Human-Readable Representation for Graph Neural Networks

Anonymous ACL submission

Abstract

This research presents an innovative method for representing nodes in graph neural networks (GNNs) using human-readable text in natural language, diverging from the traditional numerical embeddings. By employing a large language model (LLM) as a projector, we train GNNs to aggregate information from neighboring nodes and update node representations iteratively. Our experiments on the MovieLens dataset, widely used for recommendation tasks, demonstrate that human-readable representations effectively capture useful information for recommendations. This suggests that LLMs can successfully aggregate neighborhood information in a graph. Furthermore, fine-tuning the LLMs can improve their ability to generate more application-specific human-readable representations. This technique not only facilitates the incorporation of world knowledge into GNNs but also enhances their interpretability and allows for human intervention in their behavior. Our approach shows significant potential for making graph neural networks more understandable and controllable.

1 Introduction

Graph neural networks (GNNs) show effectiveness in many applications in the area of natural language processing and information retrieval (Wang et al., 2021; Wu et al., 2020; Fang et al., 2020). One of the key advantages of GNNs is their ability to exploit the relational structure in data, effectively capturing dependencies and relationships between entities. In NLP, this is extremely beneficial for tasks like semantic role labeling, entity recognition, relation extraction, and question answering, as GNNs can model the syntactic and semantic relationships between words and phrases, leading to richer contextual representations. In recommendation systems, GNNs excel by taking into account the complex interactions between users and items. For instance, in social network-based recommendations, GNNs can exploit the connections and interactions between users to provide more personalized and contextually relevant recommendations (Wu et al., 2022; Gao et al., 2022). Additionally, GNNs' capacity to aggregate information from neighbors in the graph allows for better handling of sparse data, which is a common challenge in recommendation systems. By leveraging the structural information inherent in data, GNNs offer enhanced performance and insights in tasks where relationships are key.

GNNs are adept at learning nuanced representations of nodes and edges within graph-structured data. Initially, each node is characterized by a feature vector, which could be innate attributes or embeddings. Throughout the GNN's processing, these representations are refined via an aggregation mechanism, typically employing message-passing techniques. Specifically, a node's updated representation is derived from a combination of its antecedent representation and the representations of its adjacent nodes. As the iterations progress, the GNN constructs sophisticated representations of nodes and edges, encapsulating not only the features but also the underlying structure of the graph. These representations are instrumental for an array of downstream applications such as node classification, link prediction, and graph classification.

This work introduces a novel approach for representing nodes in GNNs through human-readable text, as opposed to the conventional numerical embeddings. Utilizing a large language model (LLM) for information aggregation and node updates, this method takes advantage of LLM's natural language processing capabilities (Thoppilan et al., 2022; Shuster et al., 2022; OpenAI, 2023). Specifically, each node is represented by a description in natural language (e.g., English), and through the synergy between GNN and LLM, task-oriented descriptions for nodes will be updated during training. We conducted experiments using the MovieLens dataset (Harper and Konstan, 2015a), known for its abundant textual metadata, and showed that incorporating human-readable text elevates the network's interpretability and accessibility, facilitating deeper analysis by domain experts (Yuan et al., 2022).

Furthermore, the inclusion of human-readable text enables the incorporation of expert insights and fine-tuning, which proves invaluable in scenarios where compliance with ethical or business standards is paramount. This methodology holds significant potential in recommendation systems, where transparency is essential for earning user trust (Chen et al., 2022; Afchar et al., 2022). By understanding and having the ability to modify the network's decisions, we can achieve stronger and more dependable recommendations.

By employing human-readable text for node representation in graph neural networks, we enhance both the interpretability and manageability of these networks, laying the groundwork for the development of more responsible and transparent intelligent systems. Our contributions are three-fold as follows.

- The introduction of human-readable text as node representations in GNNs facilitates a more intuitive understanding of the network, enabling non-experts to interpret its structure and behavior more easily.

- By allowing for the integration of expert knowledge, this approach ensures that the network can be fine-tuned according to domain-specific standards or ethical considerations, making it adaptable for a wide range of applications.

- Through the cooperation between GNNs and LLMs, the method provides richer, task-oriented descriptions for nodes, which could lead to improved accuracy and performance in tasks such as recommendation systems, by harnessing the contextual information embedded in natural language.

2 Related Work

2.1 LLMs and GNNs

Recent studies have shown significant interest in integrating large language models (LLMs) with graph data. For instance, TAPE (He et al., 2024) uses LLMs to capture textual information as features, enriching graph data with semantic context. GraphAdapter (Huang et al., 2024) addresses text-attributed graphs (TAGs) by modeling textual attributes with LLMs, enhancing graph representations. GraphLLM (Chai et al., 2023) converts graphs into text, enabling LLMs to make predictions based on this transformed data, thus leveraging their language understanding capabilities.

Chen et al. (Chen et al., 2024) provide a comprehensive survey of LLM performance on graphs, identifying two main roles: enhancers, which improve input data quality by adding contextual information, and predictors, which directly generate outputs from graph data.

In contrast, our approach employs LLMs as both enhancers and projectors in training graph neural networks (GNNs). We utilize LLMs to aggregate information from neighboring nodes and update node representations iteratively, capturing rich semantic information. This method not only enhances the integration of world knowledge into GNNs but also improves interpretability and allows for human intervention. Our research contributes a novel method for leveraging LLMs' strengths, making GNNs more understandable, controllable, and effective for complex tasks.

