# Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems

Robin Schmucker, Meng Xia, Amos Azaria, Tom Mitchell

## PAPER OVERVIEW

New type of **automated** and **domain-independent** intelligent tutoring system

- Generates tutoring workflow *automatically* around existing content
- Enables *free-form* conversational tutoring in *learning-by-teaching* format
- Features two *LLM-based agents* with the roles of student and professor
- Uses *Expectation Misconception Tailoring* (EMT) to structure dialogs

## INTELLIGENT TUTORING SYSTEMS

Learning software that provides **personalized** and **adaptive** instruction

- **Content development** is **major obstacle** to widespread adoption

  -> Authoring 1 hour of content takes experts 10s to 100s of hours
- Workflow is structured into two loops
  - *Outer Loop*: Learning activity sequencing
  - *Inner Loop*: Feedback and assistance during activity

## SYSTEM EVALUATION

Online user study (N = 100) with 4 conditions:

- *Reading*: Receives lesson text (same in all conditions)
- *Teacher Q/A*: Simple chatbot with Human written questions/answers
  - Feedback: Correctness of user response + Sample solution
- *LLM Q/A*: Simple chatbot with GPT-4 generated questions/answers
- *Ruffle&Riley*: Full system (EMT + Learning-by-Teaching)

Multiple-choice Post-test

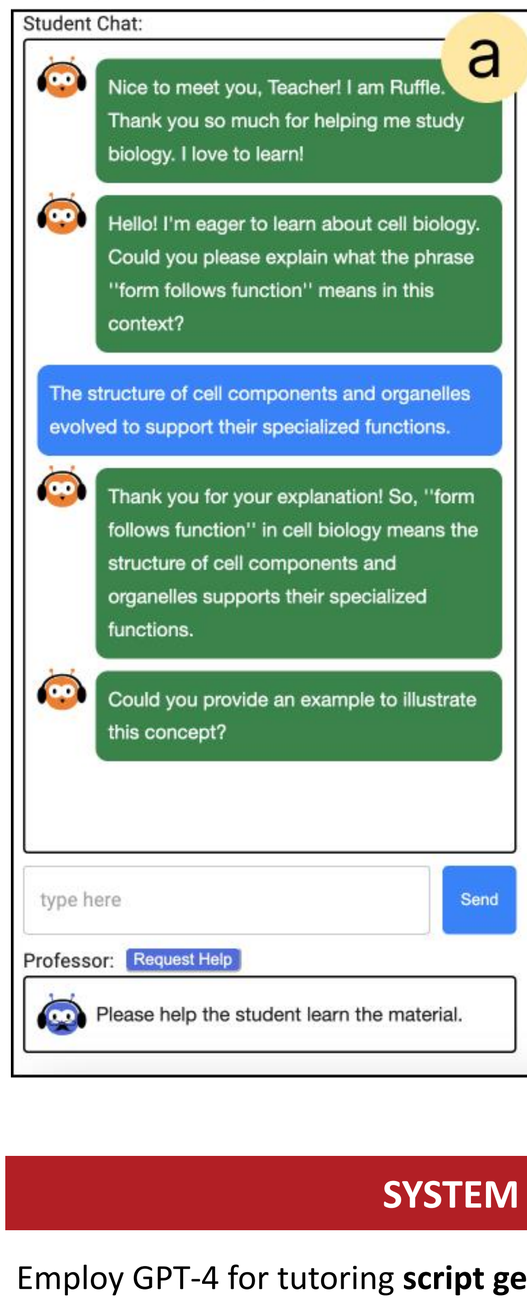
| Conditions | # of participants | | Previous Knowledge | Learning Performance |
|---|---|---|---|---|
| | Before filtering 100 | After filtering 58 | | Post-test Scores (i.e.,Multiple-Choice Questions) |
| Reading | 30 | 15 | 2.53 ± 0.41 | 5.07 ± 0.33 |
| Teacher Q/A | 17 | 7 | **3.0 ± 0.58** | 4.14 ± 0.83 |
| LLM Q/A | 23 | 15 | 2.2 ± 0.3 | 4.67 ± 0.35 |
| Ruffle & Riley | 30 | 21 | 2.67 ± 0.43 | **5.19 ± 0.25** |

