

Role-Play Prediction using Ontology-Based Graph Convolutional Network Model

Asyafa Ditra Al Hauna¹, Andi Prademon Yunus^{1*}, Siti Khomsah¹, Yit Hong Choo², Masanori Fukui³

¹Telkom University, Banyumas, Indonesia, ²Deakin University, Australia, ³Iwate Prefectural University, Japan
Email: alditra@student.telkomuniversity.ac.id, andiaiy@telkomuniversity.ac.id, sitijk@telkomuniversity.ac.id, y.choo@deakin.edu.au, fukui_m@iwate-pu.ac.jp

*Corresponding Author

Abstract

Current applications of large language models often assign tasks without consideration of how LLMs understand a given prompt. Simple commands sometimes do not guarantee desired responses, as LLMs are systems based on mathematical modeling and cannot cognitively be capable of understanding commands. Hence, a method is required to guide LLMs in performing tasks appropriately. This paper presents a method to develop model-based automation of role selection supported by ontology. This can allow for more accurate and relevant role recommendations than if done manually. As such, this optimization at hand improves the performance of LLMs for specific tasks and overcomes the limitations of previous studies that define the roles by hand.

Keywords: Graph Convolutional Network, Large Language Models, Ontology, Role-play

1. Introduction

In recent years, the development of Large Language Models (LLMs) has demonstrated remarkable progress, with researchers competing to advance their findings as breakthroughs in artificial intelligence, particularly in natural language preprocessing. This advancement is primarily driven by innovations in transformer architecture, which enables models to analyze context and semantics through the attention mechanism. Such innovations allow LLMs to exhibit reasoning abilities that can approximate human-like reasoning despite being grounded in mathematical modeling. One prominent example of an advanced LLM is GPT-4, which has made significant strides in adapting to specific tasks through fine-tuning, reinforcement learning with human feedback (RLHF), and domain-specific training [1].

Research on LLMs has shown a notable impact in the technology field and other domains. In education, LLMs promise to enable intelligent tools such as personalized learning [2]. Similarly, in medicine, LLMs can serve as agents that work alongside professionals to analyze complex problems and information [3]. Concurrently, there has been a surge in studies exploring optimization techniques for LLM performance, particularly in specialized domains. One relatively new optimization technique focuses on prompt engineering. Given that an LLM's performance heavily depends on the given prompts, prompt engineering seeks to enhance outputs by designing prompts that lead to the expected response [4][5].

A specific optimization approach within this framework is role-play prompting, which aims to improve LLM performance without requiring intensive or complex interventions. This technique narrows the scope of the LLM's response by assigning a relevant role aligned with the task. For example, a study investigating the impact of assigning roles within prompts reported

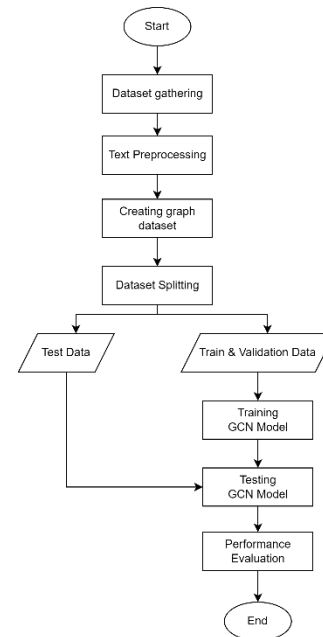


Fig. 1 Methodology flowchart

that manually assigning roles based on task benchmarks improved response accuracy by over 10% [6]. However, because the role assignment in this study was done manually, there is room for automation in determining relevant roles.

This research proposes developing a deep learning model capable of predicting roles based on tasks. Specifically, the focus is on predicting roles such as mathematician, nurse, education teacher, and recorder. A dataset comprising task-relevant information and ontology data, including skills, abilities, and role definitions, will be used to train a Graph Convolutional Network (GCN). The model aims to accurately predict the most relevant role based on the given task.

2. Methodology

The objective of this study is to train a GCN model to predict roles that are relevant to the given task. The procedures undertaken in this research are illustrated in Fig. 1.

2.1. Dataset Gathering

In line with the focus of this research, the model to be developed will be trained using datasets aimed at predicting four roles: mathematician, nurse, education teacher, and recorder. The datasets employed in this study serve as benchmarks commonly used to evaluate reasoning capabilities and the performance of large language models (LLMs). For the mathematician role, the Algebra Question Answering with Rationales (AQUA-RAT) and MultiArith datasets are utilized, containing story problems and mathematical equations [7][8]. To predict the nurse’s role, the MedQA dataset is selected as it encompasses medical terminology and concepts [9]. For the teacher’s role, the BigBench Date and StrategyQA datasets are employed because they include foundational yet comprehensive questions that reflect the broad knowledge expected of a teacher. Lastly, for the recorder role, the BigBench Object Tracking dataset is used, as it contains questions related to tracking events or occurrences. The ontology applied in this research is Occupation Ontology (OccO), which provides a structured framework consisting of the skills, abilities, and definitions specific to each role [10]. Each of those datasets was gathered from the repository listed in the corresponding paper.

