Querying the Near Future: An Exploratory Study of Interpretable Concept Spaces from Multi-Token Prediction Heads

Anonymous ACL submission

Abstract

Medusa based Multi-token prediction (MTP) language models expose several independent future-token heads after a single forward pass. This paper explores whether those heads can be used not only for generation, but also as a lightweight mechanism for constructing interpretable concept spaces over documents. Given a text and a list of concepts, we save every output head's score and assign to each concept a score by reading the probability assigned to its first future tokens by the model's MTP heads. This produces a dense document-by-concept matrix in which every column is named by a human-readable concept. Unlike conventional embeddings, the representation is immediately query-able: one can ask to which degree a document is associated with any concept in or even out of the vocabulary since the space is also modulable, as we can query for any new concept within the limit of the number of future tokens available. We conduct a proof-of-concept study on AstroConcepts, a multi-label corpus of with over 1000 keywords and show that MTP heads capture structured interpretable concept-level information while performing similar to the same base model.

1 Introduction

Multi-token prediction modules are usually treated as accelerators. We ask them to speculate candidate text and then verify it (Leviathan et al., 2023) , however, architectures such as MEDUSA (Cai et al., 2024) make another interface possible. Instead of predicting only the next token, they expose several future-token distributions after one forward pass. A model with four independent MTP heads can estimate, in one pass, the likelihood of any short continuation up to four tokens into the future.

This property is very interesting for zero-shot classification as it is an area where LLMs often perform well in (Wang et al., 2023), however, to get better results, having semantically relevant labels (Gao et al., 2023) which are longer would require having multiple forward passes to generate.

Especially with multi-label classification, the compute required explodes as it needs to compute the number of tokens required for each label multiplied by the number of labels to get. This can become expensive fairly quickly as some tasks such as keyword generation can require more than ten long specific keywords (Chalkidis et al., 2019). This method would allow use of LLMs for classification at a fraction of the cost for those extreme types of problems.

For every candidate concept, the model provides a score by multiplying the probability of each independent token that composes the concept. Repeating this over a vocabulary gives a document-by-concept matrix. Each dimension is interpretable because it is tied to a text string chosen by the researcher. This is different from an opaque embedding dimension: the axis is named before evaluation, and is flexible as we can add as many keywords to the embedding space as long as we have the multiple heads outputs in memory.

The values cannot be interpreted in isolation, as they do not represent absolute quantities, but can be compared to other values obtained in the same manner. A conceptual space can be useful for mapping the evolution of social media users, identifying clusters, and analyzing their trajectories within the space, as well as their splits and other structural changes.

AstroConcepts (Alkan et al., 2026) is a dataset that contains scientific abstracts labeled with technical keywords, a fixed controlled vocabulary. Each Abstract has up to 12 keywords. There are 1,864 observed labels in the validation vocabulary, labels are imbalanced, and exact author-assigned concepts are not always the only reasonable description of a paper. A document about exoplanet atmospheres may also be legitimately related to exoplanets, planetary atmospheres, atmospheric composition, spec-

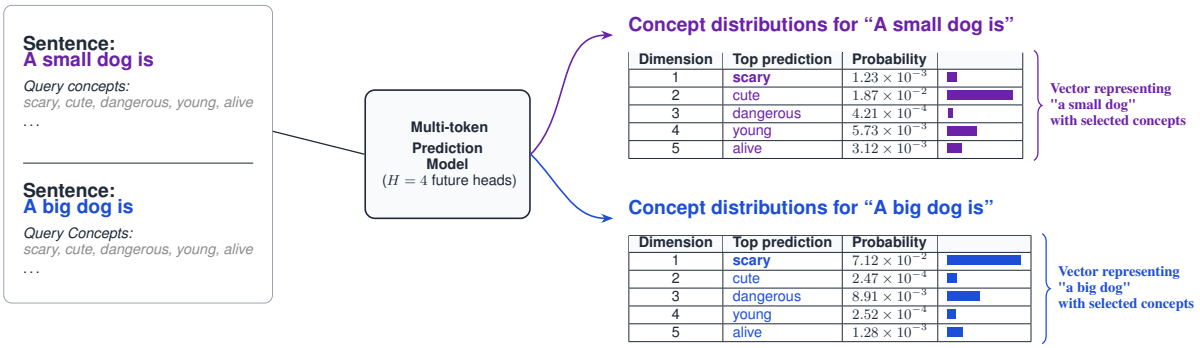

Figure 1: Illustration of concept querying through multi-token prediction heads. After defining the concepts, and tokenizing them, we obtain a vector where each dimension is interpretable and human selected.