Learning Experience Survey

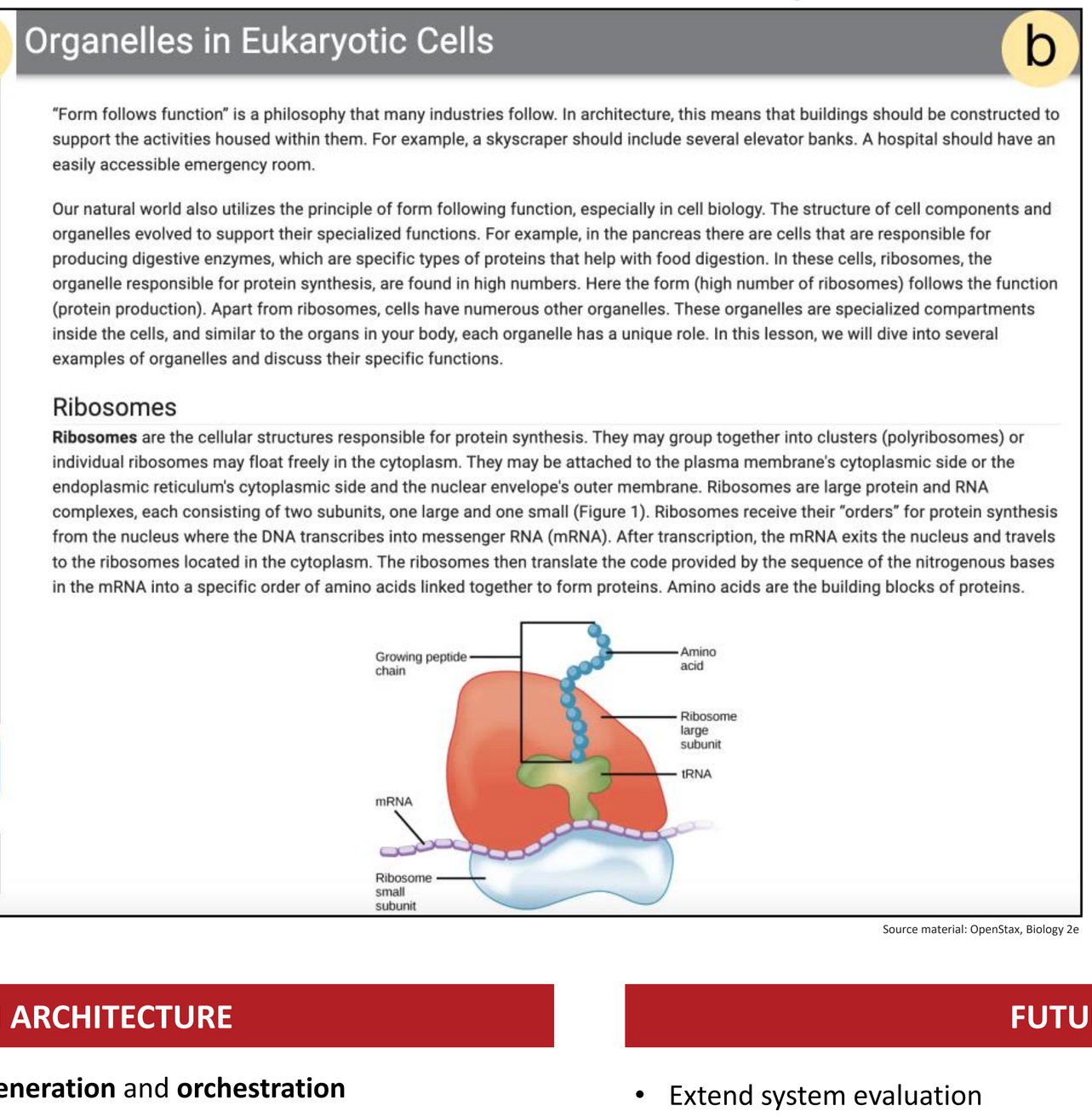
| Conditions | Learning Experience (1-strongly disagree, 7-strongly agree) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Engagement | Understanding | Remembering | Interruption | Coherence | Support | Enjoyment |
| Reading | 4.33 ± 0.52 | - | - | - | - | - | - |
| Teacher Q/A | 5.0 ± 0.53 | 4.43 ± 0.65 * | 4.43 ± 0.65 * | 2.71 ± 0.64 | 5.43 ± 0.53 | 4.57 ± 0.57 * | 3.71 ± 0.52 * |
| LLM Q/A | 4.8 ± 0.47 | 4.4 ± 0.4 * | 4.33 ± 0.42* | 2.67 ± 0.45 | 4.8 ± 0.43* | 4.0 ± 0.44* | 4.0 ± 0.44 * |
| Ruffle & Riley | **5.81 ± 0.3** | **5.81 ± 0.24** | **5.76 ± 0.22** | **2.19 ± 0.34** | **6.1 ± 0.21** | **5.9 ± 0.26** | **5.62 ± 0.31** |