2.2. Text Preprocessing

Text preprocessing is a pivotal step in the field of text data processing, as it directly impacts the overall performance and accuracy of machine learning models [7]. This process must be tailored to the unique characteristics and behaviors of the collected data to ensure effective outcomes, for the details see Table 1. One preprocessing technique involves replacing digits in text data with the word “numeric” using regular expressions. This step is specifically applied to datasets categorized under the mathematician class, as it aids the classifier in recognizing that the term “numeric” is contextually relevant to this category.

Another essential method includes the removal of dates formatted as MM/DD/YYYY, which is executed through regular expressions. This approach is applied exclusively to datasets pertaining to the teacher’s class, as dates in this context are considered irrelevant for classification tasks. Similarly, named entity recognition (NER) provided by flair is utilized to identify and remove personal names from the data [8], followed by applying regular expressions. However, this step excludes data from the MedQA class, as it lacks identifiable named subjects.

Furthermore, all datasets undergo lemmatization, a process that identifies the root form of each word, ensuring uniformity in textual representation. This step

leverages the lemmatizer provided by the Stanza [9]. Lastly, keyword extraction is conducted across applicable datasets using the KeyBERT algorithm [10] which employs sentence transformers to identify key terms effectively.

Table 1. Text preprocessing for each dataset

Stage dataset	Replace Digit	Remo- ving Date	Replacing Name	Lemmatiza- tion	Keyword Extraction
AQUA-RAT	1	0	1	1	1
Mulriarith	1	0	1	1	1
MedQA	0	0	0	1	1
StrategyQa	0	1	1	1	1
BigBench Date	0	1	1	1	1
BigBench Object Tracking	0	0	1	1	1

For ontology data, the preprocessing stage involves applying stopwords removal and lowercasing techniques. These preprocessing steps collectively optimize the datasets, preparing them for subsequent graph dataset creation and task modelling.

2.3. Creating graph dataset

All data from the text preprocessing process is represented as a graph. Each word or keyword is treated as a node in this representation, while the relationships between words and their corresponding datasets are depicted as edges. The first step involves creating six central nodes, each representing one of the dataset names. Subsequently, the text data from each dataset and ontology is iterated to construct a subgraph. In this subgraph, each word is represented as a node, which is then connected to the corresponding central node associated with the dataset.

Since the GCN model requires two inputs, numerical data representing the attributes of each node and an edge index that captures the relationships between nodes [11], the data represented by a node in this study consists of text. It will undergo a word embedding process to convert it into numerical form, this process leveraging sentence transformer [12]. Meanwhile, the edge index will be derived from the iterative process of constructing each subgraph.

2.4. Dataset splitting

In this study, data splitting was conducted using a masking technique on graph-structured data. This masking procedure involves assigning a binary value, either zero or one, to each node in the graph. A zero value indicates that the corresponding node will not be included in the training data, whereas a value of one signifies its inclusion. For the initial experimental phase, the data was divided into three subsets with a ratio of 60%, 20%, and 20%, respectively, representing the training, validation, and test datasets.

2.5. Model building and training

In the development stage of the baseline model, the primary objective is to establish a foundation for assessing the level of complexity required to learn patterns inherent in the data. This model development process leverages GCNs to capture features from graph-structured data. Batch normalization layers are employed to normalize the input of each layer for every batch during training to force the model for faster convergence and introduce slight regularization. Activation functions, such as tanh and ReLU, are incorporated to enable the model to identify non-linear patterns effectively. Additionally, dropout layers are utilized as a regularization technique to mitigate the risk of overfitting, ensuring the model generalizes well to unseen data. Experiment configuration during the training state is shown in Table 2. In addition, the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm is employed to help visualize the logit of each class generated by the model. It aims to determine how well the model improves classification during training [13].

