Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Project Plan

Cognitive-Enhanced Knowledge Tracing: A Unified Framework Integrating LLM Semantics, Adaptive Forgetting, and Mixture-of-Experts

by

Yilin Jing

Email: yilin.jing24@imperial.ac.uk
GitHub username: yj1024
Repository: https://github.com/ese-ada-lovelace-2024/irp-yj1024

Supervisors:

Ms. Wenxia Yang
Dr. Simon Warder

June 2025

Academic Integrity Declaration

I confirm that I have read, understood, and will comply with the IRP Academic Integrity requirements as outlined in the official guidelines. This project plan represents my own work, understanding, and academic effort. I acknowledge that:

- All submitted work must be my own and reflect my personal knowledge and technical skills
- I will maintain regular version control with meaningful commits throughout the project
- I will attend all scheduled meetings with supervisors and maintain a detailed logbook
- Any use of external resources, including code, literature, or AI tools, will be properly cited and disclosed
- I understand that I may be required to defend this work in a viva or authenticity interview

I commit to upholding the highest standards of academic integrity throughout this Independent Research Project.

AI Acknowledgement Statement

No generative AI tools were used in the preparation of this project plan. All content, ideas, and written work represent my own understanding and effort.

Abstract

Despite recent advances in educational AI, particularly the Knowledge Modeling and Material Prediction (KMaP) framework [1], current systems struggle with three fundamental limitations: inadequate semantic understanding of educational content, oversimplified knowledge forgetting models, and static student profiling approaches. This project proposes a novel enhanced KMaP framework that addresses these limitations through three key innovations: (1) Large Language Model (LLM) semantic embeddings using sentence-BERT to capture rich conceptual relationships in learning materials, (2) adaptive time-decay gating mechanisms that model personalized knowledge forgetting patterns based on cognitive science principles [2, 3], and (3) a gated Mixture-of-Experts (MoE) architecture for dynamic student profiling that adapts to evolving learning behaviors [4, 5]. Our approach targets measurable improvements on the EdNet dataset [6]: a minimum 1.5 percentage points (pp) enhancement in Mean Reciprocal Rank (MRR) and a 1 pp improvement in early-stage Area Under Curve (AUC) compared to baseline KMaP. This work represents a significant advancement in personalized educational technology by bridging cognitive science theories with modern deep learning architectures.

1 Introduction

Online education has experienced unprecedented growth, but ”one-size-fits-all” approaches lead to high dropout rates [7]. Personalized learning, powered by Knowledge Tracing (KT), aims to solve this by modeling student knowledge states over time. The field has evolved from early Bayesian models (BKT) [8] to deep learning architectures like DKT [9], DKVMN [10], and attention-based models such as SAKT [11] and AKT [12], which leverage Transformer architectures [13].

The current state-of-the-art is the Knowledge Modeling and Material Prediction (KMaP) framework [1], which excels at simultaneously modeling knowledge and predicting learning materials. Despite its strengths, our analysis reveals three critical gaps that limit its effectiveness, leading to suboptimal learning outcomes:

- **Semantic Representation Gap:** KMaP uses traditional embeddings that fail to capture rich semantic relationships between learning materials (e.g., ”linear equations” vs. ”algebraic expressions”). This limits recommendation accuracy for conceptually similar but presentationally different content, causing the system to miss valuable learning opportunities.
- **Oversimplified Forgetting Mechanisms:** The model lacks a sophisticated model of knowledge forgetting, a key cognitive principle [14]. It uses fixed temporal dynamics, ignoring individual differences in memory retention, which affects long-term prediction quality and fails to schedule reviews effectively, particularly for students with irregular study patterns.
- **Static Student Profiling:** KMaP’s student profiles are static. They are established once and do not adapt to a learner’s evolving behaviors, preferences, and cognitive states over time. This rigidity prevents the system from responding to changes in a student’s learning strategies or engagement levels.

