On the Use of GPT-4 for Creating Goal Models: An Exploratory Study

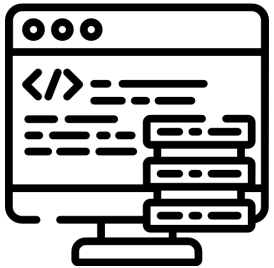
Authors: Boqi Chen, Kua Chen, Shabnam Hassani, Yujing Yang, Daniel Amyot, Lysanne Lessard, Gunter Mussbacher, Mehrdad Sabetzadeh, Dániel Varró

Presenter: Kua Chen

Table of Contents

Challenges

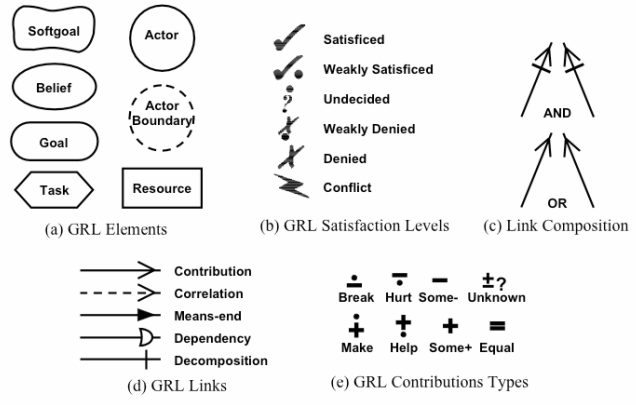
- **Goal modeling knowledge.**
  - **Describe concepts on goal modeling.**
  - Explain the difference between a contribution and a correlation in GRL.



- **Goal model creation.**
  - **Generate goal model from text description.**
  - Using the TGRL, please provide a goal model for a Kids Help Phone application.

Table of Contents

Background - GRL



- The **Goal-oriented Requirement Language** (GRL)
  - Standardized goal modeling language.
  - Includes actors, intentional elements, links.
  - Has a graphical notation and textual notation (TGRL)
- Creating a goal model from scratch can be tedious.
  - Usually done manually.

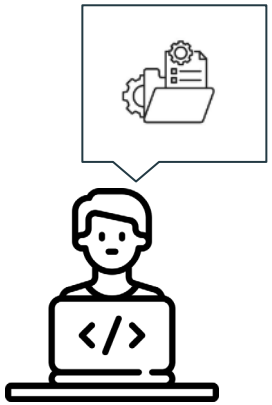


Figure top 1 from: https://www.researchgate.net/figure/Elements-of-the-Goal-Oriented-Requirements-Language_fig1_245591623

Background - Large language models

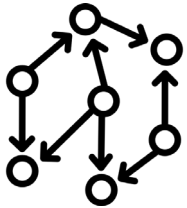
- **Large language models** (LLMs) are a type of natural language processing (NLP) application originally designed for text generation.
- LLMs are also widely used for **software engineering practices**.
- Basic mechanism: given a sequence of tokens, LLMs predict next token.
- The very initial input sequence of tokens is called **Prompt.**

Figure left 1 from: https://commons.wikimedia.org/wiki/File:ChatGPT_logo.svg
Figure right 1 from: J. Horkoff and E. Yu, "Interactive goal model analysis for early requirements engineering," Requir. Eng., vol. 21, no. 1, p. 29–61, March 2016. [Online]. Available: https://doi.org/10.1007/s00766-014-0209-8

Table of Contents

Research Questions

RQ1: How much **goal modeling knowledge** does GPT-4 preserve?



RQ2: How does GPT-4 perform in **goal model generation** from textual descriptions with different levels of detail?

RQ3: How does **immediate interactive feedback** affect the quality of goal models generated by GPT-4? (Read it in our paper)

Experiment B - Baseline Knowledge

- Check baseline knowledge through **direct questions** on goal modeling.
  - **Concept**: focus on the meaning of one or more goal-modeling concepts.
  - **Application**: apply concepts to solve some tasks.
  - **Open**: has many correct answers.
  - **Closed**: has a unique answer.

