Utrecht University

28-06-2026

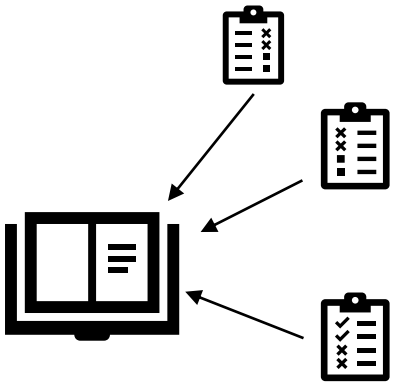
From concepts to learning objectives to questions: a pedagogically-aware prompting pipeline to generate questions from textbooks

Sergey Sosnovsky, Keyang He, Andre Bekema
Utrecht University, *Edu-Suite*

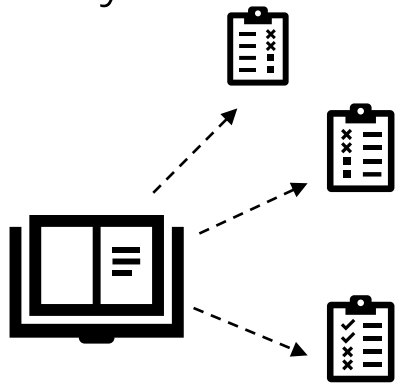
What is this paper about?

- An LLM was used to generate questions from a digital textbook
- 4 prompting strategies have been employed
- The main strategy is a 4-step chain of prompts trying to generate questions that match textbook pedagogy and semantics
- An expert-based evaluation has shown that the main strategy outperforms its counterparts

Motivation - 1: From Textbook Enrichment to Enrichment from Textbooks



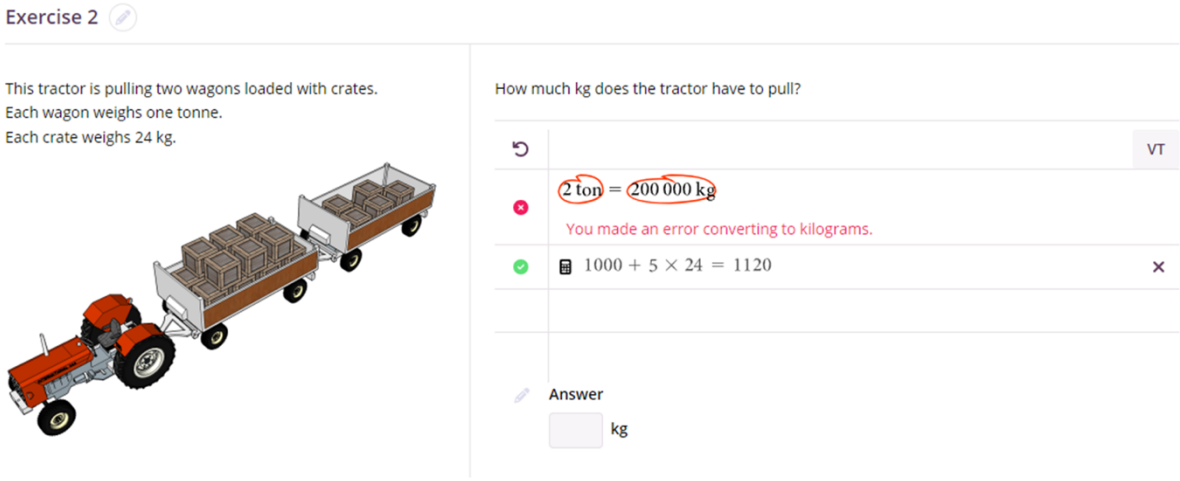
- Enrichment of digital textbooks with practice questions has many benefits
  - Natural didactic synergy
  - Evidence of learning and student modelling
  - Enables adaptive support and a wider range of pedagogies (e.g., mastery learning)



- Textbooks are a source of domain and pedagogical knowledge
  - A scalable approach towards textbook enrichment with questions should use the textbook itself to generate these questions
  - They can be generated in place
  - They can match the content, the style, the pedagogy of a textbook