troscopy, or observational techniques. The UAT hierarchy lets us distinguish exact keyword prediction from broader semantic branch recovery.

This paper does not intend on beating State-of-the-art, we do not claim that raw MTP scoring beats LLMs systems such as DeepSeek (Liu et al., 2024) combined with trained classifiers such as astroBERT (Grezes et al., 2021) as a candidate generator. Those systems are optimized for exact label prediction. Instead, we show that MTP scoring creates an interpretable concept space that reflects document content at a useful semantic granularity with very low compute requirements.

2 Related Works

Interpretable embeddings have been a subject of research and multiple architectures attempt to do just that. (Benara et al., 2024) by asking multiple questions they generate question-answering embeddings (QA-Emb), where each dimension is the answer of a yes/no question. However, that embedding space is not continuous as the values for each dimension is either 0 (No) or 1 (Yes).

(Tang et al., 2026) combines the same method as above with but by grounding the questions with an ontology. This allows to be able to get semantic relations from the extremely precise questions (atomic questions) which make each dimension clear. This method stays computationally expensive.

Wang et al. (2025) uses documents as anchors in the space to understand what a new point is within those referential points. This method does not allow to navigate the space with attention to certain specific very attributes.

Since most recommender systems use embeddings, explainable recommendation systems are also similar to our method. Sabouri et al. (2025) uses LLMs to generate a description of each user with their short and long term preferences, and only after that, embed those texts. Even though the embedding is not interpretable, this method circumvents the need for it by making explicit texts that are only then embedded. Thus every point in the embedding space has a textual description.

3 Multi-token prediction as a query interface

A standard autoregressive language model gives a distribution over the next token. A multi-token prediction model exposes multiple future-token heads. If the model has heads $h = 1, \ldots, H$, then after processing a prompt x, it produces distributions for tokens at offsets $1, \ldots, H$. In generation, these heads can help train or decode future positions. In this work, we use them differently: we treat them as a scoring device for candidate strings.

Given a candidate concept c, tokenize it into $t_1, \ldots, t_m$. With H future heads, we score up to $L = \min(m, H)$ tokens:

$$s(x, c) = \sum_{i=1}^{L} \log p_i(t_i \mid x), \qquad (1)$$

where p_i is the distribution from the i-th future head after the prompt. This score is not a full language-model probability for long phrases. It is a truncated future-continuation score. Concepts with more tokens are penalized by additional log-probability terms, and concepts whose important distinguishing word appears after the head window are only partially observed.

To reduce these effects, we compute normalized variants of the matrix. The most useful one in our experiments is per-concept percentile normalization: for each concept column, scores are ranked

across documents and mapped to $[0, 1]$. This asks a relative question: *among all documents, where does this document stand for the concept* $\boldsymbol{c}$*?*

3.1 AstroConcepts and UAT hierarchy

AstroConcepts is a multi-label classification corpus of astrophysics abstracts labeled with UAT concepts (Alkan et al., 2026). The validation split used here contains 3,025 papers, 13,239 exact ground-truth labels, and 1,864 unique observed candidate keywords. The UAT vocabulary has an 11 level hierarchical and polyhierarchical structure: a concept can have multiple parents, and concepts range from broad topics to very specific scientific terms.

UAT was designed to organize concepts at multiple levels of specificity. A paper's author-assigned labels are useful, but they are not a complete or noise-free statement of every relevant concept. Evaluating at multiple hierarchy levels is therefore a better fit for an interpretable concept-space claim.

4 Evaluation

4.1 Document prompting

For each paper, we build a prompt from its title and abstract:

```
Title: {title}
Abstract: {abstract}
Keywords:
```

The end of the prompt is voluntarily left blank so we can score candidate tokens using the MTP future heads.

