NOEM^{3}A: a Neuro-symbolic Ontology-Enhanced Method for Multi-intent understanding in Mobile Agents

Ioannis Tzachristas [1,2], **Aifen Sui** [1*]
[1]Huawei Heisenberg Research Center, Munich, Germany
[2]Technical University of Munich, Germany

Abstract

Mobile agents must map natural-language requests to executable intents under tight latency and privacy constraints. Scaling the language model is often an inefficient way to improve this component. We present NOEM^{3}A, a lightweight neuro-symbolic layer that augments compact language models with an intent ontology. For each query, NOEM^{3}A retrieves a small ontology neighborhood, injects candidate action labels into the prompt and applies a token-level decoding prior toward valid labels. This injects symbolic intent structure into both input and output representations while keeping inference suitable for local interaction. We also use Semantic Intent Similarity (SIS), a hierarchy-aware diagnostic based on ontology depth, to capture semantic proximity when predicted intents differ lexically. Experiments on MultiWOZ 2.3 dialogues show that ontology augmentation consistently improves TinyLlama and Llama-3.2-3B on exact match and Slot-F1. Our results suggest that symbolic alignment is an effective strategy for accurate, responsive and privacy-preserving on-device NLU.

1 Introduction

Mobile assistants and GUI-based agents increasingly perform actions rather than only answer questions (Tzachristas, 2024): booking travel, creating reminders, ordering food, searching products or coordinating several tools. A single utterance may contain multiple intents (Gangadharaiah and Narayanaswamy, 2019), as in *"find a hotel near the station, book a taxi and remind me before departure"*. The language component must therefore produce a valid set of tool-facing labels, and it must do so fast enough for local interaction.

Large Language Models are effective semantic parsers, but they are not always the right deployment primitive for mobile agents. Cloud inference introduces privacy exposure and tail latency, while larger local models increase memory pressure and token-by-token compute. We study a complementary efficiency strategy: keep the neural model compact and supply an explicit symbolic action space at inference time.

We propose NOEM^{3}A (Neuro-symbolic Ontology-Enhanced Method for Multi-intent understanding in Mobile Agents). Following ontology/knowledge-guided NLP work (Pinhanez et al., 2021; He et al., 2021; Yu et al., 2023; He et al., 2023), the method uses an ontology whose nodes are mobile-agent domains, categories and executable intents. At runtime, the system retrieves a query-specific subgraph and uses it as both input context and output constraint. This shifts part of the reasoning burden from parametric generation to a small, inspectable symbolic lookup. Related work further includes out-of-scope and zero-shot intent detection (Cavalin et al., 2020; Siddique et al., 2021), LLM-assisted ontology construction and graph prompting (Doumanas et al., 2025; Wen et al., 2024; Liu et al., 2025), knowledge-injected language models (Zhang et al., 2019) and dialogue-state methods used as MultiWOZ 2.3 context (Hu et al., 2022; Lee and Lee, 2024; Heck et al., 2023; Chung et al., 2023; Li et al., 2024).

This paper focuses on efficient inference. We use exact match and Slot-F1 as primary task measures. We also report Semantic Intent Similarity (SIS) as an auxiliary diagnostic for whether an error stays near the correct ontology branch, but we avoid treating it as a standalone claim of success.

Our contributions are: (i) a formalized ontology-guided inference layer for multi-intent mobile NLU; (ii) a latency-focused accounting that separates retrieval, decoding bias and neural inference; and (iii) an evaluation showing that symbolic guidance improves compact models without requiring a larger generator.

*Correspondence: aifen.sui@huawei.com.