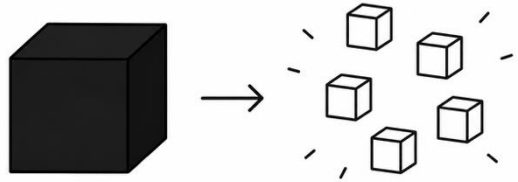
Motivation - 2: A Real Usecase

- Many publishing companies face exactly this challenge
  - How to enrich a digital textbook with in-place interactive content
- Edu-suite is a small educational publisher and an edtech service provider in the Netherlands
  - They have built a technology for converting textbooks to adaptive courses
  - Crafting exercises (and annotating them with KCs is the main bottleneck)
  - They have tried LLMs, but the quality of the generated questions is not reliably adequate



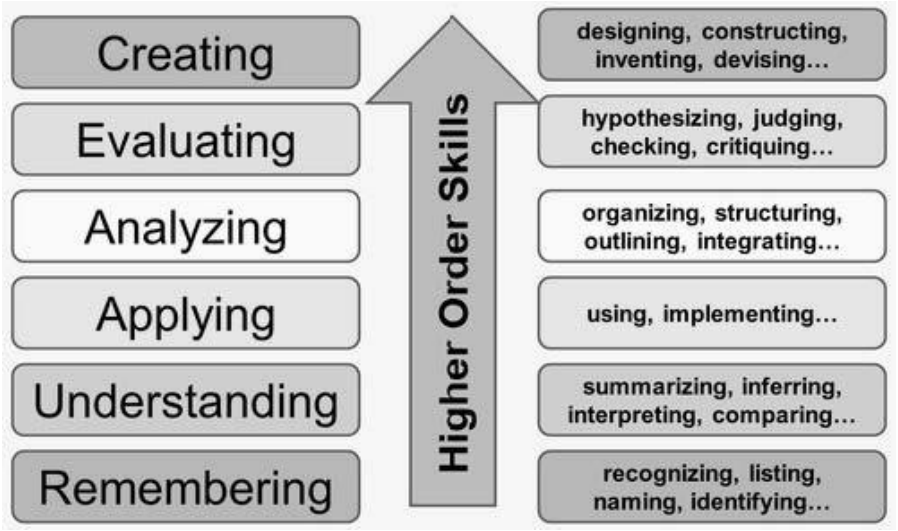
Motivation - 3: A More Reliable and Pedagogy-Driven Way to Employ LLMs

- Deconstructing the LLM blackbox
  - LLM's performance is more predictable/reliable on smaller tasks
  - A complex procedure can be broken down into a set of steps
  - Intermediate results are made explicit for quality control and symbolic inference



- Orienting LLM towards pedagogical objectives
  - Overreliance on LLM's ability to solve educational problems is dangerous
  - LLMs are optimized for next-token prediction, not learning outcomes
  - Prompts need to focus on scaling simpler text processing tasks
  - Didactic decisions need to be externalized

Proposed Question Generation pipeline

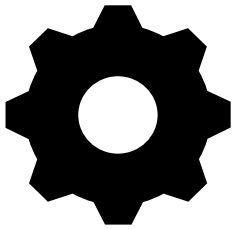


Textbook → Concept → Target Concepts → Learning Objectives → Questions

Experiment: Conditions

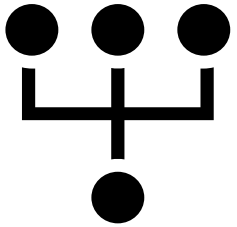
We have compared 4 prompting methods:

Monolithic Content-based prompt



VS.

Content+ Concept-based prompt pair

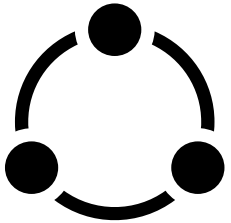


VS.