TABLE I: Example Questions for each Category for Experiment B

Concept	Open	Explain the difference between a softgoal and a goal in GRL.
	Closed	What are all the types of qualitative contributions supported by GRL? Provide a one-sentence description for each of them.
Application	Open	Give me a sample goal model in the Goal-oriented Requirement Language (GRL), with 2 actors that have several goals each, as well as relationships.
	Closed	Create a small GRL model (with one goal linked to as many indicators as you need) that determines...

Experiment B - Baseline Knowledge

- Context prompt: *You are a software engineering student on an exam for goal-oriented requirement engineering.*

  Follow the instructions in the question and answer the following question concisely.
- **R1**: no context prompt. **R2**: with context prompt.
- Four authors manually grade each response from {0,1,2,3,4,5}
- Context prompt **improves** GPT-4's performance on the question answering.

TABLE III: Average Score of all Questions for Experiment B

	Run 1	Run 2	Run 3	Run 4	SD
R1	3.00	2.97	2.39	2.44	0.29
R2	3.39	3.75	3.78	3.86	0.18

Experiment B - Baseline Knowledge

- Performance on **Concept questions** is significantly improved by the context prompt.
- **Closed Concept** questions are very challenging to GPT-4.
- GPT-4 excels at **Open** questions due to its generative nature.

TABLE IV: Results of Two Rounds for Experiment B

R1	**Application**	**Concept**
Open	4.38 ± 0.26	3.25 ± 2.30
Closed	2.13	1.60 ± 1.48
R2	**Application**	**Concept**
Open	4.23 ± 0.46	4.75 ± 0.27
Closed	2.5	3.18 ± 1.72

Experiment B - Baseline Knowledge

RQ1: How much goal modeling knowledge does GPT-4 preserve?

We find that GPT-4 preserves considerable knowledge on goal modeling.

Experiment K - Kids Help Phone

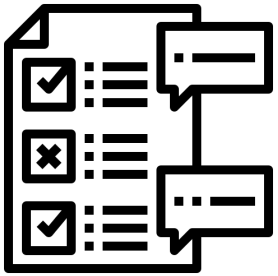
- Evaluate how well GPT-4 can create a TGRL **goal model** from scratch.
- Kids Help Phone application provides online counselling for Canadian children.
  - Prompt 1: Single Sentence
  - Prompt 2: Single Sentence + Domain Paragraph
  - Prompt 3: Syntax Description + Single Sentence (Prompt 1);
  - Prompt 4: Syntax Description + Single Sentence + Domain Paragraph (Prompt 2);
- Evaluate element based on four categories:

  Correct, Partially Correct, Incorrect, Reasonable

Experiment K - Kids Help Phone

Longer prompt

→ Better result**?**

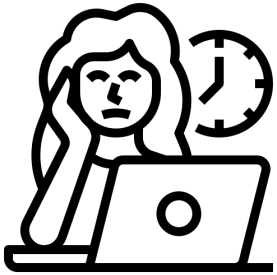


Low coverage

for elements.

Syntax prompt is

helpful.



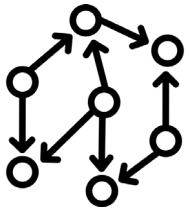
Reasonable elements:

60% too **generic**

40% **valuable**.

Experiment K - Kids Help Phone

- The **Social housing** case study confirms the findings.
- No ground-truth goal model known to GPT-4.



- **RQ2: How does GPT-4 perform in goal model generation from textual descriptions with different levels of detail?**

  GPT-4 exposes the modeler to **useful ideas** that may be non-obvious to stakeholders outside the domain.

Table of Contents

Conclusion

RQ1: Goal modeling knowledge?

Possess considerable knowledge.

RQ2: Goal model generation?

Generate useful ideas.

RQ3: Immediate interactive feedback?

Feedback can be helpful.

There is value.