2.2 LLMs for Recommendation Systems

With the advancement of large language models (LLMs), many recommendation system studies have utilized them to enhance item information. For example, IDGenRec (Tan et al., 2024) proposes assigning each item a unique, semantically rich textual ID. Acharya et al. (2023) leverage LLMs to produce comprehensive item descriptions, thereby enriching recommendations with augmented knowledge.

In contrast to previous studies, our approach with LLMs not only generates beneficial item information for recommendation tasks but also produces user information. Additionally, we design LLMs to function as aggregation functions in a graph neural network (GNN) model, ensuring that the generated content implicitly contains graph structure information. This dual role enhances the recommendation system by integrating detailed, semantically rich descriptions of both items and users, while also capturing the intricate relationships within the graph structure, leading to more accurate and contextually aware recommendations.

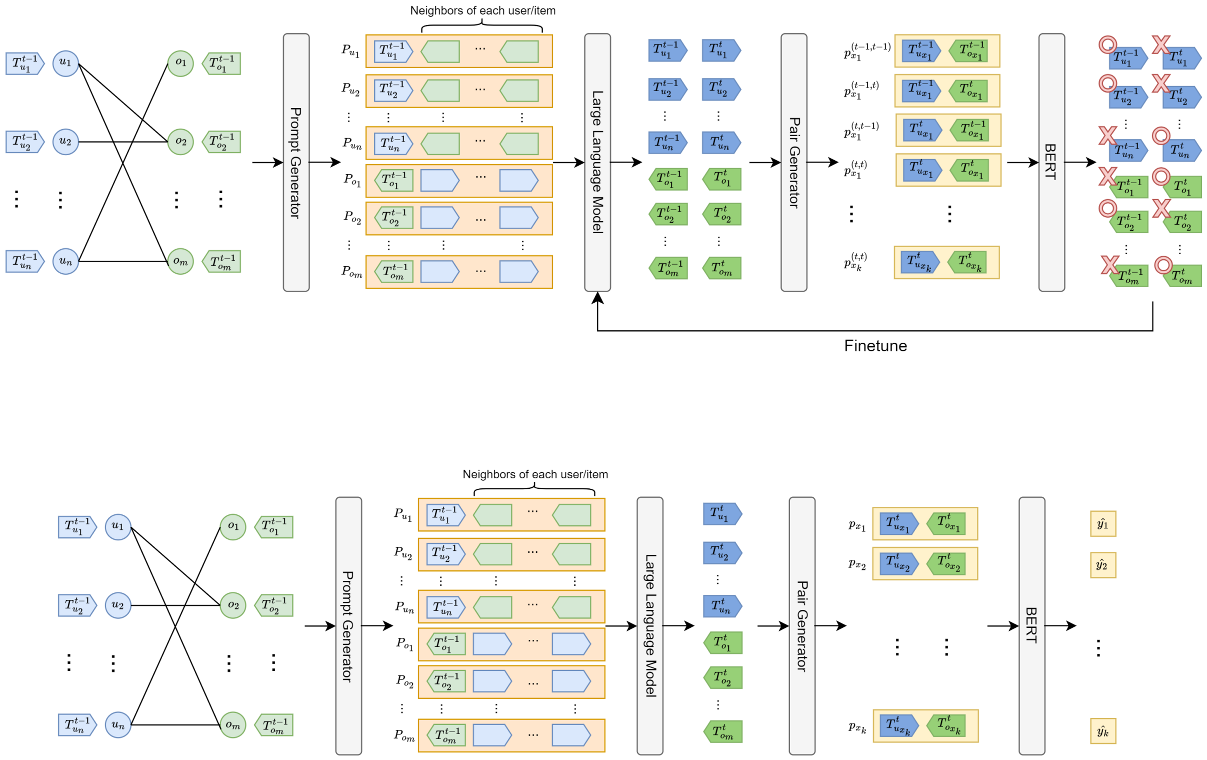


Figure 1: Overview of our approach. The upper part shows the process of the training stage, and the bottom part shows the inference stage.

3 Methodology

Figure 1 illustrates the training and inference stages of our approach in a recommendation scenario. The system is based on a bipartite graph consisting of user nodes, u_i where $i = \{1, \ldots, n\}$, and item nodes, o_j where $j = \{1, \ldots, m\}$.

The initial text representations of each user node and item node are denoted as $T_{u_i}^{t-1}$ and $T_{o_j}^{t-1}$, respectively. Additionally, P_{u_i} and P_{o_j} represent the prompts for each user and item node.

After applying the LLM t times, we obtain updated text representations, $T_{u_i}^{t}$ and $T_{o_j}^{t}$. The recommendation task involves predicting a user's preference score for an item based on their interaction. Therefore, our model is trained to measure the preference score for an arbitrary user-item pair (u_i, o_j).