Table 2. Training configuration

Parameter	Configuration
Epoch	500
GCN layer input size	768
GCN layer hidden size	64
Loss Function	Cross Entropy
Optimizer	Adamax
Learning Rate	$1e^{-3}$
Dropout Probability	0.5

2.6. Model evaluation

In the model evaluation phase, the confusion matrix and ROC-AUC curve are employed to assess the classification performance of the node classifier [14]. The confusion matrix serves as a tool to analyze the classifier's performance by evaluating predictions across four key indicators: accuracy, precision, recall, and F1 score. Accuracy measures the proportion of correctly classified data relative to the total dataset. Precision evaluates the ratio of correctly predicted positive instances to the total instances predicted as positive by the model. Conversely, recall calculates the ratio of correctly predicted positive instances to all actual positive instances in the dataset. The F1 score provides a harmonic mean of precision and recall, evaluating the model's performance when precision and recall are equally important. Additionally, the ROC-AUC curve is utilized to evaluate the classifier's performance based on the false positive rate (FPR) and true positive rate (TPR), visualized on the x-axis and y-axis, respectively, to depict the trade-off between sensitivity and specificity.

3. Results and Discussions

3.1. Training and Testing Model

The training stage is implemented according to the configuration specified in the methodology section. The t-SNE graphs before and after training are compared to suggest how well the model learns the features, as shown in Fig. 2 and Fig. 3.

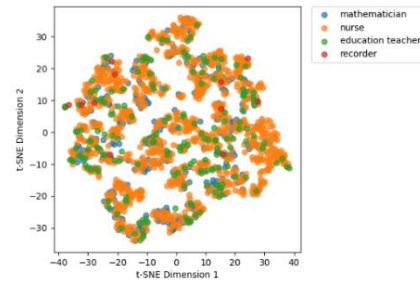


Fig. 2 t-SNE plot before training

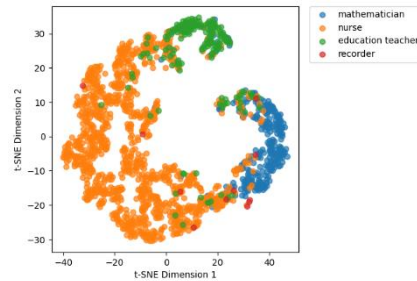


Fig. 3 t-SNE plot after training

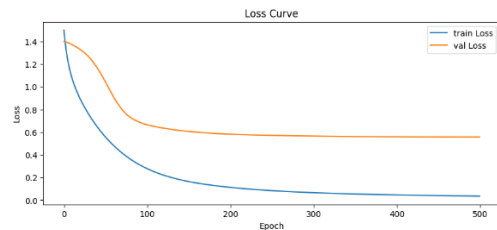


Fig. 4 Loss curve

Based on those graphs, the model can classify nodes well in mathematician, nurse, and education teacher classes but is suboptimal at classifying nodes in the recorder class. The model loss curve during training in Fig. 4 shows that the model learns well due to consistently decreasing training loss, indicating it converges sufficiently. At the beginning of training, the validation loss continues to decrease, but until around the 100th epoch, the validation loss stops decreasing even though the training loss continues to decrease.

3.2. Model Evaluation

The evaluation results on the test data, as presented in Table 3, indicate that the model demonstrates satisfactory performance in classifying data belonging to the mathematician and nurse categories. However, its performance is less than optimal when classifying instances in the education teacher category, and it proves to be particularly ineffective in accurately classifying data associated with the recorder category.

Table 3. Classification report

Class \ Metrics	Precision	Recall	F1-Score
Mathematician	0.74	0.83	0.78
Nurse	0.90	0.86	0.88
Education teacher	0.35	0.38	0.36
Recorder	0.0	0.0	0.0

Furthermore, Fig. 5 corroborates these findings, as the most prominent curves are observed in the mathematician and nurse classes. However, it is important to note that the data distribution within each class does not influence performance evaluation based on the curve area. It explains why the education teacher and recorder classes exhibit relatively good curve areas despite their lower classification performance, exceeding 0.5.

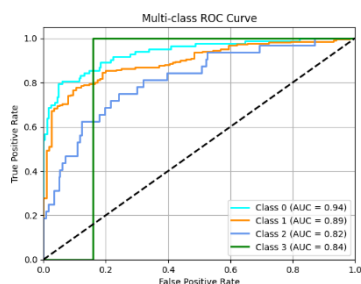


Fig. 5 ROC AUC Curve (class 0: mathematician, class 1: nurse, class 2: education teacher, class 3: recorder)

4. Conclusion

This study proposes a method to optimize prompts for Large Language Models (LLMs). As an initial experiment, the research focuses on predicting four roles: mathematician, nurse, education teacher, and recorder. A GCN algorithm was employed to develop a model capable of predicting relevant roles. Additionally, OccO, an ontology containing occupational role information, was utilized as supplementary data to help the model effectively capture features associated with specific roles. Experimental results indicate that the GCN-based model demonstrated suboptimal performance, particularly in predicting the roles of education teacher and recorder. These findings highlight limitations in the model's ability to generalize across all classes. Future work will enhance the model by integrating other advanced feature engineering techniques to capture text data representations better. Exploring various model architectures and hyperparameter configurations will also be essential to align model complexity with the dataset's characteristics. Furthermore, addressing class imbalance issues should be prioritized to ensure balanced data distribution, thereby improving the model's learning capability.

