Business Plan - ATLAS - Team Name: “The Highlights”

Title: *ATLAS - Active Tutored Learning Through Adaptive Systems*
Members: *Conrad, Jesse, Mohamed, Paul*

The purpose of ATLAS is to make learning content more fluent through content curation which leverages evidence-based learning strategies.

Our mission is to give all students the tools and knowledge to master any skill by teaching them how to be a better learner.

Initial Problem Statement

College poses a fundamental shift in the volume, intensity, and complexity of study material which overwhelms those individuals not equipped with effective study techniques. Nascent adaptive systems not only allow for a more effective storing and visualization of learned material but can also foster learning skills through algorithms that leverage evidence-based learning principles in the delivery and revision of learned content. While active learning and spaced repetition are ubiquitous in learning systems like Duolingo, Memrise, and Anki, there is a lack of systems that provide learners with customized feedback on their study habits, attempting at improving them.

Therefore, we aimed at building an adaptive tutoring system that reinforces effective study principles (e.g., active recall and spaced repetition), through providing an intuitive user interface, a flashcard and archiving system for studied content, and an insightful behavioral feedback dashboard to reflect on and improve one’s study habits.

User Interviews

We aimed at finding out whether college students are actually in need of such a product and conducted $N = 8$ user interviews with students from diverse geographical locations (US, UK, Germany) and subjects (e.g., CS, Psychology, Political Science, Engineering) in order to find out where students are frustrated about their study habits and what kind of apps would help them in their study routines. We also asked students about their willingness to pay for our product and additional features that they would require for such a financial commitment. Our full user interview guide can be found in Appendix A at the bottom of this document.

After conducting our user interviews we created affinity and emotion maps to compile repeating themes and emotions our interviewees experienced. From these themes, we developed three personas (i.e. archetypal potential customers) as well as features that our product would need to have in order to fulfill the needs and requirements that our interviewees shared with us. A Miro board, in which our design thinking process is documented, can be found here: https://miro.com/app/board/o9J_l_Y_4a8=/

Findings & Insights from User Interviews

Most notably, we found out that students are *not* in need of a tool that reinforces and provides feedback on effective studying principles (e.g., active recall and spaced repetition), but primarily struggle with two things:

First, students struggle with the generation of study questions and summaries. We especially observed this problem during our user interviews in students who attempt to learn complex, in-demand skills like data science, programming, and machine learning. Specifically, users reported that they were intimidated by the sheer amount of different resources and topics they could start learning to master these skills. Some of our interviewees reported that they were doubtful they could effectively learn complex skills such as programming because they lacked an intuition for an effective learning path. Our interviewees specifically wished for tools that would help them generate high-quality summaries and flashcards which they could revisit later.

Second, students struggle with the organization of content. Related to the problem of being overwhelmed by various options in terms of the content to start learning, students reported struggling with archiving past content they have learned in order to systematically revisit them in case they have forgotten about it or want to share material with friends. One user specifically wished to have a “global CTRL+F function” with which they could revisit summaries, flashcards, or notes that they created during learning. Some of our more advanced learners with more learning app experience also noted that they would highly appreciate an app that is able to cluster learned materials effectively. As an example, one student told us that “insight comes from unexpected links but these links are hard to find”. The student was pointing at notes and flashcards that automatically are linked with each other through algorithms. More importantly, students that already used flashcard apps often noted that their flashcards archives were difficult to search based on keywords.

Revised Problem Statement

As a consequence, we formulated a new problem statement, including two key user needs we compiled from our user interviews, that we want to tackle with ATLAS:

How can we build an application which a) effectively generates quizzes from text and b) allows for an effective storing and clustering of past learned material?

Market Size and Relevance of the Problem

Since we primarily interviewed college students, we did research about applications and tools that are similar to ATLAS and about which types of learners use them in order to identify potential market segments which could adopt our product. We found that the problem of effectively navigating digital learning content is highly relevant to large populations of learners both online and at colleges. This is because learning online and in college requires both effective study techniques to navigate content, study apps that foster long-term retention, and self-regulatory skills to continue learning despite inner resistance. We identified three key markets for our product:

First, college students (estimated more than 100 million worldwide; 19.7 million in the U.S. in 2020 according to the NCES) make up a relevant market for our product, specifically since less advanced students at the undergraduate level need to build up their study skills in order to successfully complete their degrees. With the ongoing pandemic and teaching continuing to be delivered remotely, self-study skills and effective learning tools are especially relevant and desirable for this population.

Second, users enrolled in Massive Open Online Courses (MOOCs; estimated to have around 180 million enrolled students worldwide in 2020), can also profit from tools that ease learning and foster long-term retention of study content. This is especially the case since MOOCs typically do not include content revision systems or provide readily available flashcards to users.