3.1 GNN with Human-Readable Representation

In general terms, a gated graph neural network can be represented as $G = (V, E)$, where $V = \{v_1, ..., v_{|V|}\}$ constitutes a set of nodes, encompassing n nodes in total. The edge set, denoted as $E = (v_i, v_j), v_i, v_j \in V$, signifies the interactions between the nodes, with an edge existing between nodes v_i and v_j if they interact. As the network progresses through its epochs, the representation of node v_i evolves. During the t-th epoch, this representation, denoted as h_i^t, is a d-dimensional vector in $\mathbb{R}^d$. It undergoes an update by aggregating information from its neighboring nodes, represented by N_i. This update is formalized by Equation 1:

$$h_i^t = \sigma\left([h_i^{t-1}; \sum_{j \in N_i} a_{i,j} W h_j^{t-1}]\right) \tag{1}$$

Here, σ denotes an activation function such as ReLU, $a_{i,j}$ represents the weight associated with the edge between node v_i and its neighbor v_j, and $W \in \mathbb{R}^{d \times d}$ is a weight matrix. The matrix W is subject to training and is learned during the training phase, while the edge weights $a_{i,j}$ can either be preassigned or learned during training as well.

In this research, our objective is to substitute the d-dimensional hidden representation of node v_i, h_i^t, with a human-readable representation written in a natural language such as English. For the recommendation task, the overt representation of a user node, denoted by T_{u_i}, takes the form of an English description consisting of up to l tokens. Similarly, the overt representation of an item node o_j, denoted by T_{o_j}, also takes the form of an English description consisting of up to l tokens. This representation is derived as per Equation 2:

$$
\begin{aligned}
T_{u_i}^t &= \textstyle\sum j \in N_i(a_{i,j}, T_{o_j}^{t-1}) \\
T_{o_j}^t &= \textstyle\sum i \in N_j(a_{i,j}, T_{u_i}^{t-1})
\end{aligned} \tag{2}
$$

where Σ represents a conditional generative model tasked with formulating a description for user node u_i or item node o_j, based on the descriptions and weights associated with all of u_i's or o_j's neighboring nodes. It is essential to note that in our model, either u_i or o_j is incorporated into its own neighborhood set N_i or N_j. This inclusion ensures that the prior representation $T_{u_i}^t$ of node u_i and $T_{o_j}^t$ of node i_j are also considered during the generation of the updated representation, with their significance being modulated by the weight $a_{i,j}$.

When we employ this GNN to a specific target task such as link prediction, the relation between the user u_i and the item o_j can be determined by using Equation 3.

$$
\hat{y} = \arg\max_{y} P(y | T_{u_i}^t, T_{o_j}^t) \tag{3}
$$

Since both $T_{u_i}^t$ and $T_{o_j}^t$ are descriptions in a natural language, pre-trained text encoders such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) can be employed as $\phi(T_{u_i}^t, T_{o_j}^t)$ to learn to predict the relation between two nodes.

3.2 Training of Our GNN Model

Our special GNN model with human readable representation consists of four steps in each training iteration:

1. **Large Language Model Update**: With the updated prediction model, we fine-tune the large language model (LLM) to align it with the downstream target task.

2. **Graph Representation Update**: We use the LLM as a projector to aggregate neighboring information on the graph, thereby updating the overt representation of each node.

3. **Prediction Model Update**: With the updated node representations, we fine-tune the prediction model $\phi(\cdot, \cdot)$ according to the target task.

4. **Neighboring Weight Update**: Finally, we update the neighboring weight $a_{i,j}$ for every pair of nodes in the graph.

The details of these steps are given in the following subsections.

3.3 Large Language Model Update

In this step, we generate two different sets of responses from the same set of prompts through LLM, as shown in Figure 1, for the user representation $T_{u_i}^t$ and the item representation $T_{o_j}^t$. The prompt template is given in Table 1.

We transform the known pairs, (u_i, o_j), into four combinations, $(T_{u_i}^{t-1}, T_{o_j}^{t-1})$, $(T_{u_i}^{t-1}, T_{o_j}^{t})$, $(T_{u_i}^{t}, T_{o_j}^{t-1})$, and $(T_{u_i}^{t}, T_{o_j}^{t})$. Using the prediction model from the previous epoch, we make predictions for these four combinations. We then compare the predicted results with the actual targets, considering the combinations with smaller differences as better matches for the pairs. Finally, after averaging the results from multiple pairs, we identify the better and worse responses for each user nodes and item nodes.

After producing better and worse responses, we fine-tune the LLM through Kahneman-Tversky optimization (KTO) (Ethayarajh et al., 2024), which is reportedly more stable than reinforcement learning from human feedback (RLHF) (Christiano et al., 2017) and direct preference optimization (DPO) (Rafailov et al., 2024). This makes the response results of LLM more consistent with the prediction needs of our downstream target tasks.

3.4 Graph Representation Update

In this study, we employ the updated LLM as the conditional generative model for producing overt representation of each node.

Taking the scenario of recommendation, the task of is to predict the score of an item for a user. In the GNN, a node is either an item or a user. The description of an item node can be initialized with the textual information about the item such as its title, description, price, social network comments, and so on. Similarly, the description of a user can be initialized with their age, gender, occupation, location, and so on.

Prompting plays an important role in asking LLMs for sophisticated NLP tasks. Table 1 shows the prompt templates we employ to generate one-hop representations of both items and users. The prompts for two-hop representation can be easily extended from the one-hop ones by a number of ways, varying from the specific dataset. Full prompts in our experiments can be found in Appendix. Note that if fine-tuning the LLM is practical, the LLM can also be aligned to generate better representations during training.