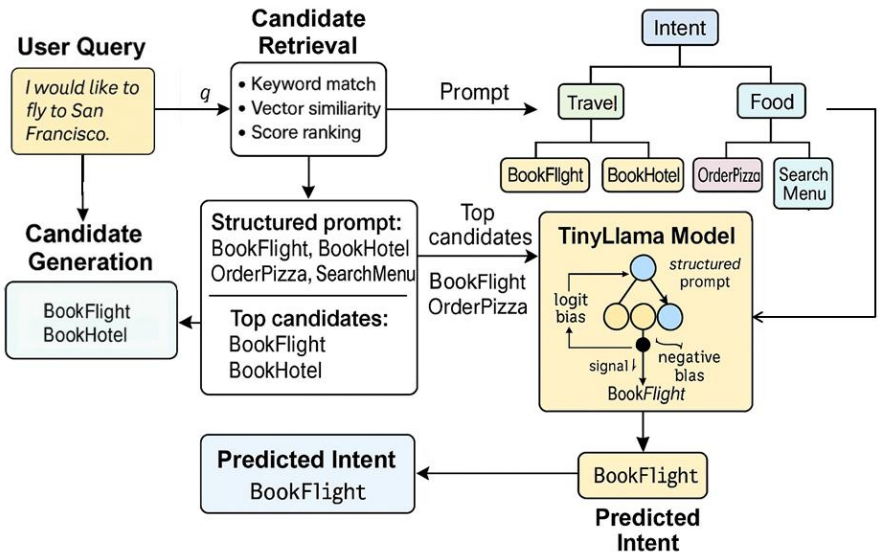

Figure 1: The NOEM^{3}A pipeline. A user query is matched to candidate ontology nodes; the retrieved candidates are serialized into a structured prompt and used as a decoding prior so that the compact model emits ontology-valid intent labels.

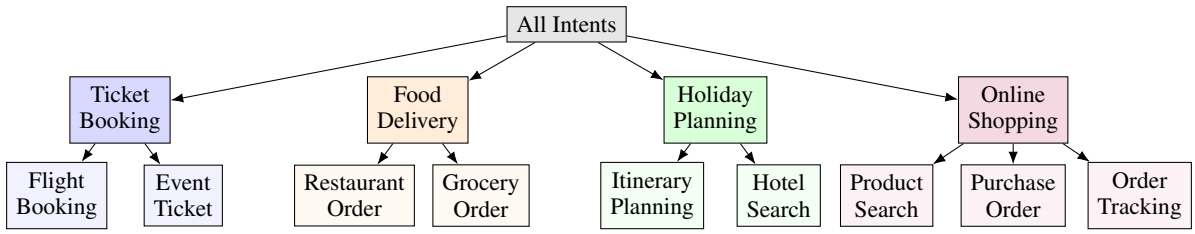

Figure 2: Excerpt from the ontology graph of mobile intents. Four top-level domains (Ticket Booking, Food Delivery, Holiday Planning, Online Shopping) are shown with some of their sub-intents. Each intent label is split over two lines for improved visual clarity.

2 Formal Method

Intent ontology. We represent the action space as a directed ontology $\mathcal{O} = (V, E)$, where each node $v \in V$ is a domain, category or executable intent and edges encode typed relations such as IS-A or RELATED-TO. Each node stores a label $\ell(v)$, description $d(v)$ and embedding

$$e_v = f\big(\ell(v) \oplus d(v)\big) \in \mathbb{R}^m, \tag{1}$$

where f is the same lightweight encoder used at inference time. The ontology therefore acts as both an action schema and a searchable memory over valid labels.

Subgraph retrieval. Given a query q, we compute $e_q = f(q)$ and rank ontology nodes by cosine similarity,

$$s(q, v) = \frac{e_q^\top e_v}{\|e_q\|_2 \|e_v\|_2}. \tag{2}$$

The raw candidate set is $C_q = \mathrm{TopK}_{v \in V}\, s(q, v)$. We then retrieve a local neighborhood

$$G_q = \mathrm{Expand}(\mathcal{O}, C_q; r), \tag{3}$$

where r controls optional parent, child or sibling expansion. In practice, G_q is small and fixed, so the downstream prompt remains short even when the ontology grows.

Ontology-augmented prompt. The retrieved labels are serialized into a compact instruction. Abstractly,

$$P_q = \big[q; \ell(v_1), \ldots, \ell(v_k)\big], \tag{4}$$

where the concrete prompt asks the model to choose one or more intents from the listed labels. This changes the task from unconstrained generation to selection among plausible ontology labels. Unlike document retrieval methods (Lewis et al., 2020; Hu et al., 2025; Edge et al., 2025), the retrieved objects are not evidence passages but executable action types. Figure 3 illustrates this token-to-node alignment.