To address these limitations, this research makes three novel contributions to educational AI. We introduce: (1) **LLM-Enhanced Content Representation** using sentence-BERT [15] to capture deep semantic meaning; (2) **Cognitive-Inspired Adaptive Forgetting** via personalized time-decay gates based on cognitive science [16, 2]; and (3) **Dynamic Student Profiling** with a gated Mixture-of-Experts (MoE) architecture [17, 18] that adapts to evolving learning behaviors. Our work pioneers the integration of these modern techniques into a unified KT framework, aiming to create a more effective and truly personalized learning experience.

2 Research Objectives and Success Criteria

2.1 Research Objectives

This project aims to enhance the KMaP framework through three specific, measurable objectives:

Objective 1: Semantic Enhancement

- Integrate sentence-BERT embeddings for material representation.
- Improve semantic similarity capture by 8-12%.
- Validate conceptual modeling via similarity analysis.

Objective 2: Adaptive Forgetting Implementation

- Develop personalized time-decay gating mechanisms.
- Improve long-term knowledge prediction accuracy by 5-10%.
- Validate forgetting patterns against psychological models.

Objective 3: Dynamic Student Profiling

- Implement a gated Mixture-of-Experts architecture.
- Improve capture of behavioral evolution by 10-15%.
- Enable real-time profile adaptation to context changes.

2.2 Quantitative Success Criteria

Building upon KMaP's baseline performance, we target specific improvements on the EdNet dataset [6]:

- **Material Recommendation:** $>$1.5 pp improvement in Mean Reciprocal Rank (MRR).
- **Knowledge Prediction:** $>$1 pp enhancement in early-stage Area Under Curve (AUC).
- **Behavioral Modeling:** 8-12% improvement in capturing student preference evolution.
- **Cross-Dataset Validation:** Consistent improvements on the Junyi Academy dataset.

2.3 Technical Validation

- **Ablation Studies:** Systematic evaluation of each innovation's contribution.
- **Statistical Significance:** All improvements validated with $p < 0.05$.
- **Computational Efficiency:** Maintain inference time within 20% of baseline KMaP.
- **Reproducibility:** Publish code and documentation for full reproducibility.

3 Methodology

3.1 Framework Architecture

Our strategy is to enhance the proven KMaP architecture modularly. We augment the material embedding layer with sentence-BERT representations, integrate adaptive time-decay gates into the LSTM-based knowledge model, and replace static clustering with a dynamic gated Mixture-of-Experts (MoE) architecture for student profiling. This approach systematically addresses KMaP's limitations while preserving its core multi-task learning structure.

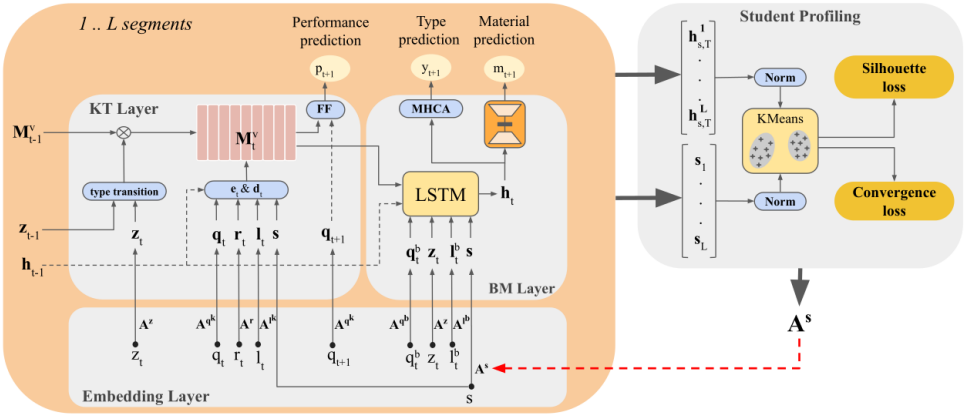

Figure 1: Original KMaP model architecture (from [1]). Our enhanced framework builds upon this foundation by integrating: (1) LLM semantic embeddings in the material representation layer, (2) adaptive forgetting mechanisms within the LSTM components, and (3) gated MoE architecture for dynamic student profiling in place of static clustering.