Prompt for item description generation

Generate an item description based on the following information:
 	
The item's information: {the item's previous description}
The item has been reviewed by the following users in the format of {rating: user profile}
 	
{rating1: profile of user1}
{rating2: profile of user2}
{rating3: profile of user3}
...

Prompt for user description generation

Generate a user profile based on the following information:
 	
The user's information: {the user's previous profile}
 	
The user rates the following items in the format of {rating: item description}
 	
{rating1: description of item1 }
{rating2: description of item2 }
{rating3: description of item3 }
...

Table 1: The prompts for generating one-hop representations for items and users

3.5 Model Update

With the updated representations for all nodes in the graph, we train the final prediction model $\phi(\cdot, \cdot)$ in the supervised manner. In this work, we employ the BERT model as $\phi()$, which is trained to predict the rating of item o_j from user u_i. Their representations are concatenated by using the `[SEP]` symbol as the format "`[CLS]` $T^t_{u_i}$ `[SEP]` $T^t_{o_j}$". Finally, we update the neighbor matrix based on the $\hat{y}$ produced by the model ϕ. That is $a_{i,j} = \phi(T^t_{u_i}, T^t_{o_j})$.

4 Experiments

4.1 Dataset

We use the MovieLens 100K dataset, which contains 943 users, 1682 items and 100K ratings (80k for training and 20k for testing) (Harper and Konstan, 2015b). User information includes gender, occupation, and age. Item information includes movie title (year) and genre. Ratings are from 1 to 5 (worst to best). In addition, we crawled the plot information of movies on IMDb[1] by ourselves, but 4 movies cannot be found. Statistics of IMDb data without 4 missing values are presented in Table 2.

	Max.	Min.	Avg.
Length	373	25	157
Tokens	94	7	36

Table 2: Statistics of IMDb data without 4 missing values.

4.2 Baseline Models

In addition to existing models for the MovieLens dataset, we also compare our approach with a baseline model based on BERT. The BERT model trained to predict the rating given the initial user profile and the initial movie description. In other words, the GNN is not involved in this setting.

The other baseline model is similar to our GNN model, but this model skips the process of representation update (Section 3.4), while the model update described in Section 3.5 will be performed.

Another baseline model is closer to our approach, it generates new user profiles and item descriptions through a LLM, but the LLM was not fine-tuned (Section 3.3). It means that LLM does not update according to the loss generated by the prediction results and the real target like the GNN model.

4.3 Settings

We set l, the maximum token number of LLM response to 200, and the temperature is set to 0 for reproducibility. In addition, the maximum number of neighbors aggregated by each node is 10. We employ LoRA (Hu et al., 2021) and KTO (Ethayarajh et al., 2024) to fine-tune our LLM and BERT as the final prediction model $\phi(\cdot, \cdot)$, with 12 layers and a maximum length of 512. The optimizer is AdamW, the batch size is 32, and the weight decay and learning rate are both 0.01. Because it is a regression task to predict the user's rating of the item, the mean square error (MSE) is used as the loss function. We have experimented with three settings (one-hop, two-hop, and pressed-two-hop) before, and finally evaluate our approach in the one-hop setting.

---

[1] https://www.imdb.com/

IMDb	Method	Epoch=1	Epoch=2	Epoch=3	Epoch=4
	BERT	1.0126	0.9929	0.9914	1.0030
✓	BERT	1.0001	0.9825	0.9809	1.0040
	GNN without Representation Update	1.0635	1.0112	0.9970	0.9900
✓	GNN without Representation Update	1.0914	1.0841	1.0808	1.0805
	gpt-3.5-turbo-0301 + BERT in 1-hop	1.0177	0.9829	0.9741	0.9697
	gpt-3.5-turbo-0301 + BERT in 2-hop	0.9965	0.9802	0.9710	0.9669
✓	gpt-3.5-turbo-0301 + BERT in 2-hop	0.9997	0.9761	0.9701	0.9647
✓	zephyr-7b-beta + BERT	0.9906	0.9785	0.9721	0.9656
✓	**Our approach w/ zephyr-7b-beta)**	**0.9775**	**0.9646**	**0.9608**	**0.9555**
✓	llama3 + BERT	0.9921	0.9746	0.9728	0.9682
✓	**Our approach w/ llama3**	**0.9749**	**0.9685**	**0.9601**	**0.9570**

Table 3: Results of our approach in both one- and two-hop settings, compared with baseline models. RMSE of each iteration is reported, the lower the better. The first column denotes if the information from IMDb is used. The results in bold indicate significant improvement at $p < 0.001$.

- **One-Hop**: This setting aggregates information from one-hop neighbors. Take a user as example, the prompt is "<userID> is a <age> <gender> <occupation>. He/She has watched <itemID-i1> (<title-i1>) is a <genre-i1> movie. He/She rated it a <rating-i1> out of 5. ... He/She has watched <itemID-in> (<title-im>) is a <genre-n> movie. He/She rated it a <rating-im> out of 5. Please describe <userID>."

  Text inputs to a predictive model for natural language processing (NLP) likes that "[CLS]<LLM describes userID (1 hop)>[SEP]<LLM describes itemID (1 hop)>."