## USER INTERFACE

### *Free-form Dialog*

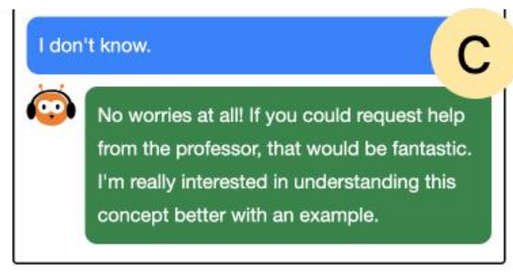



### *Pre-existing Textbook Content*

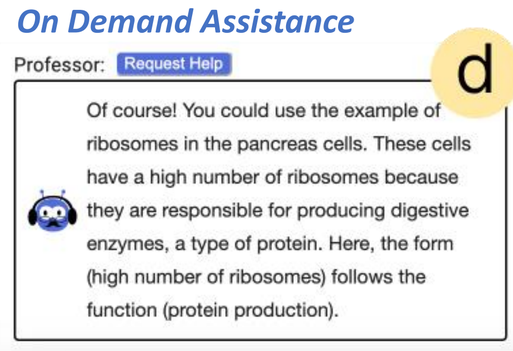



Source material: OpenStax, Biology 2e

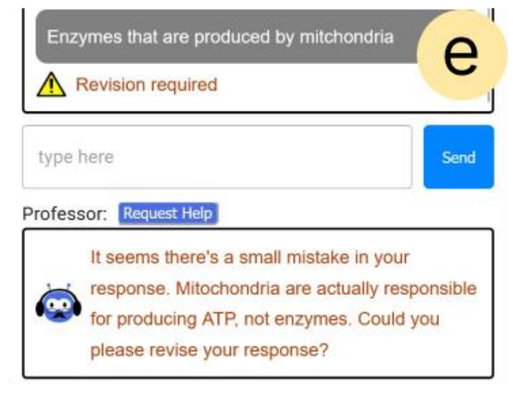
### *Intelligent Feedback*




### *On Demand Assistance*

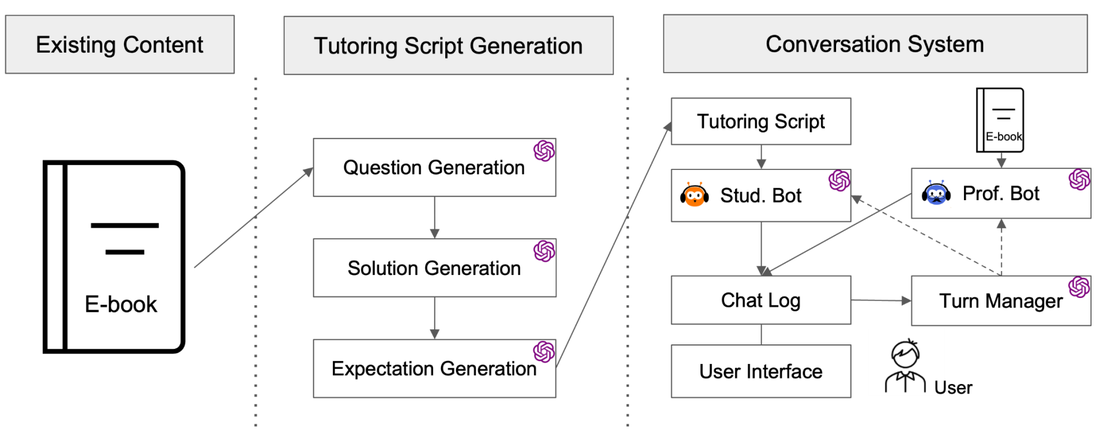



### *Misconception Correction*




## SYSTEM ARCHITECTURE

Employ GPT-4 for tutoring **script generation** and **orchestration**




**Learning-by-teaching workflow:**

*Student agent*: Asks user to teach the material; Guides conversation to cover all questions and expectations in the tutoring script

*Professor agent*: Provides feedback on user's teaching; Provides hints in response to help requests; Prompts user to correct incorrect explanations

## FUTURE WORK

- Extend system evaluation
  - Evaluate learning outcomes with deep understanding questions
  - Evaluate effects on knowledge retention
- Refine learning workflow
  - Chatbot users focused on the question content
  - Let users review content before teaching student agent
- Add human-in-the-loop capabilities (e.g., to revise tutoring scripts)
- Perform in-depth evaluation of conversation properties
  - How does explanation quality relate to post-test performance?
  - Are there questions or parts of the lesson text that cause confusion?

**Acknowledgements**: This material is based on work supported by the AFOSR under award FA95501710218 and by Microsoft as part of the Accelerate Foundation Model Academic Research Program.