3.2 Innovation 1: Semantic Embeddings with LLMs

Implementation Strategy: We replace traditional categorical encodings with 768-dimensional semantic representations from sentence-BERT [15]. We will pre-process all material texts to generate these embeddings offline. The resulting dense vectors will be concatenated with KMaP's existing categorical features and fed into the model's embedding layer, creating a hybrid representation. **Justification:** Sentence-BERT has demonstrated superior performance in capturing semantic similarity [15]. This approach enables the model to recommend materials based on deeper conceptual understanding rather than superficial categorical matches, addressing a key limitation in current systems.

3.3 Innovation 2: Adaptive Forgetting Gates

Implementation Details: Inspired by cognitive science [14, 16], particularly the testing effect [19], we implement learnable gating functions to model personalized knowledge decay. The forgetting gate f_t is defined as:

$$f_t = \sigma(W_f \cdot [h_t, \Delta t, d_m, p_s] + b_f) \tag{1}$$

where the gate is conditioned on the current hidden state (h_t), time since last interaction (Δt), material difficulty (d_m), and a student-specific forgetting profile (p_s). The profile p_s will be a learnable embedding vector, initialized by clustering students based on their historical interaction patterns. **Justification:** Psychological research shows that forgetting patterns vary by individual and content [16, 20]. Our adaptive mechanism, informed by recent computational models [2, 3], moves beyond fixed decay rates to make more accurate long-term predictions about knowledge retention and optimize review timing.

3.4 Innovation 3: Dynamic Profiling with Gated MoE

Architecture Design: To overcome the limitations of static student profiles, we employ a gated Mixture-of-Experts (MoE) architecture [21, 17, 18]. The model will consist of 4-6 specialized expert networks (small feed-forward networks) to capture distinct behavioral patterns. A lightweight, attention-based gating network will take the student's current context (e.g., recent performance, session length) and compute a soft assignment weight for each expert. The final output is a weighted sum of the expert outputs:

$$y = \sum_{i=1}^{N} G(x)_i \cdot E_i(x) \tag{2}$$

where $G(x)$ is the gating function and $E_i(x)$ are the expert networks. **Justification:** Static profiling is inadequate for the dynamic nature of student learning. MoE architectures are proven for handling heterogeneous and evolving user behaviors in recommender systems and multi-task learning [5, 22], including large-scale industrial applications like YouTube video recommendations [23]. This allows our model to adapt to individual learning preferences and behavioral shifts in real-time.

3.5 Integration and Training

Multi-task Learning Framework: The model will be trained end-to-end using a multi-task loss function. This loss will be a weighted sum of a primary cross-entropy loss for knowledge tracing, a contrastive loss for material prediction, and auxiliary losses designed to guide the forgetting and MoE components. We will employ a progressive training strategy inspired by curriculum learning [24]: first, train the baseline KMaP model; then, sequentially unfreeze and fine-tune each of the three innovative components. This staged approach ensures stable convergence and allows for clear ablation studies to validate each component's contribution.

Evaluation Protocol:

- **Datasets:** EdNet (primary), Junyi Academy (validation).
- **Baselines:** Original KMaP, DKT, AKT, SAKT.
- **Metrics:** MRR, precision@k, AUC, behavioral consistency scores.
- **Statistical Testing:** Paired t-tests for significance validation.
- Target inference time within 100ms for practical deployment.

4 Project Plan

The project follows a systematic 12-week implementation plan with four distinct phases, as illustrated in Figure 2. The timeline ensures systematic development from foundation setup through final evaluation and documentation.

The implementation strategy is designed for modularity and risk mitigation. Each of the three core innovations will be developed and validated in separate work packages before being integrated into the final framework. This approach simplifies debugging, allows for parallel development where feasible, and enables clear, incremental validation of each component's contribution to overall performance.