Third, independent learners on the internet who consume educational blogs like Medium are also in need of effective study tools to process, archive, and revise learned content. With platforms such as Medium reaching around 100 million monthly readers in 2020, this market segment is increasing in relevance.

Competition

In order to define product features for ATLAS, we thoroughly searched for existing alternative applications in order to pinpoint their limitations and define our competitive advantage.

We identified three different types of apps that try to help students leverage evidence-based study principles to become effective, independent learners:

First, we identified apps which store highlights and articles on the web and in the context of e-reading. Examples of such applications include Readwise, Pocket, and the annotation tool of Amazon Kindle. These tools are generally lacking an automated and systematic way of semantically clustering content and allowing for timely revision of content for long-term retention. At the same time, they do not transform the learned material into meaningful study material (e.g., practice questions).

Second, we identified apps that might be summarized as “flashcard apps and adaptive content revision systems”. These apps leverage active recall and spaced repetition to foster long-term retention in learners. Examples of these apps include Anki, Quizlet, Brainscape, Repetico, and GoConqr. The major drawback of these apps is that the user either has to rely on their own generation of flashcards or has to stick to community-shared flashcards. In that way, these apps lack the ability to effectively import and process external content automatically. In addition, these apps typically suffer from a steep learning curve, since the user has to search for or generate dedicated flashcards for these apps.

Third, we identified software that has attempted to automatically create flashcards from text input (e.g., PDF files, webpages, epub files), such as the “autocards” project by Psionica and web scraping tools which automatically import dictionary entries from the web (e.g., the Cambridge Dictionary) to flashcard apps like Anki. These apps lack a user-friendly design or even require the user to run software code themselves in order to work. Importantly, they do not provide an annotation tool that easily imports text from the web during learning.

In summary, there is no software available for students which combines the automatic creation of flashcards with timely revision algorithms and a user-friendly web annotation and document annotation tool.

Our Solution

As a consequence, the apps and tools currently available for students do not sufficiently address the aforementioned problem statement and user needs which we compiled from our user interviews:

How can we build an application which a) effectively generates quizzes from text and b) allows for an effective storing and clustering of past learned material?

Therefore, we came up with the following solution, including *two key features*, which we also compiled from our user interviews:

We create an application that can a) automatically generate study questions from highlighted text on the web or in PDF documents and b) archives and clusters past highlights and study questions for effective content revision and retrieval.

Our first feature includes a web annotation and document annotation tool which leverages natural language processing to automatically generate flashcards. In particular, we developed a Chrome extension that allows users to annotate any text on the browser (e.g., HTML documents and PDF files), edit and review highlights, as well as generate and export flashcards based on those highlights. In particular, the backend uses a T-5 transformers-based model (Enrico Lopez et al., 2020) to transform text into questions. The model uses four main stages that are explained more thoroughly in Appendix B at the bottom of this document. A clean frontend user interface allows users to edit their highlights prior to importing them into the model. The interface has clear indicators on how to save and delete highlights, download highlights, and finally generate content curation from the downloaded highlights. There is ample literature on the utility of engaging with content questions for effective learning. While generating and answering study questions has been shown to improve learning in postsecondary education (Luxton-Reilly et al., 2012; Shakurnia et al., 2018), the generation of high-quality questions does not come naturally to learners. This is due to the fact that creating such questions requires elaboration and recall. The automatic creation of study questions can help students apply effective study techniques more easily. With recent advancements in natural language processing, there have been systematic reviews of the literature on automatic question generation adapting to identify evidence for the efficacy of automatic questions generation for learners. As an example, an early review by Le et al. (2014) summarizes literature that showcases how offering high-quality study questions to college students, specifically when new to an area of study,

improves elaboration of content, self-explanation, and the identification of knowledge gaps. Our first feature, the automatic generation of study questions, is also the foundation for our MVP as shown in our pitch deck. A video walkthrough of our MVP can be found here: https://drive.google.com/file/d/1IdO2CNDDJ8i9Z7ZgKTb9jQYdhf2EVm6v/view?usp=sharing The code to our MVP can be found here: https://github.com/MoGaber/the-highlights-app

