section 3. In section 3 these are clarified through the methodology and the equations, but a suggestion to the authors is to reconsider some of the terminology used in the paper. The intended audience of the paper is likely to be those interested in LLM property predictions, and not those in the LDA world -- and hence terminology from that world may not be directly relevant here.
- Please start paragraph L331 with a statement like given clusters $c_i \in C$, the goal is to learn $z_j \in Z$, such that $\|Z\| \ll \|C\|$ and $Z \subset C$. A line like this would clarify what you want to do more than the explanation currently given.
- Please describe $H@1$, $H@5$ etc (Table 1).

Comments:

- "frequently requires reading-the-tea-leaves interpretations" Wonderful to see language in otherwise drab papers!
- I like the work, but I believe entire Section 3 could use another pass, or partial rewrite.

Questions:

- "our property proposal step only prompt LLMs to document detailed attributes of a single data point" But in the previous step you say that each column is a property related to the user goal. Should this not be independent of the data points? If one data point is being selected, how is it selected?
- In equation 1, why not use cosine similarity?
- Line 294 what is the data set used to train this retriever for topics?
- What does TC stand for (equation 4)? Ok I opened Steeg and it looks like it's Total Correlation.
- Equation 4 refers to $TC(X)$. I believe X is a dummy variable here, but previously defined X was a high dimensional input space. Please clarify the use of X here.

- Does the loss function (described in lines 331-351) require any input data at all, or just the knowledge of clusters C?

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 2.5

Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

Best Paper: No

Limitations And Societal Impact:

Yes

Ethical Concerns:

There are no concerns with this submission.

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Response from authors (3/3)

Official Comment

by Authors (harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=-Julian_McAuley1), Peter Clark (/profile?id=-Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=-Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

25 Nov 2024, 10:30 (modified: 02 Jan 2025, 08:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qKUw, Commitment Readers

Revisions (/revisions?id=Fm1h02CjeU)

Comment:

- [13] "our property proposal step only prompt LLMs to document detailed attributes of a single data point" But in the previous step you say that each column is a property related to the user goal. Should this not be independent of the data points? If oone data point is being selected, how is it selected?
 - The property proposal step is actually dependent on each input data point (document) and the users' instruction (goal description), as we wrote in L244: "we prompt an LLM to generate a property that describes the goal-oriented properties of this data point".
 - This means the users' goal is part of the input to LLM, which is illustrated in Figure 1 (b). See Tables 10, 11, and 12 for concrete examples.
 - We do this for every data point in the dataset
- [14] In equation 1, why not use cosine similarity?
 - Prior work proposing the original neural matrix factorization setting (He et al., 2017 at L277) optimizes for dot product, and we use the dot similarity to stay consistent with prior work.
 - But you are right that cosine similarity would work, too (depending on which similarity function to optimize for when fine-tuning the retriever) - we will add a footnote to L280 stating cosine similarity is a

potential alternative.

- [15] Line 294: what is the data set used to train this retriever for topics?
 - This is probably related to the answer for [13] above
 - Each “x” is a document, and each “c” is a property generated by LLM. A positive example is when a particular “c” is generated from “x”, while a negative example is obtained via negative sampling. Specifically, the loss function we use does in-batch negative sampling, where for a pair document and property in a batch during training, and other properties sampled in that batch is negative samples w.r.t. the document.
 - This is what we are referring to when we say “to predict whether a property is generated from a data point in the 289 property proposal stage” in L288 - we will rephrase this to directly refer to “c” and “x” to avoid confusion.
- [16] What does TC stand for (equation 4)? Ok i opened Steeg and it looks like it's Total Correlation.
 - We realized it might be better to say just Total Correlation than multivariate mutual information we currently use, so capital letters match. We will update L344.
- [17] Equation 4 refers to TC(X). I believe X is a dummy variable here, but previously defined X was a high dimensional input space. Please clarify the use of X here.
 - Thanks for the catch. We will add a clarifying sentence.
- [18] Does the loss function (described in lines 331-351) require any input data at all, or just the knowledge of clusters C?
 - It requires input data, in that we need an estimated occurrence of each property across the whole dataset.
 - Note that C is related to properties (see the answer to [3]) - the input is the data-property matrix with the estimated property scores filled in.

Please let us know if there are any further questions/concerns you have, we look forward to discussing with you.