4.2 Candidate vocabulary and matrices

The candidate vocabulary is the set of normalized UAT labels observed in the AstroConcepts train and validation splits. This gives 1,864 unique concepts in the validation score matrix. Each concept is tokenized with the MTP model tokenizer. For a four-head checkpoint, up to four tokens per concept are scored.

The scoring pipeline writes creates a matrix with shape 3025×1864. The percentile matrix is used as the primary representation because raw concept scores are not directly comparable across concepts. We do not need to know whether *Gamma-rays* has a higher absolute language-model probability than *Asteroseismology*; we need to know which papers are most associated with each concept.

4.3 Exact-label evaluation

We preserve the existing exact-label evaluator as a strict baseline. It ranks all candidate keywords for each paper and computes P@k, R@k, micro-F1@k, hit rate, mean average precision, and oracle label-count F1. The oracle label-count F1 predicts exactly as many labels as the gold label count for each document.

4.4 Hierarchy-aware evaluation

The new hierarchy-aware evaluator loads UAT 5.1 from `astrothesaurus/UAT`, resolves concept aliases, and builds parent paths. Because UAT is polyhierarchical, a concept can map to multiple ancestors at a given level. We preserve this multiplicity.

For a target level l, each candidate keyword maps to its ancestor set $A_l(c)$. Gold labels are mapped the same way. Candidate scores are aggregated upward by max pooling:

$$S_l(x, a) = \max_{c: a \in A_l(c)} S(x, c). \tag{2}$$

This creates a coarser document-by-branch matrix. A prediction is then evaluated against gold branches rather than exact leaves. We report gold-count F1, micro-F1@k, and concept-axis retrieval F1.

4.5 Experimental Setup

We use the official Hugging Face dataset `adsabs/SciX_UAT_keywords`. The validation split contains 3,025 papers, 13,239 exact ground-truth labels, 4.38 exact labels per paper, and 1,864 observed candidate keywords. The hierarchy evaluator uses UAT version 5.1 (Frey and Accomazzi, 2018).

The experiments use the `alpindale/multi-token-prediction` 7B checkpoint with four future heads. The model is not trained specifically for astrophysics and is not fine-tuned on AstroConcepts. This makes the evaluation a zero-shot test of the MTP scoring interface rather than a supervised classification experiment.

5 Results

5.1 Hierarchy-aware evaluation shows branch recovery

Despite not beating the State of the art, the results show that with a single forward pass per ab-

Granularity	Concepts	Gold	Gold F1	F1@5
Level 2	140	13,611	0.359	0.331
Level 3	606	15,570	0.222	0.203
Level 4	1,256	14,890	0.171	0.158
Level 5	1,689	14,060	0.140	0.135
Exact	1,864	13,239	0.121	0.118
No MTP Exact	1864	13239	0.125	0.120
DeepSeek	1,864	13,239	0.377	-

Table 1: Hierarchy-aware evaluation of the concept matrix. DeepSeek result is from Alkan et al. (2026) by reranking results of AstroBERT. No MTP Exact is `alpindale/multi-token-prediction` used for constrained generation

Concept	Support	Axis F1	AP
Gamma-rays	123	0.626	0.568
Supernovae	134	0.604	0.606
Biosignatures	10	0.600	0.565
Asteroseismology	20	0.600	0.508
Solar corona	116	0.578	0.552
Gamma-ray sources	96	0.563	0.515
Globular star clusters	46	0.543	0.606

Table 2: Strong level-3 concept axes for the percentile matrix and Average Precision.

stract our method can perform on par with the same model used with autoregressive constrained generation. It performs well above random chance (0.12 F1 for 1864 classes and 0.359 for 140 classes.) This serves as a proof of concept that MTP modules capture information all throughout the hierarchy without any training. When we evaluate the same percentile matrix at coarser UAT levels, performance improves substantially (Table 1).

5.2 Concept-axis retrieval

A second evaluation asks whether individual named axes retrieve documents bearing that concept or any descendant concept. For each branch, documents are ranked by the branch score. If the branch appears in m validation documents, the top m documents are predicted and F1 is hits/m.

