Additional experimental details for "Imagine a dragon made of seaweed: How images enhance learning in Wikipedia"

Anonymized for Review

Dataset

Article Selection To identify 'less known' articles, we sub-selected pages that had under 6000 views per month (9th decile) and were less than 15k characters (removing any overly detailed articles). We treated this as a proxy for articles that were less likely to be familiar to an 'average' reader.

When creating both visual and general knowledge questions, we wanted to ensure that the article text provided sufficient information to answer them. For example, the article text would need to mention that Donatello's David is in Italy and Detroit-style pizza is traditionally rectangular. In piloting question generation, we found that most information for visual knowledge questions came from specific sub-sections of the articles. Specifically, these sections were: *Description, Characteristics, Architecture, Appearance, Design, Details, Morphology, Anatomy and physiology, Construction, Size(s), and Physiology*. We retained articles that had at least one of these sections. In the end, we generated a list of 31k articles. Note that these were simply used as a seed for our question annotators.

As articles could vary significantly in detail and length, we created excerpted versions. The excerpt included the lead (top) section of the article and any section that was used to generate general or visual knowledge questions. For example, the answer to "The Père David's deer is significant to which cultural mythology?" was found under the "Legend and cultural significance" subheading. This subsection was thus incorporated into the article's excerpt (for this specific article, the except also included the Characteristics subsection). At most an except would be comprised of 4 subsections (one for question).

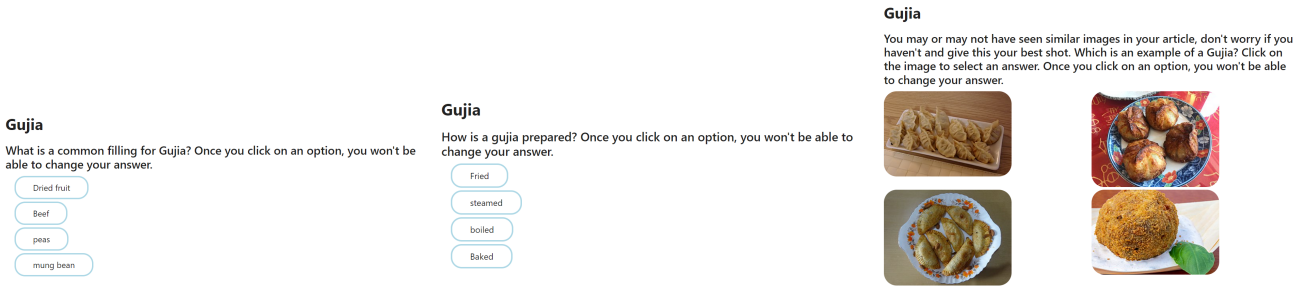
Question Generation Additionally, we developed specific guidelines for choosing general knowledge, visual, and image recognition questions. In the structure of a learning objective framework such as Bloom's (Anderson, Krathwohl, and Bloom 2001; Bloom et al. 1956)) the questions assess to test recall of specific facts (e.g., "the student will recall who commissioned Donatello's David"). Assessments at this level are more easily structured as multiple response questions (Haladyna 2004). Note that while these questions test *comprehension* (and specifically, recall), they were not designed around deeper learning tasks (e.g., summarizing, inference, etc.) or knowledge dimensions (e.g., conceptual, procedural, etc.).

Visual knowledge questions were created only using article text and did not rely on the article's images. We selected visual knowledge questions in such a way that there was *plausibly* an image that represented *the entity itself* which could be used to answer the question (e.g., a photograph or sketch of Donatello's David). Of course, this *need not be* the image that is actually used on the Wikipedia page. When crafted this way, visual knowledge questions tended to ask about visually salient features (e.g., color), visually comparable features (e.g., size in relation to similar organism), or other physical properties that could be determined with some prior knowledge (e.g, building material). Visual knowledge questions were also tagged with a binary value–consistent or inconsistent–indicating if the question could be answered by only looking at the image for the article (specifically the 'lead' or 'main' image in use on Wikipedia). For example, for the article about the Carolina Reaper pepper, the image has three examples of the pepper, but very few cues about scale. Thus, the question "When fully ripe, about how big is a Carolina Reaper pepper?" is marked as 'inconsistent.' However, the question "What texture is the Carolina Reaper Pepper?" is answerable (and thus consistent).

General knowledge questions were factual questions that did not have to do with visual properties of the subject and were not answerable by looking at images. These were intended to be simple, factual details, often with one-word answers (e.g., what part of the world or from what country is something? when was something built or created? etc.).

Main Experiment Design See Figures 1a, 1b, and 1c for examples of question format used in the study.

To create image recognition questions, our guideline suggested finding one picture of the entity that was not in use on the Wikipedia page and a number of distractors (all from the Commons). To find an appropriate target image, we suggested a number of candidates, including: lead images used in other language pages, images previously used on the Wikidata item page, or images in the same category in the Wikimedia Commons as the current image. For example, for the Turkish pastry Boyoz, all language editions utilized the same image as lead. However, the Commons cat-



(a) General knowledge question (b) Visual knowledge question (c) Image recognition question

Figure 1: Examples of (a) General knowledge questions, (b) Visual knowledge questions and (c) Image recognition questions from our experimental setting.

egory page for the image (https://commons.wikimedia.org/wiki/Category:Boyoz) provided six alternatives.