Our second feature includes an archiving, clustering, and content revision tool based on active recall and spaced repetition to help students organize learned content and foster long-term retention. We achieve this through multiple algorithms and tools, which each user has access to through dedicated user accounts connected to a database. First, long-term retention and revision recommendations are based on spaced repetition algorithms which periodically prompt the user to answer past study questions. Such content revision strategies have been shown to be the most effective study strategies based on the educational psychology literature (for an illustrative introduction, see for example Augustin, 2014). Second, clustering is achieved through user-inputted keywords associated with study questions and the content source (e.g., website URL, website domain). As an example, a user might tag the domain “towardsdatascience.com” as “data science” and “machine learning”. In addition, links between documents and content sources are automatically linked through keywords extracted through the unsupervised YAKE! and TF-IDF algorithms (Campos et al., 2020). Users are able to search for past highlights based on tags, semantic links, and relevant keywords extracted from documents through T-5 Transformers-based models, which are also connected to flashcards and documents. Finally, content revision and clustering are further optimized through the use of different highlight colors to indicate the difficulty of the material (e.g., users highlight text in red for content that requires further elaboration and green for content that is already understood well). Notably, a recent systematic review on automatic study question generation by Kurdi et al. (2020) pinpointed “generating questions with controlled difficulty” as an important area of future advancement in the field, such that ATLAS would be the first app that attempts at creating and clustering study questions of varying difficulty based on user input (i.e. different highlighting colors).

In short, automatic question generation and content clustering are ATLAS’ competitive advantage and make ATLAS favorable to customers compared to existing product solutions. These features are both backed up by empirical research as well as our user interviews. Specifically, ATLAS is the first product to successfully bridge the gap between a) evidence-based studying principles (active recall and spaced repetition) embedded in adaptive content revision systems, b) sophisticated natural language processing algorithms

for automatic question generation and content clustering, and c) powerful document and web annotation tools, which archive and annotate content across the WWW.

Business Model

We are a SaaS business operating with a Freemium Model. That means that our Basic plan is always going to be free but it comes with certain limitations. We are planning to experiment with the following types of limitations to find the most attractive combination for our users.

- Limited features: Our free plan allows learners to generate revision questions but only based on PDF upload and only paid users can use our Chrome extension.
- Limited usage: Only paid users can highlight more than 10 documents through the Chrome extension and can generate more than 10 revision questions.
- Limited capacity: We are restricting the amount of data learners can store on our platform for free users.

We did extensive research in our user interviews, and by looking at the pricing models of competing platforms, and we learned that a suitable price is in the range of USD 10 to USD 20 per month. Therefore, we will price our paid subscription at USD 15 per month. This gives us the flexibility to offer discounts to early adopters, and potentially increase the price after we achieved initial traction.

Based on our market research there is a Serviceable Obtainable Market (SOM) of 4.5M users. On average similar SaaS products are able to convert 10% of their users to paid users. This would give us a revenue potential of 6.75M MRR. In addition, we are planning to explore additional revenue streams such as partnerships with textbook providers, online course creators, and established education institutions.

Go-to-market Strategies

Our go-to-market strategy consists of four stages. However, we are constantly adopting, and therefore this strategy can change frequently.

1. We are targeting individual users that completed computer science and data science MOOCs. We are targeting them through forums such as the CS50 forum, Stackoverflow, Reddit, or other online communities. We offer them the free version

and a 50% lifetime discount for our paid version. We are also planning to work with influencers and content creators in the space to reach this audience.

2. We extend to all individual users that are outside of these online communities and have not necessarily completed a MOOC yet. We do this by creating a referral scheme and scouting program with our early adopters where they get rewarded for each user they refer. This allows us to create a referral flywheel and to grow organically. This stage can potentially be supported through paid marketing.

3. Once we have received initial traction, we can show credibility through social proof, and we can partner with educational institutions. Since we can already prove that their students are using our product, we have a high chance to close these partnerships. In particular, we are looking at MOOC providers such as Coursera or Udemy, Cohort-based courses platforms such as Maven or Coleap, or traditional educational institutions such as universities and high schools.

4. As a next step, we are planning to sell our product to educational content creators such as textbook providers or bigger technology companies that can use it as a supplement for their products. An example would be Amazon Kindle. This allows us to have larger deals, and to reach more users easily. Finally, this would offer certain exit opportunities for our business since we could be a suitable acquisition target for these companies.

Risk Assessment

There are two main risks in our business. Firstly, there is a technological risk. Our product builds on cutting-edge innovation in the fields of learning sciences and natural language processing. While the research results are promising, there is a risk that the technology and the underlying research are difficult to apply to a product at scale that has to have certain quality standards and user experience. However, we have a very strong team with deep technical knowledge that is perfectly suited to bring cutting-edge research to production.

Secondly, there is the risk of not reaching enough users, and not converting them to paying customers. While we have unique features that sets us apart from the competition, the SaaS independent learning tool market is very crowded. Therefore, it can be difficult to attract users and to minimize churn. In addition, many users try out different platforms but are ultimately not paying for them. However, we have a very strong go-to-market strategy and

growth hacking expertise in our team and we believe that our product is unique and exciting enough to convince the free users of its value.