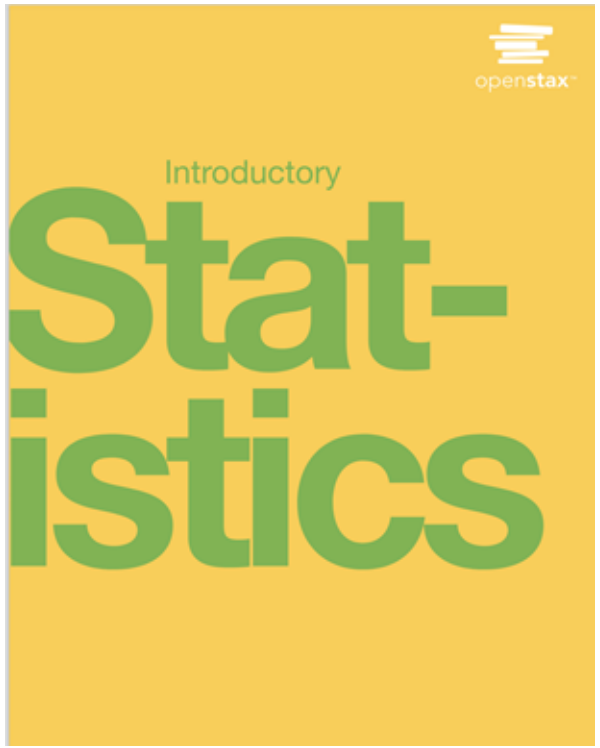
Target LO-based prompting pipeline

VS.

Chain-of-thought-based Super-prompt



Experiment: Content



- Single textbook on Statistics | Two sections
- For each section:
  - A set of questions are generated by each method
  - 40 questions are selected – 10 from each method

2.5 | Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common

13.2 | The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

NOTE

The F distribution is derived from the Student's t-distribution. The values of the F distribution are squares of the corresponding values of the t-distribution. One-Way ANOVA expands the t-test for comparing more than two groups. The scope of that derivation is beyond the level of this course. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise t-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

Experiment: Experts

13.2 | The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the ***F* distribution**, named after Sir Ronald Fisher, an English statistician. The *F* statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if *F* follows an *F* distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

NOTE

The *F* distribution is derived from the Student's t-distribution. The values of the *F* distribution are squares of the corresponding values of the *t*-distribution. One-Way ANOVA expands the *t*-test for comparing more than two groups. The scope of that derivation is beyond the level of this course. It is preferable to use ANOVA when there are more than two groups instead of performing pairwise *t*-tests because performing multiple tests introduces the likelihood of making a Type 1 error.

2.5 | Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common

Experiment: Instrument

Question Quality Questionnaire:

- 7 criteria on a 5-point Likert scale:
  - *Grammaticality, Relevance, Answerability, Interest, Importance, Utility, Sufficiency*
- *Difficulty* on 5-level scale (from much too easy to much too hard)
- *Cognitive level alignment* on a 4-category scale: Lower / Comparable / Higher / Too high

Experiment: Data pre-processing

- Overall data sample:
  6 ratings per question across 2 sections across 10 questions across 4 conditions = 480 ratings for each criterion
- Altogether, 440 individual ratings had to be removed due to abnormal rating patterns (straight-lining)
- We also estimated reliability of our quality criteria using a linear-mixed-model-based intra-class correlation (LMM-based ICC). We removed:
  - Grammaticality due to a ceiling effect (almost all answers “agree” or “strongly agree”)
  - Sufficiency due to a very low reliability

Results: Marginal Means of Ratings on Quality Criteria per Method

Criterion	*Content*	*Concepts*	*LOs*	*SuperPrompt*
Relevance	4.16 (0.89)	4.10 (0.93)	**4.44** (0.75)	4.14 (0.81)
Answerability	3.91 (1.31)	4.01 (1.23)	**4.63** (0.61)	4.14 (1.14)
Interestingness	3.47 (1.26)	3.12 (1.30)	**3.72** (1.12)	3.48 (1.15)
Importance	3.86 (1.03)	3.81 (1.00)	**4.21** (0.89)	3.88 (0.92)
Utility	3.68 (1.08)	3.57 (1.12)	**4.02** (1.06)	3.83 (1.04)

Across Criteria, the Ratings for  are higher than for other methods

Results: Marginal Means of Ratings on Quality Criteria per Method

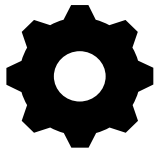
Criterion	*Content*	*Concepts*	*LOs*	*SuperPrompt*
Relevance	4.16 (0.89)	4.10 (0.93)	**4.44** (0.75)	4.14 (0.81)
Answerability	3.91 (1.31)	4.01 (1.23)	**4.63** (0.61)	4.14 (1.14)
Interestingness	3.47 (1.26)	3.12 (1.30)	**3.72** (1.12)	3.48 (1.15)
Importance	3.86 (1.03)	3.81 (1.00)	**4.21** (0.89)	3.88 (0.92)
Utility	3.68 (1.08)	3.57 (1.12)	**4.02** (1.06)	3.83 (1.04)

Across Criteria, the Ratings for  are higher than for other methods

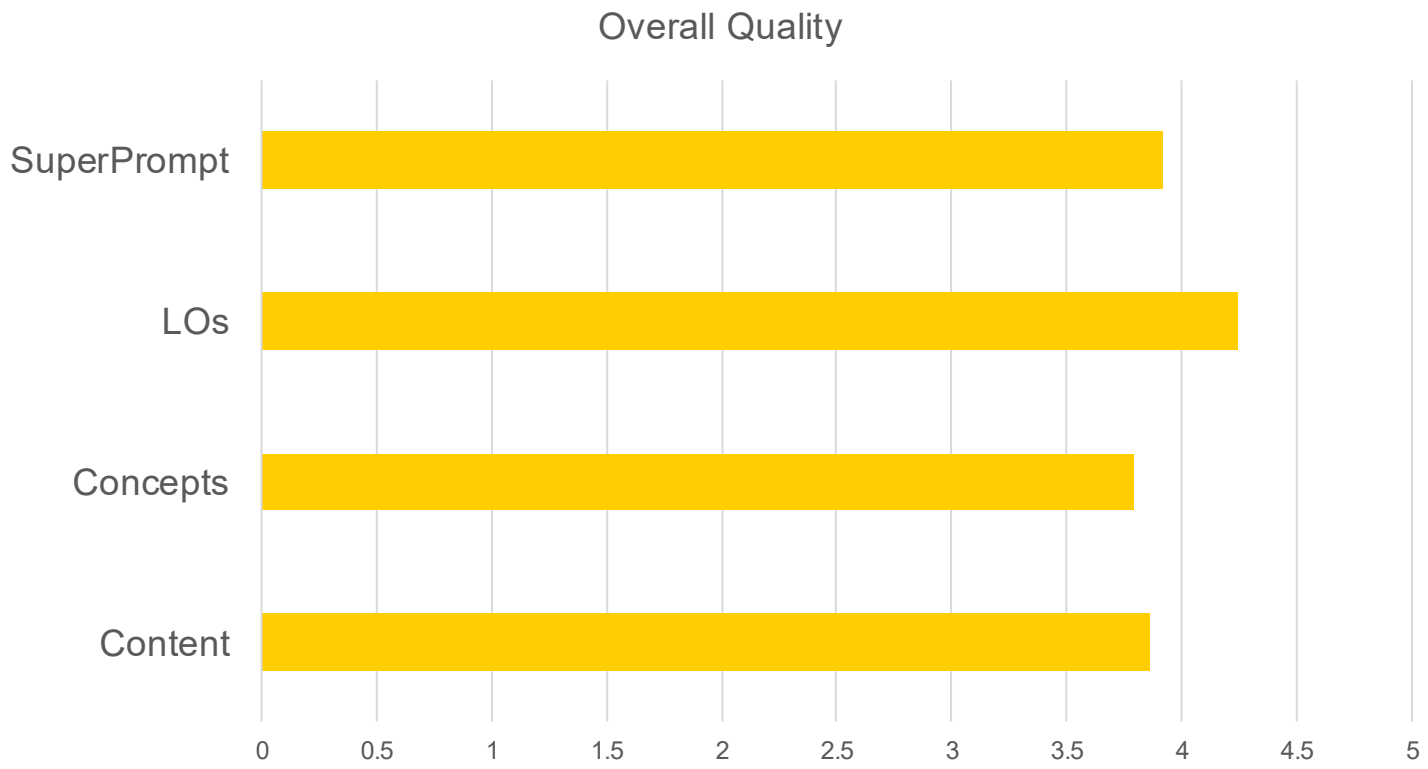
Results: Comparing the Target Method (LOs) with the Baseline (Monolithic Prompt)