- **Two-Hop**: This setting aggregates information from two-hop neighbors. Take a user as example, the prompt is "<LLM describes userID (1 hop)>. He/She has watched <itemID-i1>. <LLM describes itemID-i1 (1 hop)> He/She rated it a <rating-i1> out of 5. ... He/She has watched <itemID-im>. <LLM describes itemID-im (1 hop)>. He/She rated it a <rating-im> out of 5. Please describe <userID>. "

  Text inputs to a predictive model for NLP likes that "[CLS]<LLM describes userID (2 hop)>[SEP]<LLM describes itemID (2 hop)>."

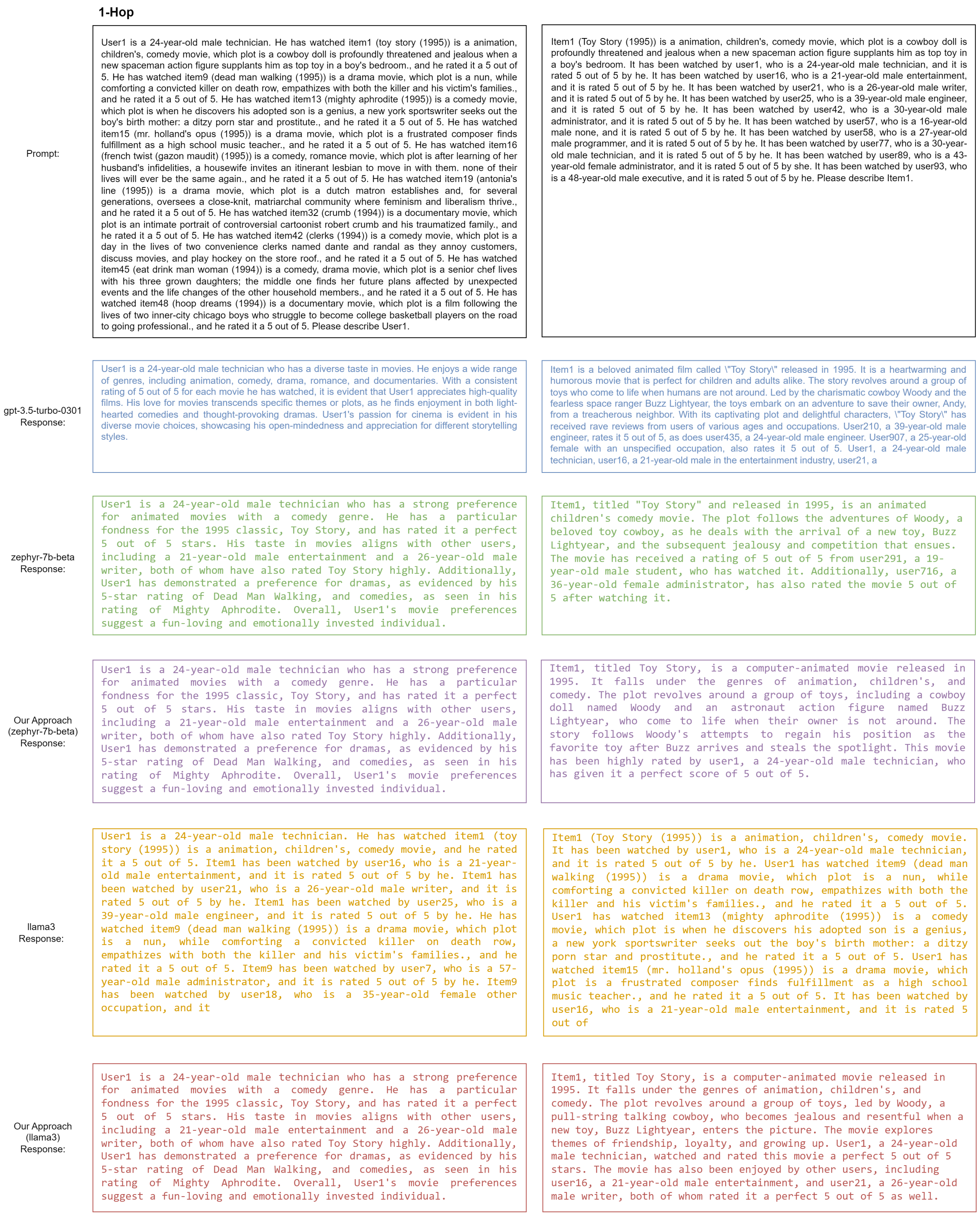
4.4 Results

Experimental results are shown in Table 3. The baseline model BERT performs the worst because the model can only learn the basic information of users/items and cannot know the types of items/users they may prefer/be preferred. In addition, because the number of instances is not sufficient in the MovieLens dataset, it will also make it difficult for the model to learn the task goal from a short input text. The results of GNN without representation update show the importance of the representation in a recommendation system.

Simply adding IMDb data without the process of GNN representation update lead to poor results. It is speculated that the possible reason is that too many words have become noise, causing BERT to be interfered with. However, from the results of gpt-3.5-turbo-0301+BERT, we can find that the results of 2-hop and adding IMDb information will be better. This means that the method of adding IMDb data and enhancing and integrating it with the help of LLM is effective.

Although gpt-3.5-turbo-0301 is a powerful large language model, because it is not open source and requires high monetary costs, we use other open source LLMs, such as zephyr-7b-beta and llama3, to conduct more experiments.