For the percentile matrix, mean axis F1 at support at least 10 is 0.248 at level 2, 0.214 at level 3, 0.200 at level 4, and 0.207 at level 5. For illustration purposes, strong level-3 axes are shown in Table 2. Strong performance on precise concepts suggests that the method can perform well with concepts that are very long: "Globular star clusters" and "Asteroseismology" are tokenized into 5 tokens. and the rest of the list are under four, but their length shows that having access to an estimation of 4 tokens allows long technical keywords.

6 Discussion

The method described is good at producing a named, queryable concept space. Once a document has been scored, any concept column can be inspected directly. This is useful for analysis tasks where interpretability is crucial even at the cost of classification performance.

Another aspect worth-mentioning is that we have used only a 7b model in this exploratory study, whereas the original article uses DeepSeek V3 - a 671B parameters model- with a restrained generation from a selection made by AstroBert (a model trained specifically to be proeficient on astronomical content).

This method seems interesting to analyze the evolution over time of many attributes to find correlations, pioneers, scissions from explainable clusters.

7 Future Work

Several directions follow naturally. Evaluate checkpoints with more future heads. A twelve-head MTP model could score longer scientific concepts more faithfully. Use our method as a candidate generator for a constrained LLM. Fine-Tune a model explicitly for this task by dividing the loss for multi-label prediction. Optimize prompting and explore other forms such as *This user likes* or *This user hates* to explore different types of spaces for explainable user profiling and/or explainable recommendations.

8 Conclusion

This paper investigated whether multi-token prediction heads can be used as a lightweight interface for constructing interpretable concept spaces. On AstroConcepts, exact author-keyword prediction is weak: the best exact gold-count F1 is 0.121 and performs similarly to the same model with autoregressive constrained generation.

The conclusion is that even though MTP concept scoring does not replace trained classifiers, they provide a promising mechanism for building named, query-able, interpretable spaces. Hierarchical evaluation reveals that these spaces often recover the right semantic branch even when exact labels are missed. For applications such as exploratory analysis, user modeling, temporal profiling, and concept retrieval, that tradeoff may be needed.

Limitations

The method requires a predefined candidate vocabulary, and the four-head checkpoint truncates longer concepts, furthermore, since the heads are independent, the predictions cannot be used for certain syntaxic operations such as negation, coordination, and requires vigilance since a first token can lead into many others.

Finally, the model used here is not astrophysics-specific; the results should be interpreted as a proof of concept for the scoring interface rather than a final benchmark result.

References

Atilla Kaan Alkan, Felix Grezes, Sergi Blanco-Cuaresma, Jennifer Lynn Bartlett, Daniel Chivvis, Anna Kelbert, Kelly Lockhart, and Alberto Accomazzi. 2026. Astroconcepts: A large-scale multi-label classification corpus for astrophysics. *arXiv preprint arXiv:2604.02156*.

Vinamra Benara, Chandan Singh, John X Morris, Richard J Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings for language neuroscience by asking llms questions. *Advances in neural information processing systems*, 37:124137.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in eu legislation. In *Proceedings of the natural legal language processing workshop 2019*, pages 78–87.

Katie Frey and Alberto Accomazzi. 2018. The unified astronomy thesaurus: Semantic metadata for astronomy and astrophysics. *The Astrophysical Journal Supplement Series*, 236(1):24.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 13823–13844.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, and 1 others. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Milad Sabouri, Masoud Mansoury, Kun Lin, and Bamshad Mobasher. 2025. Towards explainable temporal user profiling with llms. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 219–227.

Yixuan Tang, Zhenghong Lin, Yandong Sun, Wynne Hsu, Mong Li Lee, and Anthony KH Tung. 2026. Qime: Constructing interpretable medical text embeddings via ontology-grounded questions. *arXiv preprint arXiv:2603.01690*.

Yile Wang, Zhanyu Shen, and Hui Huang. 2025. Ldir: Low-dimensional dense and interpretable text embeddings with relative representations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14397–14409.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.