Criterion	Diff	95% CI	p	g
Relevance	+0.28	[0.00, 0.56]	**.047**	+0.56
Answerability	+0.72	[0.24, 1.20]	**.003**	+1.16
Interestingness	+0.25	[-0.17, 0.56]	.248	+0.35
Importance	+0.35	[0.01, 0.75]	**.047**	+0.61
Utility	+0.34	[-0.14, 0.59]	.079	+0.53

On three criteria (Relevance, Answerability and Importance),

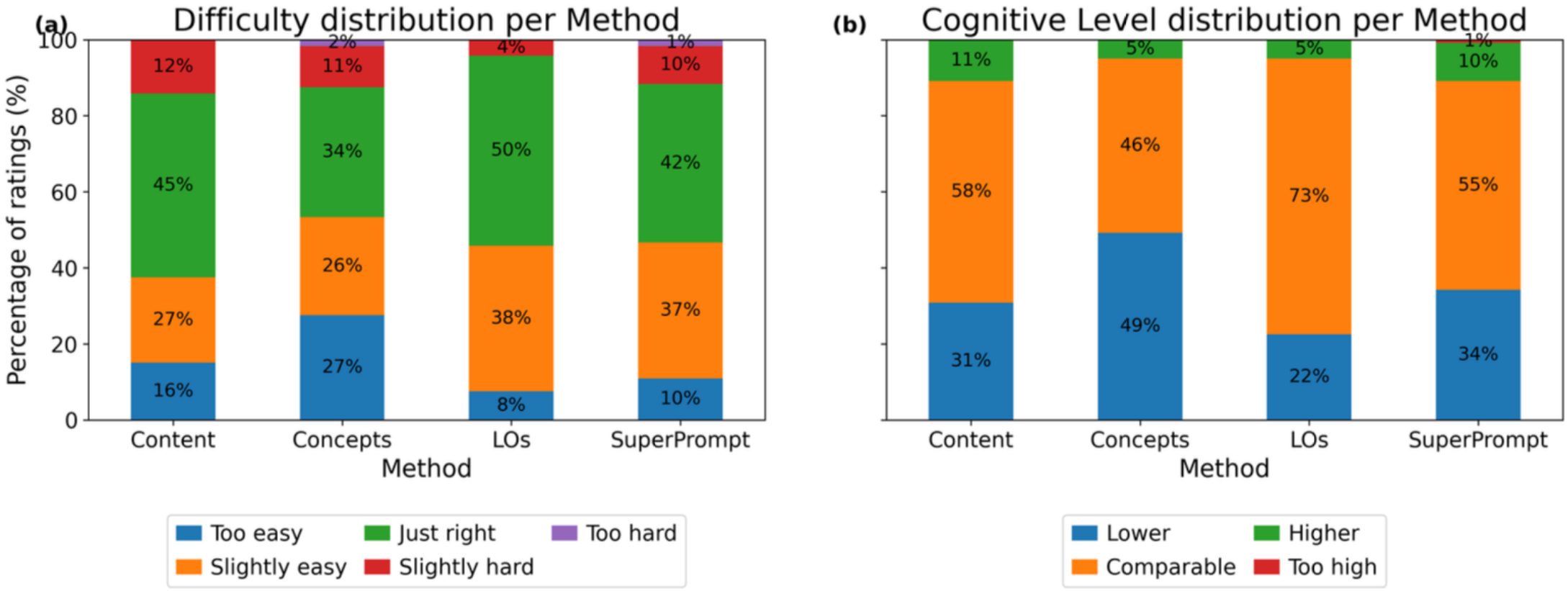
 significantly outperforms 

Results: Comparing Overall quality across methods



Based on a combined quality metric, 🎯 significantly outperforms the rest

Results: Difficulty and Cognitive Level



🎯 generates questions that are ranked most often as:

- Just Right on Difficulty
- Comparable on Cognitive Level

Utrecht University

Thank you for attending
Questions?