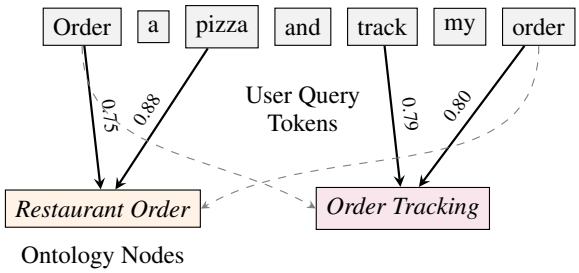

Figure 3: Semantic alignment between query tokens and ontology node embeddings. For the query "Order a pizza and track my order", food-ordering tokens align with *Restaurant Order*, while the tracking context aligns with *Order Tracking*. This alignment allows the retriever to select both intent nodes as relevant.

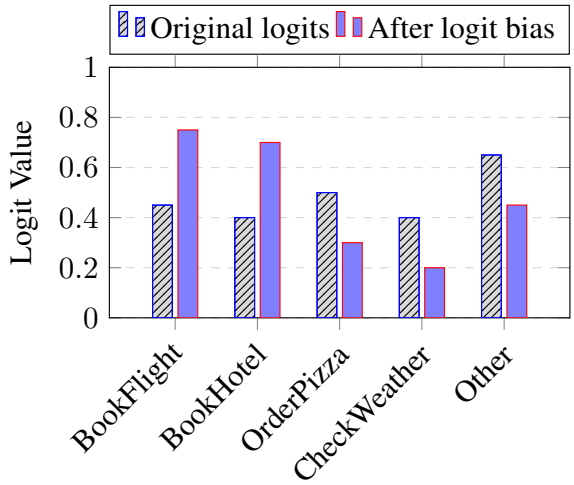

Figure 4: Effect of logit biasing in ontology-augmented inference. Ontology-relevant intents like `BookFlight` and `BookHotel` receive higher logits, while less relevant or unrelated intents are suppressed.

Token-level decoding prior. Let $T(G_q)$ be the set of tokenizer IDs appearing in labels from G_q. During decoding, we add a small bias to candidate-label tokens and a mild penalty to unrelated label tokens:

$$b_t = \begin{cases} \beta, & t \in T(G_q), \\ -\gamma, & t \in T(V \setminus G_q), \\ 0, & \text{otherwise}, \end{cases} \qquad z'_t = z_t + b_t. \tag{5}$$

The next-token distribution is then $p_t = \mathrm{softmax}(z')_t$. We tune β and γ conservatively; the goal is to regularize labels rather than force an intent when the query does not support it. Figure 4 shows the intended effect of the decoding prior.

Auxiliary classifier. When labeled data are available, we attach a multi-label classifier to the pooled hidden state h_q:

$$p^{\text{clf}} = \mathrm{sigmoid}(Wh_q + a), \tag{6}$$

$$\mathcal{L}_{\text{clf}} = -\sum_{v \in V} \big[y_v \log p_v + (1 - y_v) \log(1 - p_v) \big]. \tag{7}$$

Generated labels and classifier labels are merged as

$$\hat{Y} = Y_{\text{gen}} \cup \{ v \in V : p_v^{\text{clf}} > \tau \}. \tag{8}$$

This optional head is useful when a mobile platform has curated logs or simulator-generated labels.

Hierarchy-aware diagnostic. For analysis only, we compute a taxonomy score inspired by Wu and Palmer similarity (Wu and Palmer, 1994). For predicted node u and gold node v,

$$\mathrm{SIS}(u, v) = \frac{2\,\mathrm{depth}(\mathrm{LCA}(u, v))}{\mathrm{depth}(u) + \mathrm{depth}(v)}. \tag{9}$$

For multi-intent sets, we compute a weighted bipartite matching between predicted and gold nodes and average matched similarities. Figure 5 gives an example calculation. Because SIS depends on the chosen ontology topology, it is a debugging signal rather than a cross-system leaderboard metric.