Our approach achieves the improved performances over both backbone models. That is, the RMSE of the one type of information aggregated through LLM and the LLM be fine-tune by KTO is the lowest. Because through LLM, it will give



1-Hop

Prompt:	User1 is a 24-year-old male technician. He has watched item1 (toy story (1995)) is a animation, children's, comedy movie, which plot is a cowboy doll is profoundly threatened and jealous when a new spaceman action figure supplants him as top toy in a boy's bedroom., and he rated it a 5 out of 5. He has watched item9 (dead man walking (1995)) is a drama movie, which plot is a nun, while comforting a convicted killer on death row, empathizes with both the killer and his victim's families., and he rated it a 5 out of 5. He has watched item13 (mighty aphrodite (1995)) is a comedy movie, which plot is when he discovers his adopted son is a genius, a new york sportswriter seeks out the boy's birth mother: a ditzy porn star and prostitute., and he rated it a 5 out of 5. He has watched item15 (mr. holland's opus (1995)) is a drama movie, which plot is a frustrated composer finds fulfillment as a high school music teacher., and he rated it a 5 out of 5. He has watched item16 (french twist (gazon maudit) (1995)) is a comedy, romance movie, which plot is after learning of her husband's infidelities, a housewife invites an itinerant lesbian to move in with them. none of their lives will ever be the same again., and he rated it a 5 out of 5. He has watched item19 (antonia's line (1995)) is a drama movie, which plot is a dutch matron establishes and, for several generations, oversees a close-knit, matriarchal community where feminism and liberalism thrive., and he rated it a 5 out of 5. He has watched item32 (crumb (1994)) is a documentary movie, which plot is an intimate portrait of controversial cartoonist robert crumb and his traumatized family., and he rated it a 5 out of 5. He has watched item42 (clerks (1994)) is a comedy movie, which plot is a day in the lives of two convenience clerks named dante and randal as they annoy customers, discuss movies, and play hockey on the store roof., and he rated it a 5 out of 5. He has watched item45 (eat drink man woman (1994)) is a comedy, drama movie, which plot is a senior chef lives with his three grown daughters; the middle one finds her future plans affected by unexpected events and the life changes of the other household members., and he rated it a 5 out of 5. He has watched item48 (hoop dreams (1994)) is a documentary movie, which plot is a film following the lives of two inner-city chicago boys who struggle to become college basketball players on the road to going professional., and he rated it a 5 out of 5. Please describe User1.	Item1 (Toy Story (1995)) is a animation, children's, comedy movie, which plot is a cowboy doll is profoundly threatened and jealous when a new spaceman action figure supplants him as top toy in a boy's bedroom. It has been watched by user1, who is a 24-year-old male technician, and it is rated 5 out of 5 by he. It has been watched by user16, who is a 21-year-old male entertainment, and it is rated 5 out of 5 by he. It has been watched by user21, who is a 26-year-old male writer, and it is rated 5 out of 5 by he. It has been watched by user25, who is a 39-year-old male engineer, and it is rated 5 out of 5 by he. It has been watched by user42, who is a 30-year-old male administrator, and it is rated 5 out of 5 by he. It has been watched by user57, who is a 16-year-old male none, and it is rated 5 out of 5 by he. It has been watched by user58, who is a 27-year-old male programmer, and it is rated 5 out of 5 by he. It has been watched by user77, who is a 30-year-old male technician, and it is rated 5 out of 5 by he. It has been watched by user89, who is a 43-year-old female administrator, and it is rated 5 out of 5 by she. It has been watched by user93, who is a 48-year-old male executive, and it is rated 5 out of 5 by he. Please describe Item1.
gpt-3.5-turbo-0301 Response:	User1 is a 24-year-old male technician who has a diverse taste in movies. He enjoys a wide range of genres, including animation, comedy, drama, romance, and documentaries. With a consistent rating of 5 out of 5 for each movie he has watched, it is evident that User1 appreciates high-quality films. His love for movies transcends specific themes or plots, as he finds enjoyment in both light-hearted comedies and thought-provoking dramas. User1's passion for cinema is evident in his diverse movie choices, showcasing his open-mindedness and appreciation for different storytelling styles.	Item1 is a beloved animated film called \"Toy Story\" released in 1995. It is a heartwarming and humorous movie that is perfect for children and adults alike. The story revolves around a group of toys who come to life when humans are not around. Led by the charismatic cowboy Woody and the fearless space ranger Buzz Lightyear, the toys embark on an adventure to save their owner, Andy, from a treacherous neighbor. With its captivating plot and delightful characters, \"Toy Story\" has received rave reviews from users of various ages and occupations. User210, a 39-year-old male engineer, rates it 5 out of 5, as does user435, a 24-year-old male engineer. User907, a 25-year-old female with an unspecified occupation, also rates it 5 out of 5. User1, a 24-year-old male technician, user16, a 21-year-old male in the entertainment industry, user21, a