Acknowledgement

We extend our heartfelt gratitude to Telkom University for its unwavering support during this research. It is important to note that this work did not receive any financial assistance from funding agencies. Nonetheless, we deeply appreciate the researchers and

collaborators whose invaluable contributions and insights greatly enhanced the quality of this study. Their dedication and support played a vital role in the successful completion of this research.

References

1. OpenAI *et al.*, GPT-4 Technical Report, 2023, doi: <https://doi.org/10.48550/arXiv.2303.08774>.
2. W. Gan, Z. Qi, J. Wu, and J. C. W. Lin, Large Language Models in Education: Vision and Opportunities, in *Proceedings - IEEE International Conference on Big Data*, 2023, pp. 4776–4785, doi: <https://doi.org/10.48550/arXiv.2311.13160>.
3. X. Tang *et al.*, MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning, *ACL*, 2024, pp. 599–621, doi: <https://doi.org/10.18653/v1/2024.findings-acl.33>.
4. Y. Zhou *et al.*, Large Language Models Are Human-Level Prompt Engineers, *ICLR*, 2023, doi: <https://doi.org/10.48550/arXiv.2211.01910>.
5. H. Sun *et al.*, AutoHint: Automatic Prompt Optimization with Hint Generation, 2023, doi: <https://doi.org/10.48550/arXiv.2307.07415>.
6. A. Kong *et al.*, Better Zero-Shot Reasoning with Role-Play Prompting, *Proc. NAACL*, vol. 1, pp. 4099–4113, 2024, doi: <https://doi.org/10.48550/arXiv.2308.07702>.
7. A. R. Baskara, M. A. S. Jati, M. Maulida, Y. Sari, N. F. Mustamin, and E. S. Wijaya, Classification of User Reviews for Software Maintenance in Indonesian Language Using IndoBERT-BiLSTM (Case Study: MyPertamina), *8th International Conference on Informatics and Computing, ICIC 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, doi: <http://dx.doi.org/10.1109/ICIC60109.2023.10381946>.
8. A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP, *NAACL*, 2019, pp. 54–59, doi: <https://doi.org/10.18653/v1/N19-4010>.
9. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, *ACL*, 2020, pp. 101–108, doi: <https://doi.org/10.18653/v1/2020.acl-demos.14>.
10. Maarten Grootendorst, KeyBERT: Minimal keyword extraction with BERT, *Zenodo*: 0.8.0, 2020, 10.5281/zenodo.4461265.
11. W. L. Hamilton, Graph Representation Learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2020, Vol. 14, No. 3, pp. 47–53, isbn: 978-3031004605.
12. N. Reimers and I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019, doi: <https://doi.org/10.48550/arXiv.1908.10084>.
13. W. Li, J. E. Cerise, Y. Yang, and H. Han, Application of t-SNE to Human Genetic Data, *J Bioinform Comput Biol*, 2017, <http://dx.doi.org/10.1101/114884>.
14. T. Fawcett, An introduction to ROC analysis, *Pattern Recognit Lett*, 2006, vol. 27, pp. 861–874, doi: <https://doi.org/10.1016/j.patrec.2005.10.010>.

Authors Introduction

Mr. Asyafa Ditra Al Hauna



He is currently pursuing the Bachelor of Informatics Engineering with Honours in Faculty of Informatics, Telkom University, Indonesia. His research interests are Machine Learning, Deep Learning, and Natural Language Processing.

Dr. Andi Prademon Yunus



He is an Assistant Professor at Telkom University, and he received his PhD in Engineering from Mie University, Japan. His research focuses on applied and fundamental machine learning for motion and behavior computing. He also collaborates with industry partners to develop AI-based tools for language modeling and image analytics.

Ms. Siti Khomsah, M. Cs



She is an Lecturer in Faculty of Informatics at Telkom University in Indonesia. She received her Master in Computer Science from Universitas Gajah Mada. Her research interests are Text Mining, Data Analytics, and Big Data Analytics.

Dr. Yit Hong Choo



He has completed his PhD and is now a Research Fellow in Operations Analytics at Deakin University's Institute for Intelligent Systems Research and Innovation (IISR). His research focuses on advanced multi-objective optimisation algorithms for complex maintenance scheduling in rolling stock. He also collaborates with transportation industry partners to develop AI-based tools for video and image analytics.

Dr. Masanori Fukui



He is an associate professor at Iwate Prefectural University and received his PhD in Engineering from Hiroshima University, Japan. His research interests include creativity education, problem-posing, computational thinking, and learning technology.