Team

The ATLAS team brings very complementary skill sets and is in a perfect position to create a product that helps students to become better learners.

Paul brings extensive start-up experience and deep knowledge in product and business development. He co-founded a start-up that provided eight-week virtual learning experiences for hundreds of students and he was Head of Product for the venture-backed start-up ed-tech coleap.com.

Conrad has a Psychology and Data Science background and will start his MSc in Social Data Science at the Oxford Internet Institute this fall. In addition, he contributes a strong understanding of the empirical learning sciences and scientific research. He is responsible for making sure that our product is pedagogically sound.

Jesse studied Computer Science at Stanford and brings strong software engineering skills. He is our lead developer and is responsible for integrating the learning science and algorithms into the product.

Mohamed brings experience in backend development and data science. He is the wizard behind our Natural Language Processing algorithms.

Together we bring all the necessary skills to create a world-class learning product.

References

Augustin, M. (2014). How to learn effectively in medical school: test yourself, learn actively, and repeat in intervals. The Yale journal of biology and medicine, 87(2), 207.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. Information Sciences, 509, 257-289.

Enrico Lopez, L., Cruz, D. K., Blaise Cruz, J. C., & Cheng, C. (2020). Transformer-based End-to-End Question Generation. arXiv e-prints, arXiv-2005.01107.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30(1), 121-204.

Le, N. T., Kojiri, T., & Pinkwart, N. (2014). Automatic question generation for educational applications–the state of art. In Advanced computational methods for knowledge engineering (pp. 325-338). Springer, Cham.

Luxton-Reilly, A., Bertinshaw, D., Denny, P., Plimmer, B., & Sheehan, R. (2012, February). The impact of question generation activities on performance. In Proceedings of the 43rd ACM technical symposium on Computer Science Education (pp. 391-396).

Shakurnia, A., Aslami, M., & Bijanzadeh, M. (2018). The effect of question generation activity on students' learning and perception. Journal of Advances in Medical Education & Professionalism, 6(2), 70.

Appendix A: User Interview Guide

Consent

To record this interview, I need your explicit consent that will also be recorded. Is that okay for you?

- Could you state your full name for the recording?
- Do you give me permission to interview you?
- Do you give me permission to record this interview?
- Do you give me permission to quote you directly using a pseudonym?
- Are you happy to take part?

Background

- Where do you go to school and what is your class year?
- What are you studying? Major/minor/master's?
- How many Computer Science courses have you taken? [Statistics??] Which courses have you taken specifically?
- What are your post-college plans? What job/career do you plan to pursue?

Project Concept

- Our vision = chrome extension and/or web application
- Highlight text on the web or in a pdf (e.g., lecture slides) -> obtain keywords, definitions, and flashcard questions automatically
    - Content curation and generation based on Wikipedia articles
- Collect these condensed concepts and questions over time, sort by topic, and quiz yourself on the contents regularly
- Having a personal account: be able to view past highlights and make edits when necessary

Questions

Each section should have a combination of open and closed questions. Open questions could be followed up by further questions based on that person's response. Potential follow-up questions should also be noted.

Preparing for Homework, Quizzes/Exams, Projects

- How do you typically go about processing and learning new information while studying?
    - Which learning strategies do you know of and use during studying?
    - Which apps might you already use?
    - What app or functionality might help you in facilitating the way you learn?
- What have you found hard about completing the homework assignments?
- What have you found hard about performing on quizzes and/or exams?
- What have you found hard about completing the projects?
- Do you consider yourself a good student and why?

Product

- Would you be interested in taking advantage of a tool as we have described? What would you most likely use?
    - Would you see a use case through simply just learning the material?
- What would you improve about it?
- What would you not use?
- How much would you be willing to pay for a service such as this (think monthly rate of a premium account)?
    - In your opinion, is this something your university/college should invest in?
    - Is it more valuable than a Netflix ($9 USD/month in US)?

Final Thoughts

- Do you see any use cases for our product beyond what we've talked about?
- Any thoughts or suggestions you would like to end with?