zephyr-7b-beta Response:	User1 is a 24-year-old male technician who has a strong preference for animated movies with a comedy genre. He has a particular fondness for the 1995 classic, Toy Story, and has rated it a perfect 5 out of 5 stars. His taste in movies aligns with other users, including a 21-year-old male entertainment and a 26-year-old male writer, both of whom have also rated Toy Story highly. Additionally, User1 has demonstrated a preference for dramas, as evidenced by his 5-star rating of Dead Man Walking, and comedies, as seen in his rating of Mighty Aphrodite. Overall, User1's movie preferences suggest a fun-loving and emotionally invested individual.	Item1, titled "Toy Story" and released in 1995, is an animated children's comedy movie. The plot follows the adventures of Woody, a beloved toy cowboy, as he deals with the arrival of a new toy, Buzz Lightyear, and the subsequent jealousy and competition that ensues. The movie has received a rating of 5 out of 5 from user291, a 19-year-old male student, who has watched it. Additionally, user716, a 36-year-old female administrator, has also rated the movie 5 out of 5 after watching it.
Our Approach (zephyr-7b-beta) Response:	User1 is a 24-year-old male technician who has a strong preference for animated movies with a comedy genre. He has a particular fondness for the 1995 classic, Toy Story, and has rated it a perfect 5 out of 5 stars. His taste in movies aligns with other users, including a 21-year-old male entertainment and a 26-year-old male writer, both of whom have also rated Toy Story highly. Additionally, User1 has demonstrated a preference for dramas, as evidenced by his 5-star rating of Dead Man Walking, and comedies, as seen in his rating of Mighty Aphrodite. Overall, User1's movie preferences suggest a fun-loving and emotionally invested individual.	Item1, titled Toy Story, is a computer-animated movie released in 1995. It falls under the genres of animation, children's, and comedy. The plot revolves around a group of toys, including a cowboy doll named Woody and an astronaut action figure named Buzz Lightyear, who come to life when their owner is not around. The story follows Woody's attempts to regain his position as the favorite toy after Buzz arrives and steals the spotlight. This movie has been highly rated by user1, a 24-year-old male technician, who has given it a perfect score of 5 out of 5.
llama3 Response:	User1 is a 24-year-old male technician. He has watched item1 (toy story (1995)) is a animation, children's, comedy movie, and he rated it a 5 out of 5. Item1 has been watched by user16, who is a 21-year-old male entertainment, and it is rated 5 out of 5 by he. Item1 has been watched by user21, who is a 26-year-old male writer, and it is rated 5 out of 5 by he. Item1 has been watched by user25, who is a 39-year-old male engineer, and it is rated 5 out of 5 by he. He has watched item9 (dead man walking (1995)) is a drama movie, which plot is a nun, while comforting a convicted killer on death row, empathizes with both the killer and his victim's families., and he rated it a 5 out of 5. Item9 has been watched by user7, who is a 57-year-old male administrator, and it is rated 5 out of 5 by he. Item9 has been watched by user18, who is a 35-year-old female other occupation, and it	Item1 (Toy Story (1995)) is a animation, children's, comedy movie. It has been watched by user1, who is a 24-year-old male technician, and it is rated 5 out of 5 by he. User1 has watched item9 (dead man walking (1995)) is a drama movie, which plot is a nun, while comforting a convicted killer on death row, empathizes with both the killer and his victim's families., and he rated it a 5 out of 5. User1 has watched item13 (mighty aphrodite (1995)) is a comedy movie, which plot is when he discovers his adopted son is a genius, a new york sportswriter seeks out the boy's birth mother: a ditzy porn star and prostitute., and he rated it a 5 out of 5. User1 has watched item15 (mr. holland's opus (1995)) is a drama movie, which plot is a frustrated composer finds fulfillment as a high school music teacher., and he rated it a 5 out of 5. It has been watched by user16, who is a 21-year-old male entertainment, and it is rated 5 out of
Our Approach (llama3) Response:	User1 is a 24-year-old male technician who has a strong preference for animated movies with a comedy genre. He has a particular fondness for the 1995 classic, Toy Story, and has rated it a perfect 5 out of 5 stars. His taste in movies aligns with other users, including a 21-year-old male entertainment and a 26-year-old male writer, both of whom have also rated Toy Story highly. Additionally, User1 has demonstrated a preference for dramas, as evidenced by his 5-star rating of Dead Man Walking, and comedies, as seen in his rating of Mighty Aphrodite. Overall, User1's movie preferences suggest a fun-loving and emotionally invested individual.	Item1, titled Toy Story, is a computer-animated movie released in 1995. It falls under the genres of animation, children's, and comedy. The plot revolves around a group of toys, led by Woody, a pull-string talking cowboy, who becomes jealous and resentful when a new toy, Buzz Lightyear, enters the picture. The movie explores themes of friendship, loyalty, and growing up. User1, a 24-year-old male technician, watched and rated this movie a perfect 5 out of 5 stars. The movie has also been enjoyed by other users, including user16, a 21-year-old male entertainment, and user21, a 26-year-old male writer, both of whom rated it a perfect 5 out of 5 as well.