Risk Mitigation Strategies:

- **Computational Complexity:** Pre-compute sentence-BERT embeddings offline; implement efficient batch processing and sparse MoE routing.
- **Model Convergence:** Use established initialization strategies for MoE; maintain fallback to simpler profiling if needed.
- **Performance Validation:** Implement incremental improvement tracking; prepare alternative optimization strategies if targets are not met.
- **Timeline Management:** Maintain a schedule buffer; design simplified fallback implementations for complex components.

4.1 Supervision and Project Management

Meeting Schedule: I will meet with my main supervisor Dr. Wenxia Yang weekly and with co-supervisor Dr. Deborah Pelacani Cruz monthly. All meetings will be conducted in person when possible, with

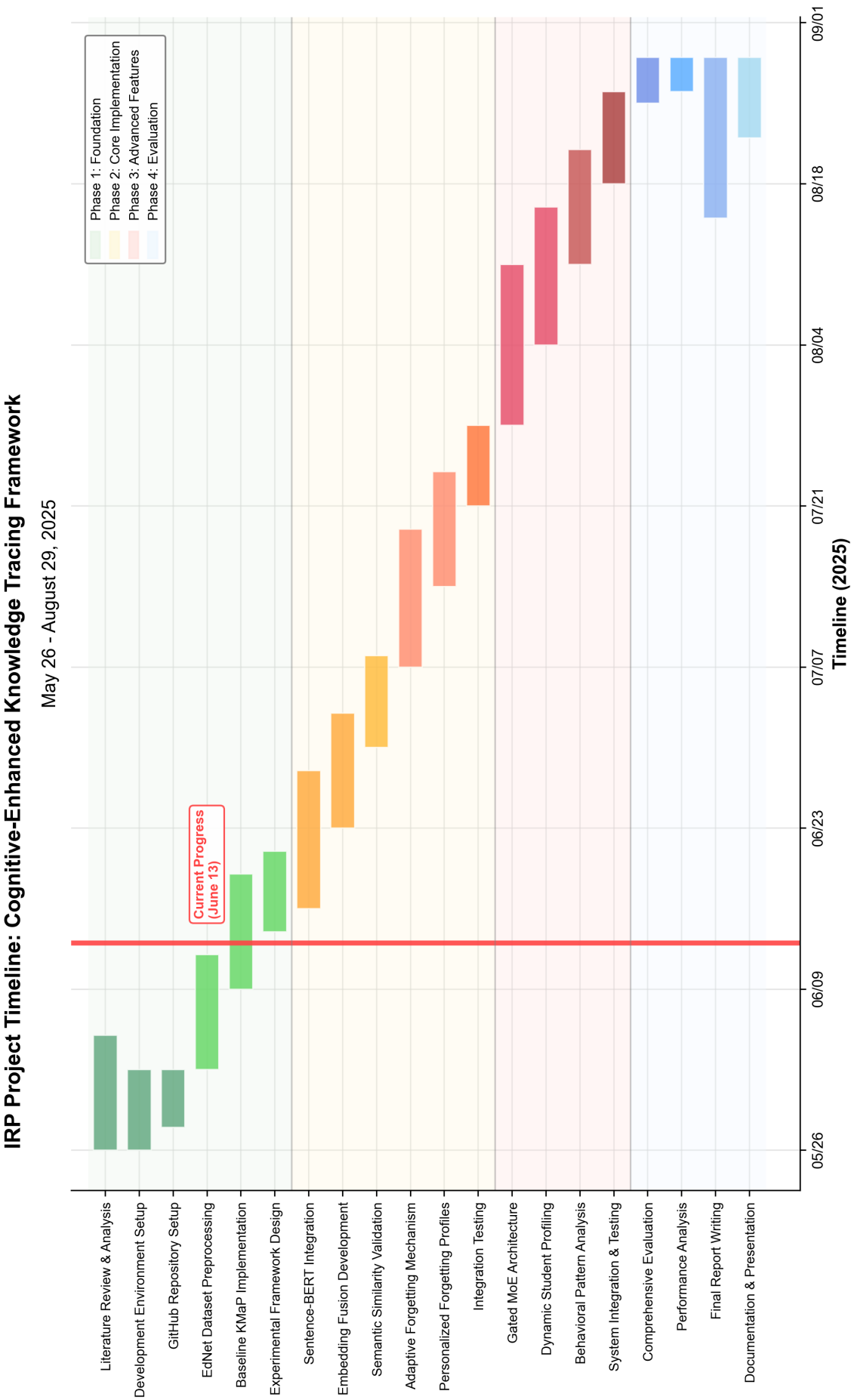

Figure 2: Project timeline showing four-phase implementation with milestone tracking.