Latency accounting. We report the symbolic layer's added latency separately from neural inference. Exact retrieval is $O(|V|m)$ for embedding dimension m; an approximate nearest-neighbor index makes lookup sublinear in practice. Prompt overhead is $O(|G_q|)$ labels and the decoding prior is $O(|T(G_q)|)$ token updates, both bounded by a small candidate set. This accounting keeps the deployment claim explicit without changing the underlying neural model.

3 Experiments

Setup. We evaluate on MultiWOZ 2.3 turns (Budzianowski et al., 2018; Han et al., 2021). The dataset is mapped to mobile-intent labels using an LLM annotator followed by ontology consistency checks. This setup is intended to test efficient multi-intent grounding, not to replace the standard MultiWOZ 2.3 dialogue-state leaderboard.

We compare TinyLlama (Zhang et al., 2024) and Llama-3.2-3B (Meta AI, 2024) with and without NOEM^{3}A. We report exact match (EM), micro Slot-F1 and the auxiliary hierarchy similarity described above. The most relevant comparison is within each compact model family, because those pairs isolate the cost and benefit of adding the symbolic layer. We set $k = 5$ retrieved candidates and

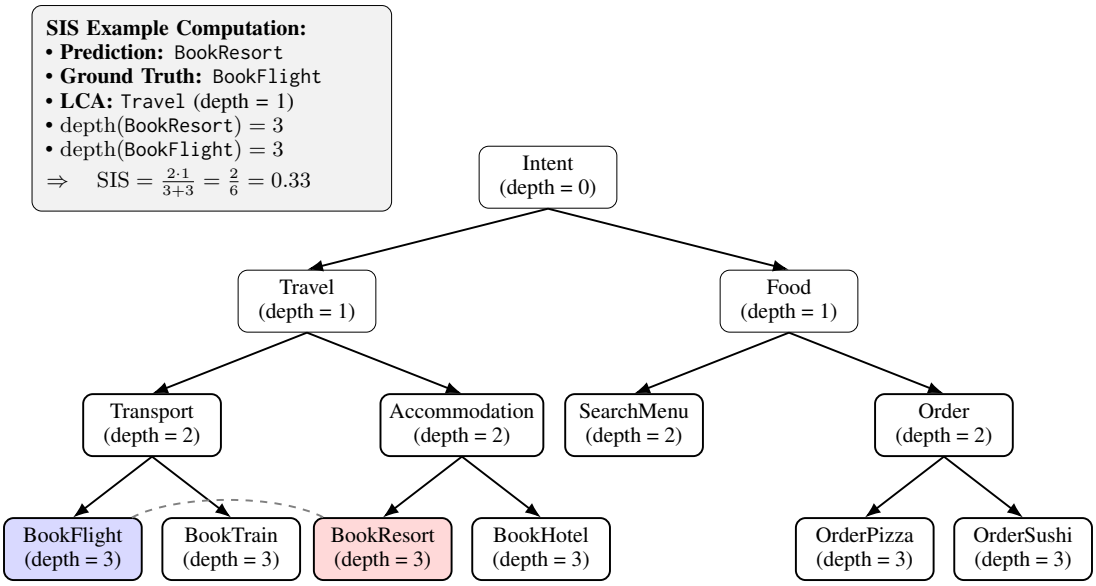

Figure 5: Semantic Intent Similarity (SIS) example. Predicting `BookResort` instead of `BookFlight` yields a similarity score of 0.33 due to their shared ancestor `Travel`.

Model / configuration	EM	Slot-F1	SIS
TinyLlama	14.6	36.7	0.37
TinyLlama + NOEM^{3}A	23.2	55.3	0.62
Llama-3.2-3B	22.8	58.8	0.63
Llama-3.2-3B + NOEM^{3}A	**35.7**	**72.8**	0.85
Llama + SI only	30.1	66.4	0.79
Llama + LB only	26.3	61.2	0.71
Llama + SI + LB	34.5	71.8	0.83
Llama + SI + LB + CLF	35.7	72.8	0.85

Table 1: Compact-model results and Llama-3.2-3B ablations. SI is ontology retrieval plus prompting, LB is logit biasing and CLF is the auxiliary classifier. The last column is reported only as a hierarchy-aware diagnostic.

tune the decoding prior on development data, fixing $\beta = 0.3$ and $\gamma = 0.2$ after a small grid search found a broad plateau around those values.