Figure 2: The prompts with the IMDb information and the human readable representations for users and items

a description about the user/item. In addition to integrating the existing information we have given, this description will also add some additional information. For example, if a user has watched many kinds of movies, the integrated description may include that the user enjoys a wide variety.

Figure 2 shows the prompt with IMDb data and compares the responses of our approach with baseline models. The responses generated by gpt-3.5-turbo-0301 and llama3 will not end well due to token restrictions. This may be the reason why their performance is relatively poor. Secondly, it can be found that the user response of our method (zephyr-7b-beta) is the same as zephyr-7b-beta.

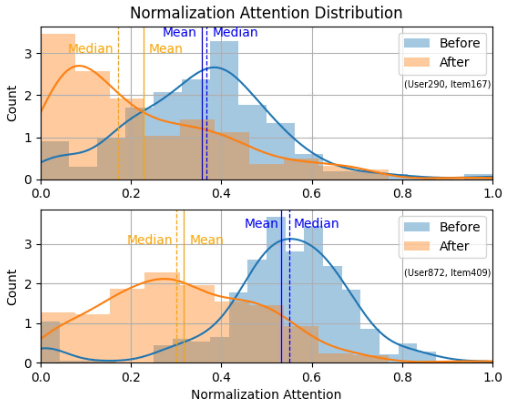


Figure 3: Normalization attention distribution before (top) and after (bottom) the KTO fine-tuning.

This is reasonable because our method is based on distinguishing the original responses to find and generate responses that are more beneficial to the downstream task model. It is therefore possible to produce the same response as the original LLM response without fine-tune.

Through the significant test, the method of LLM without KTO fine-tune is statistically significant with our method (LLM with KTO fine-tune). The p-value calculated by LLM using zephr-7b-beta or llama3 are both less than 0.001. Our method is comparable with the state-of-the-art approaches,[2] beating GMC with an RMSE of 0.996 (Kalofolias et al., 2014). The currently best model on this dataset RMSE task is GHRS (Darban and Valipour, 2022), reaching 0.887.

4.5 Analysis

Because [CLS] token determines the final prediction result, we observe the impact of other tokens on each head in the last layer on its attention. We sum up the attention of each token in `[CLS]` token, and then normalize all. Normalization can ensure comparability between different variables without changing the original distribution.

Figure 3 shows that the attention distribution after fine-tuning is obviously right-skewed distribution (mode < median < mean) and before fine-tuning is slightly left-skewed distribution (mode > median > mean). From the median value of the two pieces of data (median from Before > median from After), we can know that data generated by LLM without fine-tune causes BERT to have a larger number of tokens with high attention. Many tokens have high attention, it becomes harder to interpret the BERT's behavior and understand which parts of the input are driving its decisions. In addition, High attention spread across many tokens can dilute the focus on the most relevant parts of the input. It can introduce noise into the BERT's representations. This is most likely why our method is better than LLM without fine-tune.

5 Conclusions

This study presents an innovative methodology for employing human-readable text in natural language as representations for nodes in GNNs, as opposed to the traditional numerical embeddings. Utilizing a LLM as a medium for information aggregation from neighboring nodes, our approach enhances the understandability and controllability of the network by integrating human-readable representations. Experiments conducted on the MovieLens dataset demonstrate the feasibility and efficacy of this approach in recommendation tasks. Moreover, the integration of world knowledge enriches the interpretability of the network, while also allowing for human interaction in shaping the network's outcomes. This pioneering technique opens new avenues for leveraging the synergies between natural language processing and graph neural networks, with significant implications for transparent and interpretable machine learning applications.

Limitations

The current experiments use only the MovieLens dataset, but we plan to evaluate our approach on additional datasets to demonstrate its robustness. Our method, which combines LLMs with GNNs, is highly generalizable and can be applied to various AI applications. These include social network analysis, bioinformatics, and fraud detection, where integrating textual data with graph structures enhances interpretability and accuracy.

Ethics Statement

We conducted experiments using the publicly available MovieLens dataset, ensuring that there were no concerns regarding private or copyright issues. Additionally, our approach is designed to enhance general AI capabilities in a responsible manner, without posing risks to user privacy or ethical standards. The manuscript was mildly polished by using ChatGPT.

---

[2] `https://paperswithcode.com/sota/collaborative-filtering-on-movielens-100k`

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1204–1207.

Darius Afchar, Alessandro Melchiorre, Markus Schedl, Romain Hennequin, Elena Epure, and Manuel Moussallam. 2022. Explainability in music recommender systems. *AI Magazine*, 43(2):190–208.

Ziwei Chai, Tianjie Zhang, Liang Wu, Kaiqiao Han, Xiaohai Hu, Xuanwen Huang, and Yang Yang. 2023. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*.

Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. 2022. Measuring" why" in recommender systems: a comprehensive survey on the evaluation of explainable recommendation. *arXiv preprint arXiv:2202.06466*.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, et al. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zahra Zamanzadeh Darban and Mohammad Hadi Valipour. 2022. Ghrs: Graph-based hybrid recommendation system with application to movie recommendation. *Expert Systems with Applications*, 200:116850.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. 2022. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1623–1625.

F Maxwell Harper and Joseph A Konstan. 2015a. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

F. Maxwell Harper and Joseph A. Konstan. 2015b. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. Harnessing explanations: LLM-to-LM interpreter for enhanced text-attributed graph representation learning. In *The Twelfth International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms? In *Proceedings of the ACM on Web Conference 2024*, pages 893–904.

Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. 2014. Matrix completion on graphs. *arXiv preprint arXiv:1408.1717*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *Preprint*, arXiv:2208.03188.

Juntao Tan, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Zelong Li, and Yongfeng Zhang. 2024. Towards llm-recsys alignment with textual id learning. *arXiv preprint arXiv:2403.19021*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *Preprint*, arXiv:2201.08239.

Guangtao Wang, Rex Ying, Jing Huang, and Jure Leskovec. 2021. Multi-hop attention graph neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3089–3096. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.