Add:

Author-Editor Confidential Comment



Response from authors (2/3)

Official Comment

by Authors (harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=-Julian_McAuley1), Peter Clark (/profile?id=-Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=-Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

25 Nov 2024, 10:24 (modified: 02 Jan 2025, 08:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qKUw, Commitment Readers

Revisions (/revisions?id=AL4jp44SyT)

Comment:

Suggestions section:

- [8] A section on the classical Latent Factor algorithm may be due.
 - Thanks for the suggestion. We agree that including an introduction to the classical latent factor algorithm is helpful - we will add an additional introduction to the appendix and point to it in the related work section.
 - However, Instruct-LF, while trying to address tasks similar to those of these prior algorithms, is not a direct methodology modification. Thus, we currently opted to discuss latent factor discovery algorithms as relevant works (in section 2), since we do not need to “set up” the classical algorithm before a reader can understand instruct-LF (unlike papers that are direct methodological extensions of prior works).
- [9] Re. the term goal-oriented data transformation, why not simply say we learn property scores for the input sentences Thanks for the suggestion. We agree that the suggested phrase is helpful in explaining our concept.

- We do, meanwhile, think there is value in stating the first step is “goal-oriented data transformation”, which explains the aspect of the algorithm that “amplifies” features in the original dataset by turning each unstructured dataset into a data-property “tabular” matrix, where each feature is potentially useful w.r.t. the users’ goal.
- In contrast, even if we only say we “ learn property scores for the input sentences/documents”, we’d still have to explain what a property is and what a score represents (currently line 194+)
- We will incorporate the suggested expression by stating “...and transform the input unstructured dataset into a data-property matrix by learning a property-compatibility score for each document” in line 197+.
- [10] Calling input sequences as data points does not contribute to the understanding of the work, especially when used before Section 3. In section 3 these are clarified through the methodology and the equations, but a suggestion to the authors is to reconsider some of the terminology used in the paper. The intended audience of the paper is likely to be those interested in LLM property predictions, and not those in the LDA world.
 - Thanks for the suggestions; we will add an explanation when we refer to data point prior to section 3
 - Specifically
 - L72: “sampled data points” -> “sampled documents”
 - L78: “from data” -> “from documents”
 - L107: “based on input data point” -> “based on input documents”
- [11] Adding a line using formal notations in L331
 - Thanks for the suggestion; we will update accordingly
- [12] Re. H@K
 - Thanks for the catch - these are Hit@K; we will add a notation in the caption

Add: **Author-Editor Confidential Comment**



Response from authors (1/3)

Official Comment

by Authors (harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=~Julian_McAuley1), Peter Clark (/profile?id=~Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=~Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

25 Nov 2024, 10:22 (modified: 02 Jan 2025, 08:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qKUw, Commitment Readers

Revisions (/revisions?id=sGYfdBQSiq)

Comment:

Thank you for reviewing our work; we appreciate your insightful and detailed feedback. For your convenience, we numbered each bullet in your original comments and addressed them as follows:

On Weakness:

- [1,2] Additional ablations on encoder in property score prediction/other models tried in link prediction:
 - Thanks for the suggestions. We will add the suggested ablations but note that we don't expect the findings to be too different. We used a relatively small model (smaller than BERT-small) that is popularly used (95M downloads in Huggingface) to verify our core idea. This shows that our general formulation can get much better results than using LLMs or topic models alone. Other models will likely further validate our findings.
 - In particular, when we try the Contriever model, as suggested by another reviewer, the H@{1, 5, 20} is 3.3/15.8/25.0 on Inpsired (with GPT-3.5-Turbo generated properties). The performance is slightly higher than that from all-minilm-l6-v2, and the trend where Instruct-LF has better performance is indeed consistent.
- [3] Why is C/Z not defined before? Writings of L331+:

- Sorry about the confusion - while we defined these terms in lines 206/215/221, we will cross reference these definitions in section 3.2 for clarity.
- [4] Actual loss function
 - Thanks, we agree that it'd be helpful to have the actual loss function in the paper.
 - Let $\nu_{C_i|Z}$ be the conditional mean of C_i given Z under the factorization $p(c_i|z) = p(c_i)/p(z)\prod_j p(z_j|c_i)$, which is implied by the modular $TC(Z|C_i) = 0$ constraint, the loss function that Linear Corex (Steeg's method) optimizes for is:
 - $\sum_{i=1}^p Q_i \frac{1}{2} \log E [(C_i - \nu_{C_i|Z})^2] + \sum_{j=1}^m \frac{1}{2} \log E [Z_j^2]$
 - We will add this information to the appendix and link to the explanation at eq. 4 so readers can easily access this information.
- [5] Re. ablations for, e.g., PCA and LLM-based cluster proposal
 - We wanted to start by clarifying that our primary goal here is to verify the merit in combining LLM and gradient-based optimization, which our particular initiation of the framework using Linear Corex (Steeg et al.) proves. We do not want to deny the possibility of discovering a better formulation in the future, but to this end, finding the best gradient-based method for latent factor discovery frameworks aligns more with an extension of this work than the weakness of the current work.
 - In the meantime, in our framework, a set of properties is proposed based on each document, and thus, the number of total observed properties (all properties that have ever been proposed) can be high.
 - For example, in the Alfworld dataset, 41k unique properties from which to discover latent factors are observed (see line 1017+ for more details).
 - Thus, we mentioned that the reason why we use Linear Corex is that it scales better to large amounts of observed variables (compared to, e.g., PCA, which is already verified as a baseline in the Steeg paper): we had a brief note on line 338+ "...that scales well with input dimensionality."
 - For similar reasons, recent work shows LLM ignore content in long context (Lost in the Middle: How Language Models use Long Context, ACL 24), and thus, it is unclear how to use LLM to process this long list of properties to generate clusters. Further, after proposing the clusters, we'd also have to find a way to link each property to the cluster (since LLMs cannot easily generate, e.g., a series of clusters from 41k properties). Thus, LLM-based property-based clustering is actually not an obvious baseline for the problem to be tackled here.
- [6] Can latent factor be linked to data X?
 - Yes, we could see the "compatibility scores" of each property w.r.t. in each document (data point using our current terminology) using the estimated property score in the data-property matrix. Since each property will be assigned to a latent factor, we can thus also see the most salient latent factors w.r.t. each document.
- [7] Re. alternative titles: why is "instruction" in the title
 - Thanks for the suggestion. W.r.t. the use of the word "instruction", we want to clarify that in our framework, users' goals written in natural language are a form of instruction (see concrete example in Tables 10, 11, and 12). This notion is similar to Instructor (One Embedder, Any Task: Instruction-Finetuned Text Embeddings, Su et al., ACL 23) and INBEDDER (Answer is All You Need: Instruction-following Text Embedding via Answering the Question, Peng et al, ACL 24), in that there is a model that adapts its behavior based on user-specified purpose in natural language.

Add:

Author-Editor Confidential Comment

Official Review of Submission1340 by Reviewer 43cV

Official Review by Reviewer 43cV 📅 19 Nov 2024, 02:29 (modified: 20 Dec 2024, 12:12)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 43cV, Commitment Readers

📄 Revisions (/revisions?id=WcLepBePpf)

Paper Summary:

The paper proposes Instruct-LF, a latent factor discovery framework that extracts latent concepts from unstructured text related to the user's instructions given in natural language. The approach combines instruction-following ability of LLMs and the scalability of gradient-based latent factor models. The approach is evaluated against multiple baselines for multiple datasets and tasks showing an improvement in performance for the tested tasks.

Summary Of Strengths:

1. Very well written paper and well justified contributions
2. Detailed analysis and evaluation, including human-based evaluation showing advantages of the approach versus baselines

Summary Of Weaknesses:

A lack of discussion on limitations of the approach related to applicability to domains/tasks

Comments Suggestions And Typos:

It could be nice to include a little bit about generalisation of approach to different domains/tasks

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.

Best Paper: No

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**



Official Comment by Authors

Official Comment

by Authors (harshits@allenai.org (/profile?id=harshits@allenai.org), Julian McAuley (/profile?id=-Julian_McAuley1), Peter Clark (/profile?id=-Peter_Clark1), Bodhisattwa Prasad Majumder (/profile?id=-Bodhisattwa_Prasad_Majumder1), +3 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission1340/Authors))

25 Nov 2024, 10:32 (modified: 02 Jan 2025, 08:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 43cV, Commitment Readers

Revisions (/revisions?id=OBM5IRjtzN)

Comment:

Thank you for your valuable insight into our work. Regarding your concern about the discussion on limitations, we will add the following sentence to the end of the limitations section (L648):

- Finally, our framework assumes LLM can propose or extract properties from input textual documents and thus cannot generalize to domains where input documents are beyond LLMs' pre-training distribution.

Please let us know if there are any further questions/concerns you have, we look forward to discussing with you.

Add:

Author-Editor Confidential Comment

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[Frequently Asked Questions](#)

[\(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

[Feedback](#)

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors). © 2025 OpenReview