Accuracy gains under compact inference. Table 1 shows that the ontology layer improves both compact models. TinyLlama gains 8.6 EM points and 18.6 Slot-F1 points. Llama-3.2-3B gains 12.9 EM points and 14.0 Slot-F1 points. The ablation indicates that retrieval plus ontology-conditioned prompting is the largest single contributor; logit biasing improves label fidelity, and the classifier adds a small final gain. This pattern is consistent with the intended efficiency role of the method: the ontology narrows the output space so that a compact model need not infer the full action inventory from parameters alone.

Latency overhead. Table 2 reports the added latency of the symbolic layer. Querying a 5k-node ontology took 2.1 ± 0.3 ms. Scaling the graph to 50k nodes increased lookup to 3.8 ms. The logit update touched at most 32 label tokens and remained below 0.1 ms. These numbers should not be read as universal device benchmarks, but they show that lookup and logit control add only millisecond-scale delay.

Operation	Latency	Notes
5k lookup	2.1 ± 0.3 ms	ANN/top-k
50k lookup	3.8 ms	scaled index
Logit mask	< 0.1 ms	$\leq$32 tokens

Table 2: Measured symbolic-layer latency. Retrieval and logit control are separated from neural decoding.

Prompt-template check. We also tested whether gains depend on the exact wording of the candidate-list prompt. Five development-set variants changed label order, added a task cue, inserted one example, replaced *intent* with *option* or removed most meta-language. EM stayed within roughly one point of the baseline prompt, suggesting that the main signal is the retrieved label set rather than a fragile phrasing choice.

4 Discussion

Why this helps efficiency. The method does not make the language model itself cheaper. Instead, it improves the amount of task accuracy obtained from a small model by providing a compact symbolic prior. This is useful for mobile agents because supported actions are usually known in advance: the ontology can double as the planner's schema and the NLU model's constrained label space. In

other words, the system spends a small amount of compute to recover the valid action vocabulary before asking the model to decide among actions. This is a better match to mobile agents than open-ended generation over arbitrary text.

Failure localization. The symbolic interface makes errors easier to inspect. If an intent is missed, developers can check whether it was absent from retrieval, present but ignored during generation, or filtered by the classifier. These failure modes suggest different fixes: improving node descriptions, adjusting retrieval expansion, tuning β, γ or changing merge thresholds. This is also why we keep the hierarchy-aware score as a diagnostic: it can reveal whether the system selected the wrong leaf inside a relevant branch or moved to an unrelated branch, but it should not replace EM or Slot-F1.

Relation to efficient NLP. The approach is complementary to compression, pruning and mobile-specific models (Sun et al., 2020; Yi et al., 2024; Tan et al., 2024). Those methods reduce neural cost directly; ontology guidance reduces the need to scale model size for a structured intent task. The two strategies can be combined: a quantized or mobile-specialized model can still benefit from ontology retrieval, while the ontology layer remains independent of the model's internal architecture.

5 Conclusion

We presented NOEM^{3}A as an efficient ontology-guided layer for multi-intent understanding in mobile agents. The central contribution is practical rather than metric-centric: a small symbolic lookup and ontology-guided logit biasing can improve compact local models while adding only millisecond-scale latency. The formalization above makes the mechanism inspectable, and the ablation indicates where gains come from. Future work should evaluate the same mechanism on real mobile-agent traces, richer GUI context and broader hardware profiles.

Limitations

Our evaluation uses the MultiWOZ 2.3 dataset rather than a naturally collected mobile-agent corpus. The annotations and ontology require human auditing; poor ontology design can affect retrieval and the auxiliary hierarchy diagnostic. Logit biasing assumes that intent labels map to short token spans, so long labels may require special tokens or classifier-only handling. The latency observations come from one device class and should be replicated across hardware and runtime stacks before deployment claims are generalized.