Appendix B: The Four Stages of the ATLAS Model

Keywords Extraction

The first step in the pipeline is keyword extraction. Out of all the provided highlights, we need to extract the main keywords for which the model would then generate concepts. We implemented keyword extraction using Yake and TF-IDF. The algorithm simply compares the frequency of every word in the provided highlights to the frequency of that word in a larger corpus. If the word has a high frequency in the highlights as compared to the training corpus, the word will be considered a keyword. We used the publicly available Wikipedia corpus to train the model: https://zenodo.org/record/3631674#.YQmvto4zZPY

Keywords Tailoring

Although the TF-IDF model is very powerful at extracting keywords, it still has a few limitations. Some of the keywords that get extracted are not very relevant. For example, look at an excerpt taken from the model:

```
'regression attempts',
'machine',
'Machine Learning',
'variable',
```

The model indicated that the phrase “regression attempts” is a keyword. This phrase has indeed been referenced several times in the highlighted text but it’s not a meaningful keyword.

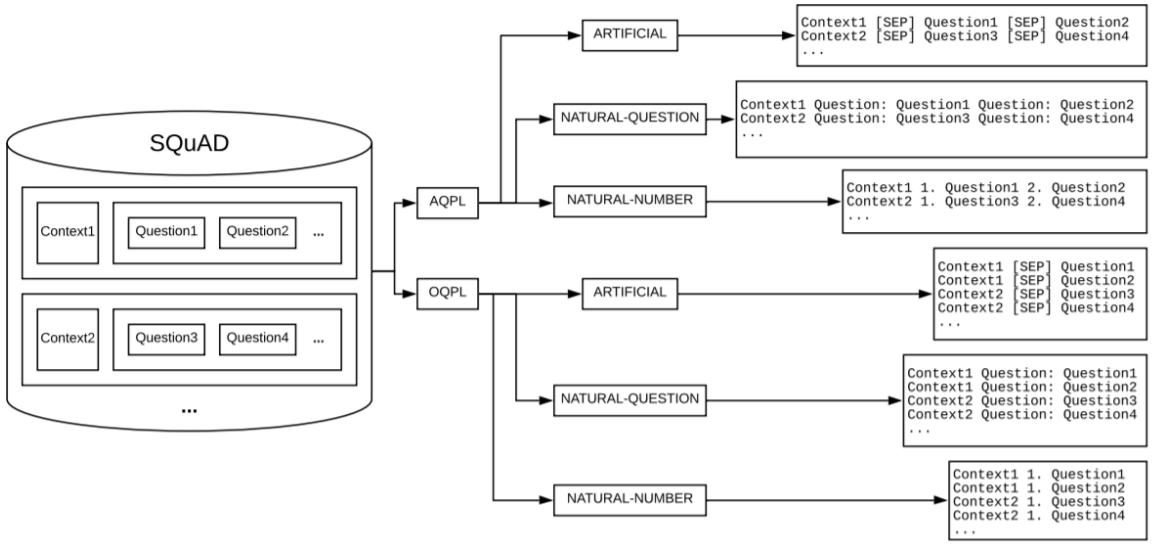
To overcome this limitation, we had included a filtering layer that would include only the meaningful keywords. We used the module ‘Wikipedia’ in Python that directly loads concepts from Wikipedia articles. We then passed each extracted keyword from the model to Wikipedia ‘Titles” and we took only the keywords that had a suitable match with the titles.

Concepts Generation

After generating the final list of keywords, we generate a list of definitions for each keyword. This is done in two different approaches. The first approach uses Wikipedia scraping. In that approach, we search the keyword in Wikipedia Corpus and we extract the definition text provided. This works well most of the time since the keywords are already filtered by Wikipedia before getting selected. However, sometimes Wikipedia has several definitions for the same concept. In that case, we use Google Developers “Machine Learning Glossary” to scrape definitions: https://developers.google.com/machine-learning/glossary. We plan to expand this approach to be applicable for any topic and not just machine learning.

Quiz Questions Generation

The final step in the pipeline is to generate quiz questions. The most common approach of generating questions using ML models is Answers-Aware questions generation. This approach works by feeding the model with answers and the text, then the model is supposed to move through the text and tokenize the phrases before and after the answer to turn them into questions. We found that this approach will not be applicable in our case since we don't have the answers available. We then decided to use an End-to-End question generation model. The model is discussed in this paper by Enrico Lopez et al. (2020): https://arxiv.org/pdf/2005.01107v1.pdf. The model is a T-5 Based Transformers model. The mechanism is summarized in the following image:



Briefly, the model is fed with a context in the form of a text. The text gets broken into sentences separated by tokens just like the usual BERT embedding process. The model then starts by extracting keywords that would indicate the beginning of an answer (e.g., the word 'by' can be turned into a question of the format 'by who' or 'by what'). The model would then attempt a few different training mechanisms (AQPL and OQPL) which basically assigns either one or multiple questions per line to the provided context.

We used a pre-trained version of the model provided by the NLP company 'Hugging Face' available here: https://github.com/patil-suraj/question_generation.git