Distractors were generated by members of our research team. Following suggested guidelines (Haladyna 2004), the distractors for visual and general knowledge questions were in the same class as the true answer (e.g., other country names in similar areas of the world to the true answer). However, distractors were not intended to create trick questions or confusion. Distractor for the image recognition questions were found by looking at related or higher-level categories to the entity. For example, the page for Boyoz links to https://en.wikipedia.org/wiki/List_of_pastries. From here we see other stuffed pastry types (e.g., the Scottish Bridie), which have a related but distinct appearance.

Experimental Metrics

Participants' Characteristics For each participant, we collected:

- **Participant Id**, recorded anonymously.
- **English Proficiency**, taken from the demographics form, can take one of the following values: *beginner*, *basic*, *intermediate*, *upper-intermediate*, and *advanced*.
- **Education Level**, taken from the demographics form: *pre-high school*, *high school*, *college*, *graduate school*, *professional school*, *postdoctoral*, *PhD*.
- **Age** was taken from the demographics form.
- **Platform**, either *LabintheWild*, or *Prolific*. LabintheWild was used for participants that organically found and took the study through LabintheWild platform. While Prolific was used for participants that took the experiment for paid compensation using the Prolific platform.

Additionally, data for gender, country of origin, reasons for data exclusion (e.g., self-reported cheating), and indication of technical issues were recorded from the demographics form and the self-reported questionnaire at the end of the experiment. Devices and browsers used by the participants were also logged after participants agreement.

Article Properties Article metrics included:

- **Article Id**, a unique ID for each of the 94 articles
- **Article Time** taken by a participant to read an article[1].
- **Article Knowledge**, the answer to a five-nominal scale familiarity question regarding the topic of the article. The scale categories were: "I've never heard of it", 'I've heard of it but I don't know what it is", 'I have some knowledge about this topic", 'I know a lot about this topic", and 'I'm an expert on this topic".
- **Article Word Count** for each of the 94 articles.
- **Image Presence** a binary variable corresponding to the presence/absence of an image in the viewed article.
- **Readability** computed via the Flesch-Kincaid metric.

Question Properties We record the following data for questions in our experiment:

- **Question Type** identifies the type of question the participant saw: *general*, *visual*, or *image*.
- **Individual Question Score**, a binary variable that identifies whether the question was answered correctly or not.
- **Question Time**, the difference between timestamps of when the participant first saw the question and when they clicked a blue arrow to see the next slide.
- **Visual Knowledge Question Consistency**, an indicator (only for visual knowledge questions) of whether the question could be answered by looking at the image.

Experiment Design

At the end of the main and baseline experiment we asked participants to answer demographic questions about their level of education, country of origin, age, gender, native language, if they have done the study before, and English reading proficiency. The next page asked if they had technical issues or cheated at any point in the study to determine whether their data should be excluded. Finally, they saw a results page including a personalized "Wiki Knowledge Score" that took into account how long they took and how well they answered each question compared to others. An extra slide was used for Prolific participants to get their Prolific ID, and to display a completion code.

[1]before they clicked a blue arrow to see the next slide which included the first question for that article

Participant Demographics for Baseline Of the 47 usable participants, 47.82% self-identified as female, 50% as male, and 2.17% participants answered using other expressions (eg. non-binary, bigender, etc.). Ages ranged from 18 to 69 (M = 36.23, SD = 12.35). Two different countries of origin were reported by the participants: United Kingdom (91.5%), and United States (8.5%). A majority (91.11%) were English speakers, 4.44% Lithuanian, and 4.44% German speakers. The distribution of English proficiency was 86.95% advanced, 10.86% upper-intermediate, and 2.17% intermediate. The distribution of education levels was 21.73% high school, 26.08% college, 10.86% professional school, 34.78% graduate school, and 6.52% Ph.D.

Main Experiment Participants Among our 704 participants, 47.57% self-identified as female, 48.71% as male, and 3.70% participants answered using other expressions (e.g., non-binary, bigender, etc.). Participants were between 10 and 89 years old (M = 32.23, SD = 23.60). In total, 53 different countries of origin were reported by the participants. The top five were the United Kingdom (44.79%), United States (29.10%), Canada (4.27%), Germany (3.28%), Australia (2.13%), and France (1.42%). Thirty-one different native languages were recorded to be spoken among the participants, the majority being English (77.23%). The distribution of English proficiency was 82.91% advanced, 12.26% upper-intermediate, 3.06% intermediate, 1.02% beginner, and .07% basic. The distribution of education levels was 3.43% pre-high school, 15.04% high school, 35.81% college, 9.31% professional school, 27.07% graduate school, 7.16% PhD, and 2.14% postdoctoral.

References

Anderson, L.; Krathwohl, D.; and Bloom, B. 2001. *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives (Complete Edition)*. Longman.

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; and Krathwohl, D. R. 1956. *Taxonomy of educational objectives, handbook I: The cognitive domain*, volume 19. New York: David McKay Co Inc.

Haladyna, T. M. 2004. *Developing and validating multiple-choice test items*. Routledge.