Ethical considerations

The system is designed for privacy-preserving local intent understanding, but mobile agents can execute consequential actions. Deployments should require confirmation before purchases, bookings, data sharing or other high-impact actions. Ontology construction can encode omissions or cultural assumptions, so the intent inventory and LLM-assisted annotations should be audited before production use.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3952–3961, Online. Association for Computational Linguistics.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. InstructTODS: Large language models for end-to-end task-oriented dialogue systems. In *Proceedings of the Second Workshop on Natural Language Interfaces*, pages 1–21, Bali, Indonesia. Association for Computational Linguistics.

Dimitrios Doumanas, Andreas Soularidis, Dimitris Spiliotopoulos, Costas Vassilakis, and Konstantinos Kotis. 2025. Fine-tuning large language models for ontology engineering: A comparative analysis of gpt-4 and mistral. *Applied Sciences*, 15(4).

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 564–569, Minneapolis, Minnesota. Association for Computational Linguistics.

Ting Han, Ximing Liu, Ryuichi Takanobu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie

Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. *Preprint*, arXiv:2010.05594.

Ting He, Xiaohong Xu, Yating Wu, Huazhen Wang, and Jian Chen. 2021. Multitask learning with knowledge base for joint intent detection and slot filling. *Applied Sciences*, 11(11).

Yacheng He, Qianghuai Jia, Lin Yuan, Ruopeng Li, Yixin Ou, and Ningyu Zhang. 2023. A concept knowledge graph for user next intent prediction at alipay. In *Companion Proceedings of the ACM Web Conference 2023*, page 45–48. ACM.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. GRAG: Graph retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4145–4157, Albuquerque, New Mexico. Association for Computational Linguistics.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jihyun Lee and Gary Geunbae Lee. 2024. Inference is all you need: Self example retriever for cross-domain dialogue state tracking with chatgpt. *Preprint*, arXiv:2409.06243.

Patrick 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A. Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8688–8704, Bangkok, Thailand. Association for Computational Linguistics.

Runxuan Liu, Bei Luo, Jiaqi Li, Baoxin Wang, Ming Liu, Dayong Wu, Shijin Wang, and Bing Qin. 2025. Ontology-guided reverse thinking makes large language models stronger on knowledge graph question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15269–15284, Vienna, Austria. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

Claudio Santos Pinhanez, Heloisa Candello, Paulo Cavalin, Mauro Carlos Pichiliani, Ana Paula Appel, Victor Henrique Alves Ribeiro, Julio Nogima, Maira de Bayser, Melina Guerra, Henrique Ferreira, and Gabriel Malfatti. 2021. Integrating machine learning data with symbolic knowledge from collaboration practices of curators to improve conversational systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

A. B. Siddique, Fuad Jamour, Luxun Xu, and Vagelis Hristidis. 2021. Generalized zero-shot intent detection via commonsense knowledge. *Preprint*, arXiv:2102.02925.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Fuwen Tan, Royson Lee, Łukasz Dudziak, Shell Xu Hu, Sourav Bhattacharya, Timothy Hospedales, Georgios Tzimiropoulos, and Brais Martinez. 2024. MobileQuant: Mobile-friendly quantization for on-device language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9761–9771, Miami, Florida, USA. Association for Computational Linguistics.

Ioannis Tzachristas. 2024. Creating an llm-based ai-agent: A high-level methodology towards enhancing llms with apis. *Preprint*, arXiv:2412.13233.

Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. MindMap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Rongjie Yi, Xiang Li, Weikai Xie, Zhenyan Lu, Chenghua Wang, Ao Zhou, Shangguang Wang, Xiwen Zhang, and Mengwei Xu. 2024. Phonelm:an efficient and capable small language model family through principled pre-training. *Preprint*, arXiv:2411.05046.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. FolkScope: Intention knowledge graph construction for E-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1173–1191, Toronto, Canada. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.