camera-on video calls as backup. I will maintain an active role in discussions, present progress updates, and seek guidance on technical challenges.

Version Control Strategy: I commit to regular, meaningful commits to the IRP repository with descriptive messages. The repository will include:

- Weekly commits of code development and experimental results
- Bi-weekly commits of report drafts and documentation updates
- Detailed commit messages describing specific changes and progress
- Proper branching strategy for different development phases
- Regular backup of all project materials to ensure reproducibility

Logbook Maintenance: After each supervisory meeting, I will update the `logbook.md` file in the repository with meeting details, feedback received, and action items for the next period. This will serve as a transparent record of project evolution and supervisory engagement.

References

[1] Soroush Hashemifar and Sherry Sahebi. Personalized student knowledge modeling for future learning resource prediction. *arXiv preprint arXiv:2505.14072*, 2025.

[2] Alyssa Shuang Sha, Bernardo Pereira Nunes, and Armin Haller. "forgetting" in machine learning and beyond: A survey. *arXiv preprint arXiv:2405.20620*, 2024.

[3] Shanshan Wang, Ying Hu, Xun Yang, Zhongzhou Zhang, Keyang Wang, and Xingyi Zhang. Personalized forgetting mechanism with concept-driven knowledge tracing. *arXiv preprint arXiv:2404.12127*, 2024.

[4] Zhaoxing Li, Vahid Yazdanpanah, Jindi Wang, Wen Gu, Lei Shi, Alexandra I Cristea, Sarah Kiden, and Sebastian Stein. Tutorllm: Customizing learning recommendations with knowledge tracing and retrieval-augmented generation. *arXiv preprint arXiv:2502.15709*, 2025.

[5] Jie Zou, Cheng Lin, Weikang Guo, Zheng Wang, Jiwei Wei, Yang Yang, and Hengtao Shen. Multi-type context-aware conversational recommender systems via mixture-of-experts. *arXiv preprint arXiv:2504.13655*, 2025.

[6] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. Ednet: A large-scale hierarchical dataset in education. *Artificial Intelligence in Education*, pages 69–73, 2020.

[7] George Siemens and Ryan SJD Baker. Learning analytics: The emergence of a discipline. *American behavioral scientist*, 57(10):1380–1400, 2013.

[8] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[9] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

[10] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. *Proceedings of the 26th international conference on world wide web*, pages 765–774, 2017.

[11] Shalini Pandey and George Karypis. Self-attentive model for knowledge tracing. *Proceedings of the 12th International Conference on Educational Data Mining*, pages 384–389, 2019.

[12] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. Context-aware attentive knowledge tracing. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2330–2339, 2020.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Teachers college, Columbia university*, 1885.

[15] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[16] John T Wixted. The psychology and neuroscience of forgetting. *Annual review of psychology*, 55:235–269, 2004.

[17] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformer: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[18] Evan Shelhamer, Hanzi Zhang, Sheng Liu, Bowen Deng, Li Fei-Fei, and Ranjay Krishna. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*, 2022.

[19] Jeffrey D Karpicke and Henry L Roediger. The critical importance of retrieval for learning. *Science*, 319(5865):966–968, 2008.

[20] Jaap MJ Murre and Joeri Dros. Replication and analysis of ebbinghaus' forgetting curve. *PLoS one*, 10(7):e0120644, 2015.

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

[22] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939, 2018.

[23] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Anirudh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. Recommending what video to watch next: a multitask ranking system. *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 43–51, 2019.

[24] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.