

Master Thesis Review

Charles University, Faculty of Mathematics and Physics

Thesis author Karen Jia-Hui Li

Thesis title A Diet-Coaching Chatbot with Neural Language Models

Submitted 2025

Program Computer Science **Specialization** Computational Linguistics

Review author Dr. Saad Mahamood **Role** reviewer

Position Lead Data Scientist, trivago N.V.

Review text:

This thesis builds upon existing work of using a chatbot to communicate personalised diet and nutrition information to recipients. The main contribution in this thesis is the construction and evaluation of two new system variants, which explores the possible impact large models can have in generating such messages to recipients in terms of diet outcomes, emotional well-being, and whether the response generated are seen as improvement over the existing template-based system. The first variant leverages a rule-based chat-bot with a fine-tuned large language model for nutritional counselling, while the second variant leverages prompt-engineering to rephrase templated messages from the existing chatbot system that this work builds upon. The student performs an extrinsic task-based evaluation with human participants to evaluate the efficacy, for different diet related goals, of each of the three systems against each other through a seven-week randomised controlled trial (RCT). While the results are inconclusive, as there are no clear system differences in terms of dietary improvements or emotional-wellbeing for the participants, the student demonstrates conclusively that the limitations of the nutritional chatbot architecture is the main factor for the limited impact with the trial participants.

The thesis on the whole is well written and well structured. In chapter 2, the student introduces background literature on the architecture of pre-trained language models for natural language generation (NLG) with discussion on transformer based models and the use of fine-tuning and prompt engineering. The background literature section concludes with a discussion on evaluating NLG systems with a particular focus comparing automatic with human based approaches. Overall, there are no problems in this chapter, however the statement that “Automatic metrics provide a cheap and quick method for intrinsically evaluating NLG systems...” is an overgeneralisation. A distinction needs to be applied between reference-based and reference-less automatic metrics. Only the latter, not the former, are quick and cheap to run.

The next chapter, chapter 3, focuses on background literature on the healthcare side, focusing

squarely on digital health interventions (DHI) due to increasing shift towards more accessible patient-centred healthcare. However, what is not mentioned is the broader context of demographics. With aging populations, longer-term there is an intrinsic need for more DHIs to serve this population. Whilst the chapter correctly focuses on the use of chatbots and evaluation within the nutrition and counselling domain, I find the lack of discussion of past NLG based research in healthcare a significant omission from this chapter and leaves chapter 2 and chapter 3 disconnected from each other.

Whilst chapter 4 discussed the system architecture of the original baseline system, chapter 5, focuses on the experimentation and intrinsic evaluation of both the fine-tuned and rephrasing system variants. The discussion in section 5.3 about optimising decoding and the decision to use a 4-bit quantised Llama 3 model, as the most performant model, is interesting. However, there is no exploration on whether the choice to optimise for speed had lead to any other potential tradeoffs in terms of output quality.

The human evaluations conducted in chapter 5 and 6 also have some missing details. For example, the intrinsic human evaluation provides no details with respect to the compensation paid to the recruited crowd workers. Additionally, in both the evaluations in chapters 5 and 6 don't declare whether any piloting was done by the student prior to launching the evaluations.

Overall I consider this a good thesis, irrespective of the few issues pointed out above and shows an extensive amount of experimental work for a master thesis. The motivation and approach undertaken is robust and shows a good methodological approach. Conducting extrinsic task-based evaluations are unfortunately uncommon in our field and the fact this was undertaken in this work should be strongly commended. The results obtained, whilst negative, are still informative and demonstrates the challenge of building meaningful digital healthcare solutions. Such solutions need to solve problems of significant importance, relative to the recipient, by providing insights and information that the recipient cannot obtain or infer themselves. The student recognise some of these short comings of the current system with their recommendations for incorporation of more information, personalisation, and the possibility of adding some gamification elements.

I recommend the thesis for defense.

I suggest to not consider the thesis for the annual award.

In Düsseldorf, Germany, 26.01.2025

Signature:

