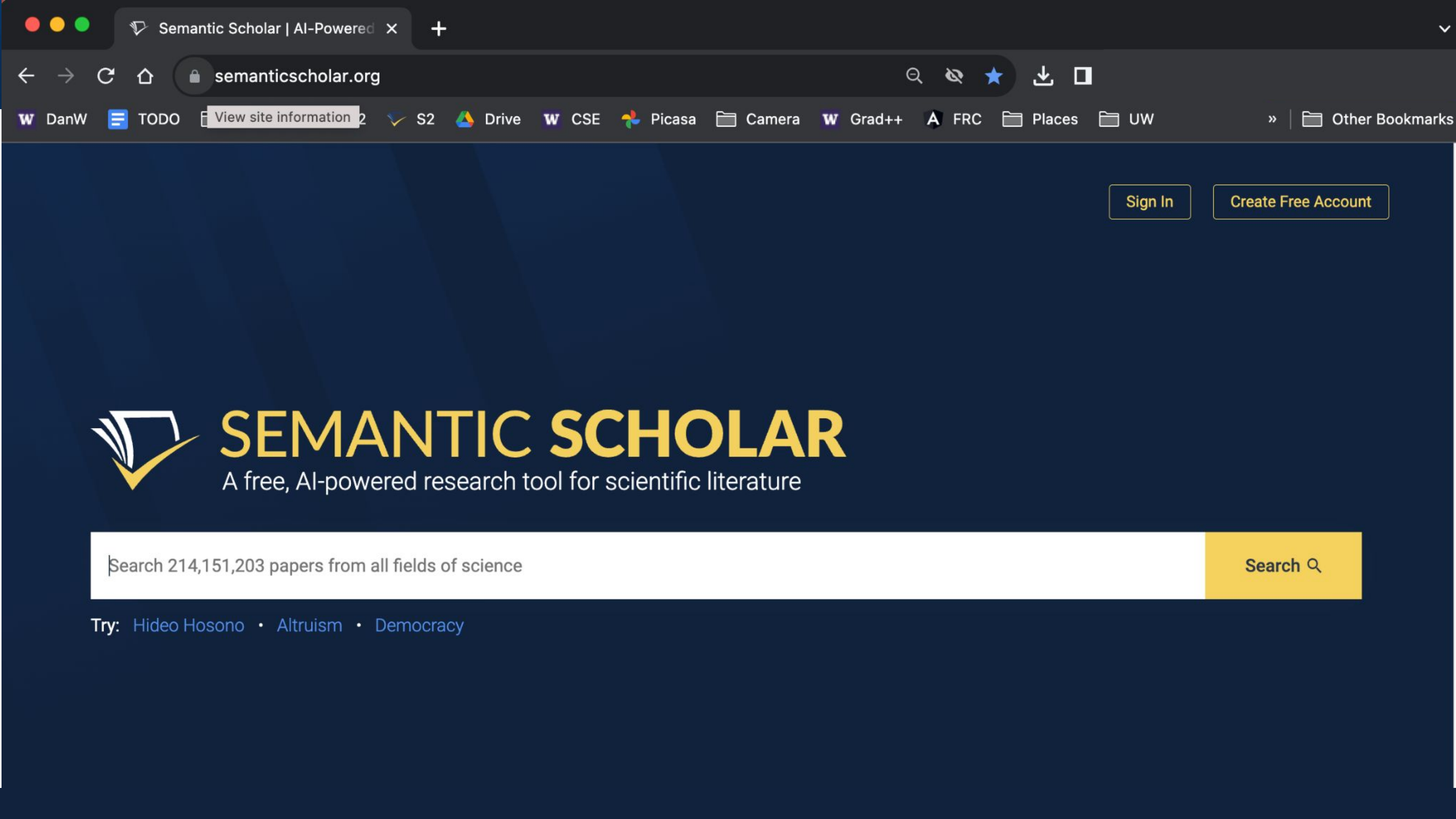
AI for Scientific Assessment

Daniel S. Weld

Allen Institute for Artificial Intelligence
University of Washington

Ai2

September 2024


Semantic Scholar | AI-Powered
semanticscholar.org
DanW
TODO
View site information
S2
Drive
CSE
Picasa
Camera
Grad++
FRC
Places
UW
Other Bookmarks
Sign In
Create Free Account
SEMANTIC SCHOLAR
A free, AI-powered research tool for scientific literature
Search 214,151,203 papers from all fields of science
Search
Try: Hideo Hosono • Altruism • Democracy

Semantic Scholar Mission:

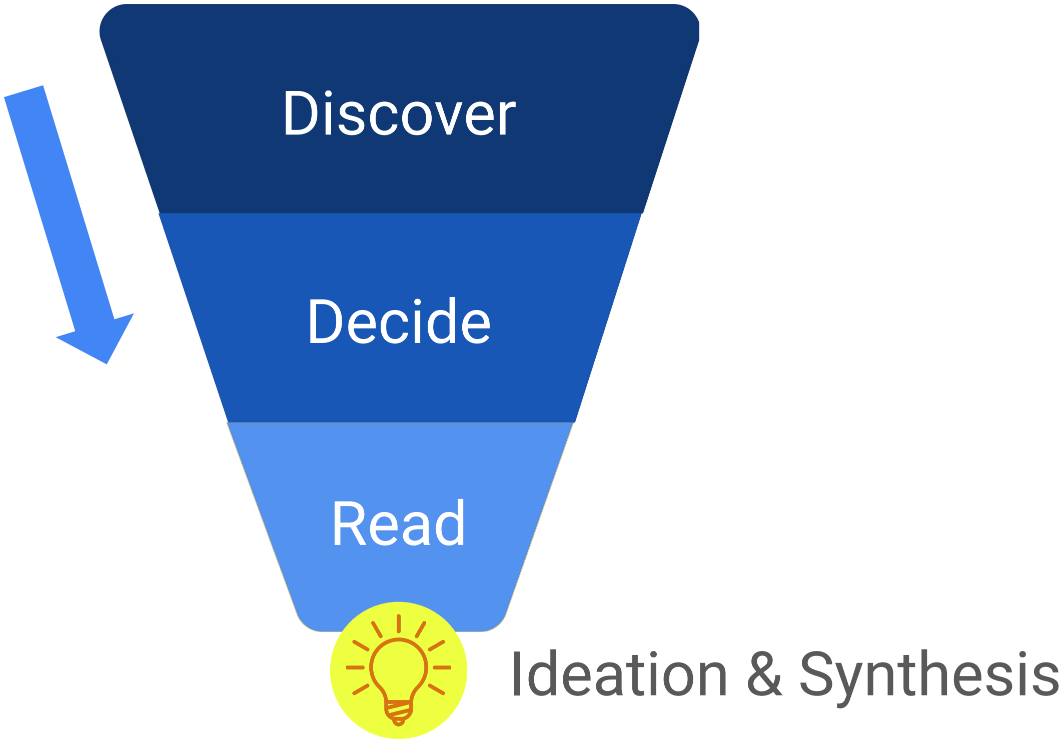
Accelerate Science Breakthroughs Using AI

Ai2 is a non-profit & Semantic Scholar is a free service

8M	monthly active users
221M	scientific paper index
30+	data sources
25+	ML models

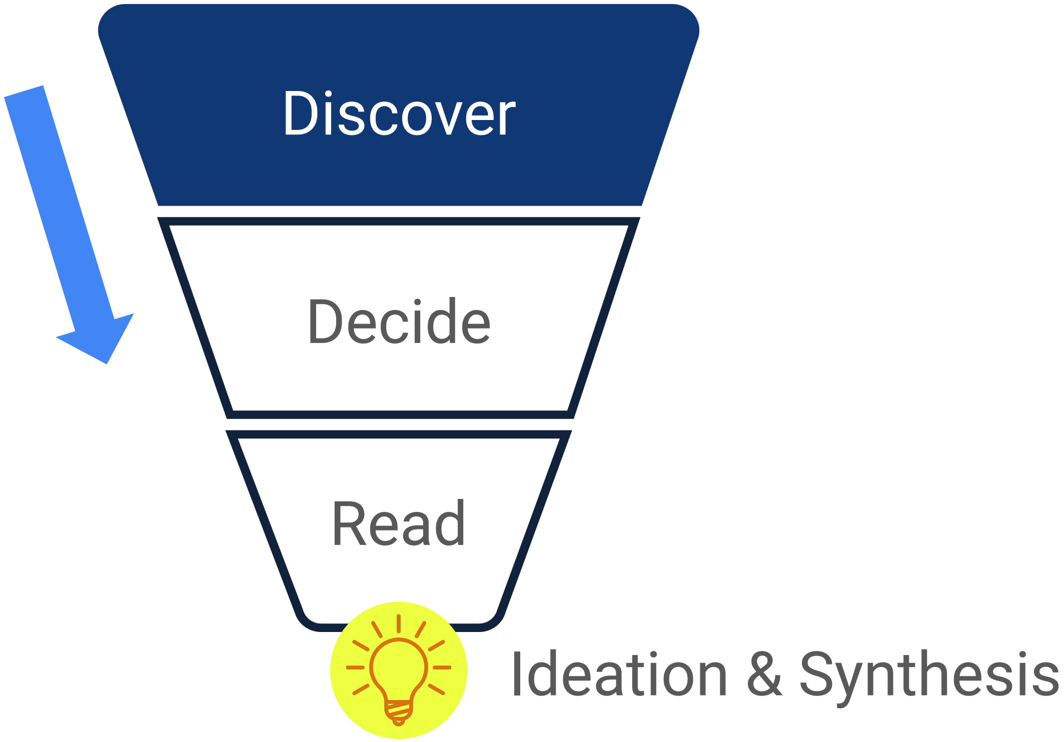
Ai2

The Scholarly Funnel



Ai2

Discover!



Ai2

Discovery - Our Research

	Traditional	Via Papers		Via People
Discovering Papers	Socially	**Research Feed**	**Synergi** UIST'23	**WhoYouKnow** CHI'22
	Searching	**Threddy** UIST'22	**CiteSee** CHI'23	**LitTweeTure** CHI'24
	Browsing	**Relatedly** CHI'23	**CitanceRetrieval** Ongoing	
	Check References & Citations	**Topics** Ongoing	**PaperWeaver** CHI'24	
Discovering People	Socially	**FeedLens** UIST'22		**S2AMP** JCDL'22
	Searching	**Bridger** CHI'22		**ComLitttee** CHI'23

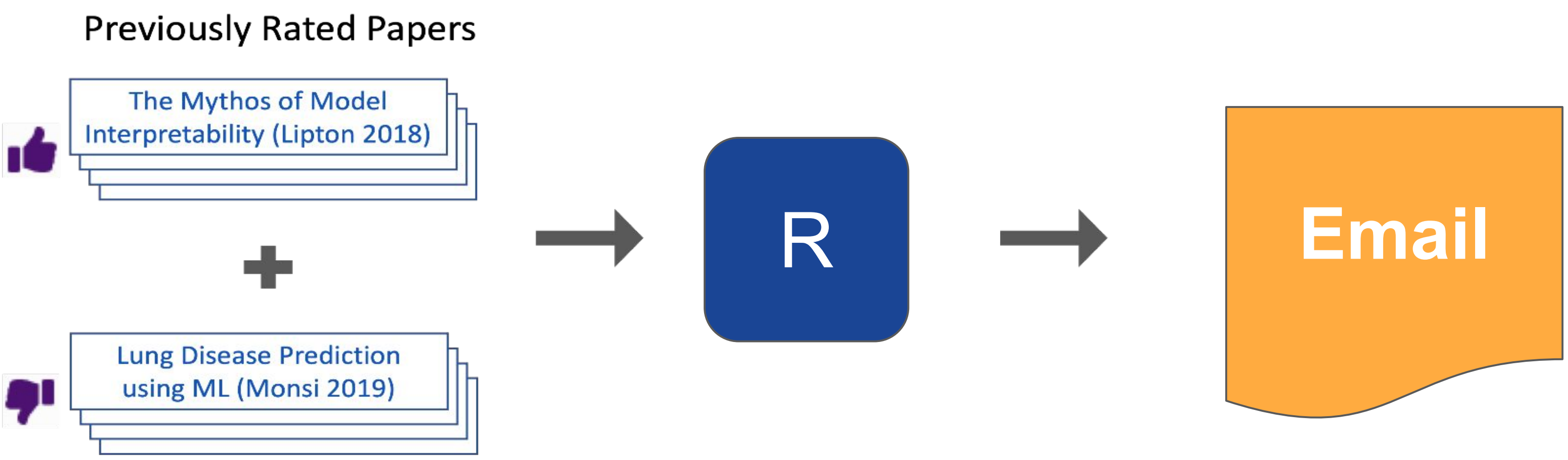
Traditional Methods · Published Research · Ongoing Research

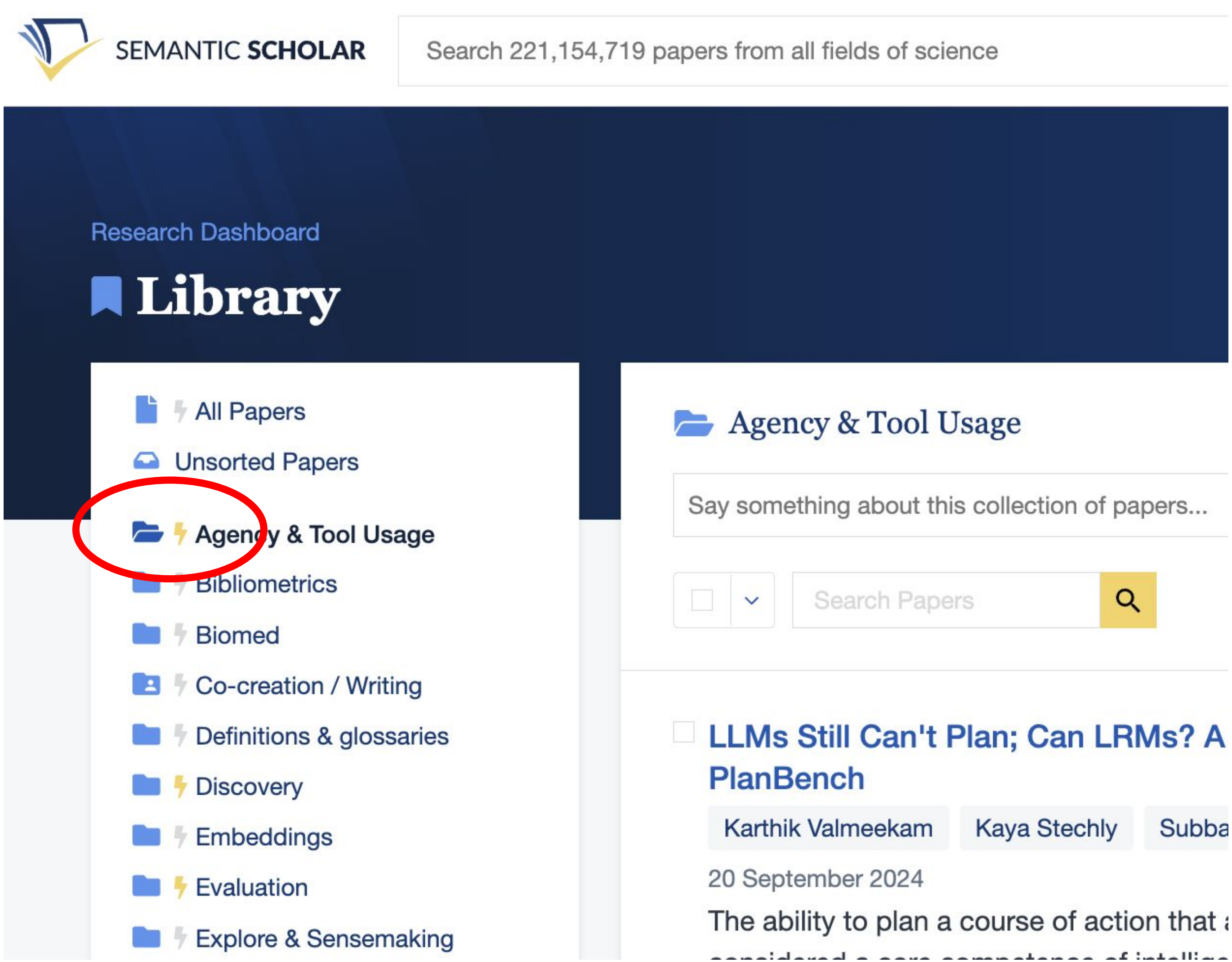
Adaptive Recommendations

Papers you've rated, saved to library, or browsed

S2 learns recommender

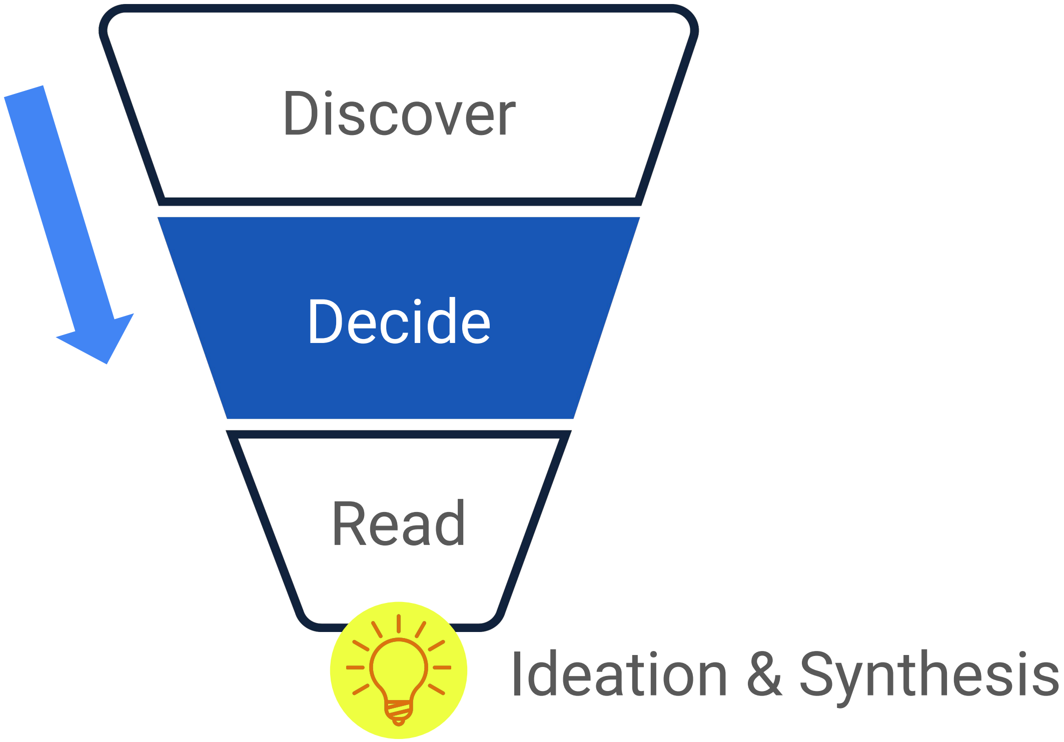
S2 suggests new papers that user will find relevant




SEMANTIC SCHOLAR
Search 221,154,719 papers from all fields of science
Research Dashboard
Library
All Papers
Unsorted Papers
Agency & Tool Usage
Bibliometrics
Biomed
Co-creation / Writing
Definitions & glossaries
Discovery
Embeddings
Evaluation
Explore & Sensemaking
Agency & Tool Usage
Say something about this collection of papers...
Search Papers
LLMs Still Can't Plan; Can LRMs? A
PlanBench
Karthik Valmeekam
Kaya Stechly
Subba
20 September 2024
The ability to plan a course of action that

Ai2

Decide!



Decide - Our Research

Discover

One Size Fits All

- Extract Figures, Tables
- TLDRs Findings’20
- Qlarify UIST’24
- Identifying Influential Cit’ns AAAI wkshp’15
- Intel.Tables Ongoing
- CorpusQA Ongoing

Personalized

- WhoYouKnow CHI’22
- ACCoRD EMNLP’22 Demo
- PaperWeaver CHI’24

Traditional Methods | Published Research | Ongoing Research

Ai2

Extreme Summary TL;DRs

Decide

[Cachola *et al.* Findings EMNLP20]

Ai2

UW

TL;DRs *vs.* Abstracts

Decide

TLDR We develop a new topological complexity measure for deep neural networks and demonstrate that it captures their salient properties.

Abs

more fathoma

terrogating the network with input data

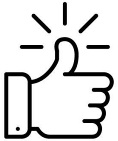
terizing and monitoring structural properties, however, have not been developed. In this work, we propose neural persistence, a complexity measure for neural network architectures based on topological data analysis on weighted stratified graphs. [...]

Intro [...] In this work, we present the following contributions: We introduce neural persistence, a novel measure for characterizing the structural complexity of neural networks that can be efficiently computed. We prove its theoretical properties, such as upper and lower bounds, thereby arriving at a normalization for comparing neural networks of varying sizes. [...]

Conclusion [...] However, this did not yield an early stopping measure because it was never triggered, thereby suggesting that neural persistence captures salient information that would otherwise be hidden among all the weights of a network [...]

How to Evaluate TLDRs

- **Time on page?**
- **Click-thru rate?**
- **Bounce rate?**

- **User feedback:** 

Then… We added TLDRs to Email Alerts

- **Tested:** Impact of replacing truncated abstracts with model generated TLDRs in alert ***emails***.

- **Expected:** An increase in clicks and saves.
- **Hoped:** Unsubscribe rate doesn't increase.

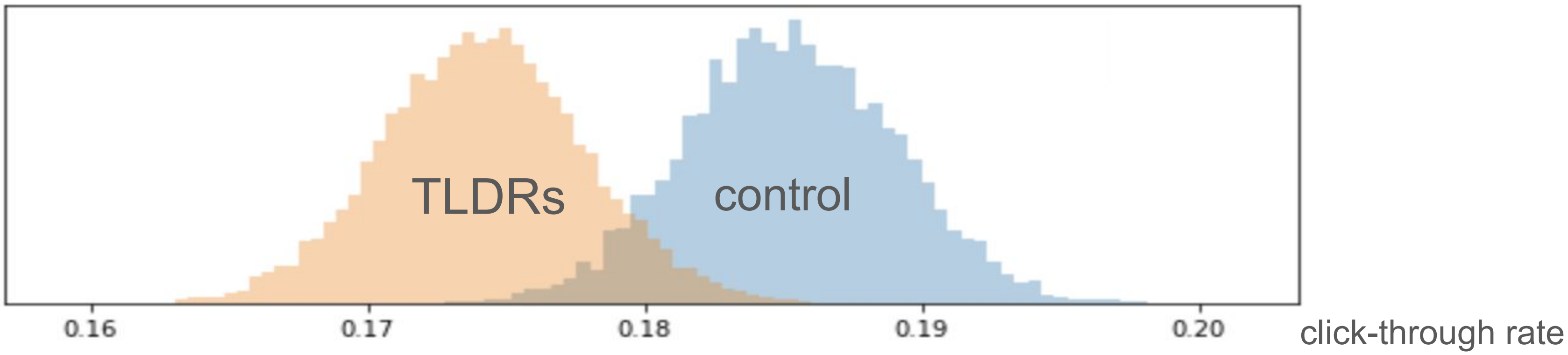
Dany Haddad. Medium, March 2023

https://medium.com/ai2-blog/tldrs-help-to-cut-through-the-clutter-3ad802caed93

Evaluating TLDRs in Alert Emails

Decide

- No increase in unsubscribe rate 😃
- Stat significant 6% ***decrease*** in CTR 😬

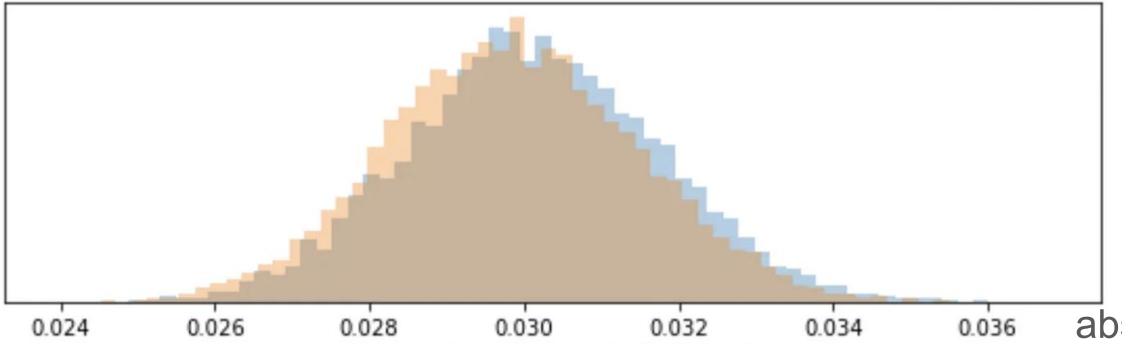


Why? Lower engagement? Greater efficiency?

Click-thru Rate ***Conditioned on Action***

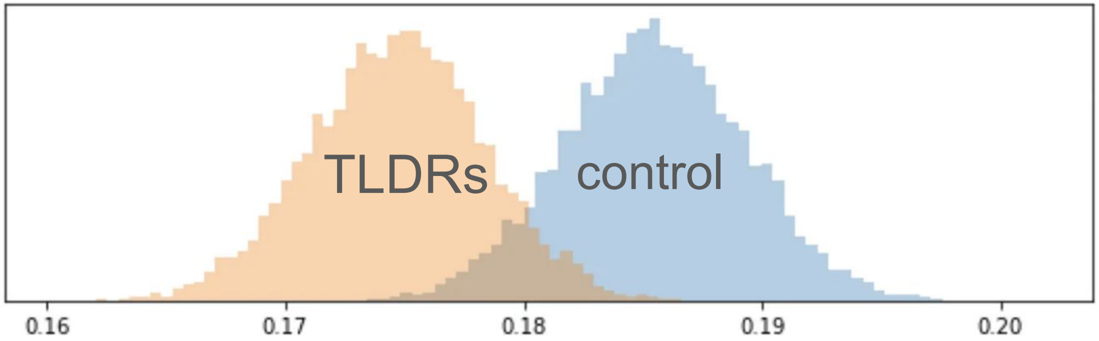
Decide

Email → Click-through → Save to Library



absolute click-through rate

Email → Click-through → ***NO*** Save to Library



abs. click-through rate

Ai2

AI-Powered Question Answering

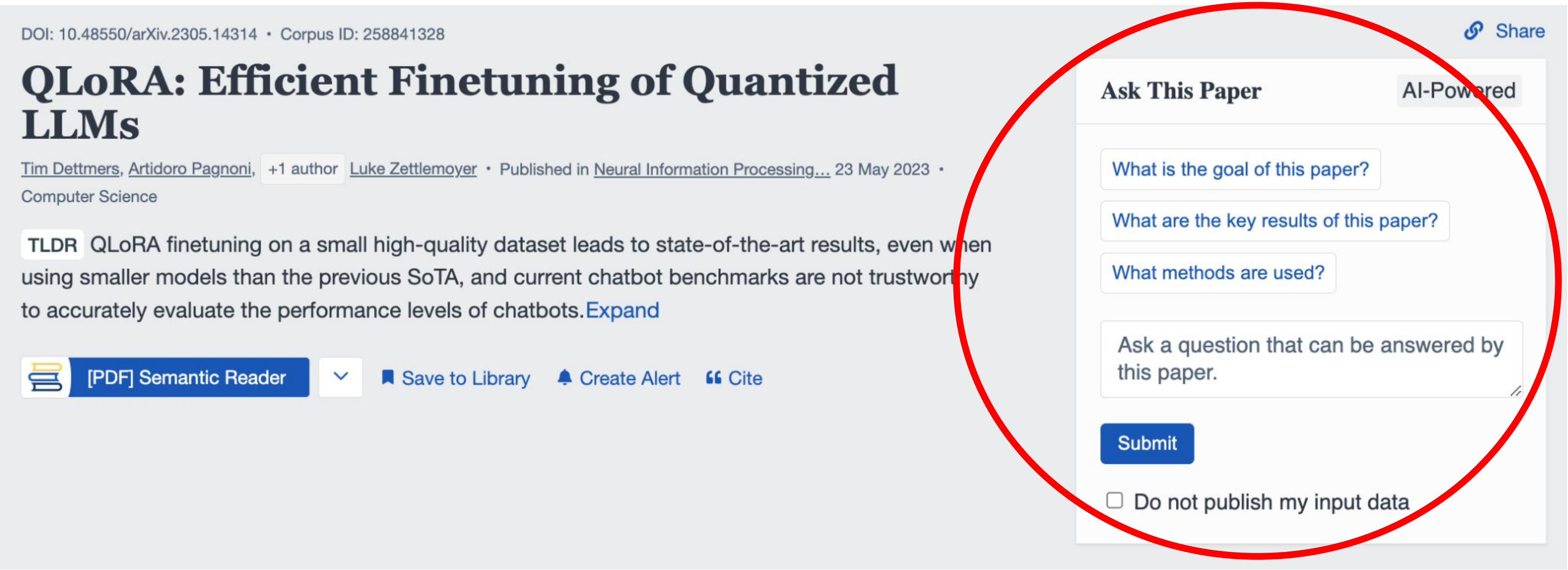
Decide

Paper-Specific QA (asking about specific paper) ✅

vs.

CorpusQA (asking a question over millio
of papers) ⚠️

Paper-Specific Question Answering v1



Paper-Specific Question Answering v2

Decide

How does double quantization reduce..

Define Expand ? Why

QLoRA introduces a number of innovations to s
performance: (a) 4-bit NormalFloat (NF4), a nev
theoretically optimal for normally distributed weights (b) double quantization to reduce the average memory footprint by quantizing the quantization constants, and (c) paged optimizers to manage memory spikes.

[Fok *et al.* UIST'24]

Ai2

Paper-Specific Question Answering ... v2

Decide

QLoRA i
perform
theoreti
reduce t
and (c) p

Define "double quantization"

Do
co
pa
fo

k-bit NormalFloat (NFk), since the data type is information-theoretically optimal for zero-centered normally distributed data. The exact values of this data type can be found in Appendix E.

Double Quantization We introduce *Double Quantization* (DQ), the process of quantizing the quantization constants for additional memory savings. While a small blocksize is required for precise 4-bit quantization [13], it also has a considerable memory overhead. For example, using 32-bit constants and a blocksize of 64 for $\mathbf{W}$, quantization constants add $32/64 = 0.5$ bits per parameter on average. Double Quantization helps reduce the memory footprint of quantization constants.

More specifically, Double Quantization treats quantization constants c_2^{FP32} of the first quantization as inputs to a second quantization. This second step yields the quantized quantization constants c_2^{FP8} and the second level of quantization constants c_1^{FP32}. We use 8-bit Floats with a blocksize of 256 for the second quantization as no performance degradation is observed for 8-bit quantization, in line with results from Dettmers and Zettlemoyer [13]. Since the c_2^{FP32} are positive, we subtract the mean from c_2 before quantization to center the values around zero and make use of symmetric quantization. On average, for a blocksize of 64, this quantization reduces the memory footprint per parameter from $32/64 = 0.5$ bits, to $8/64 + 32/(64 \cdot 256) = 0.127$ bits, a reduction of 0.373 bits per parameter.

Paged Optimizers use the NVIDIA unified memory [3] feature wich does automatic page-to-page transfers between the CPU and GPU for error-free GPU processing in the scenario where the GPU occasionally runs out-of-memory. The feature works like regular memory paging between CPU RAM and the disk. We use this feature to allocate paged memory for the optimizer states which are then automatically evicted to CPU RAM when the GPU runs out-of-memory and paged back into GPU memory when the memory is needed in the optimizer update step.

QLoRA. Using the components described above, we define QLoRA for a single linear layer in the quantized base model with a single LoRA adapter as follows:

$$\mathbf{Y}^{\text{BF16}} = \mathbf{X}^{\text{BF16}}\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{NF4}}) + \mathbf{X}^{\text{BF16}}\mathbf{L}_1^{\text{BF16}}\mathbf{L}_2^{\text{BF16}}, \tag{5}$$

where doubleDequant$(\cdot)$ is defined as:

$$\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{k-bit}}) = \text{dequant}(\text{dequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}), \mathbf{W}^{\text{4bit}}) = \mathbf{W}^{\text{BF16}}, \tag{6}$$

We use NF4 for $\mathbf{W}$ and FP8 for c_2. We use a blocksize of 64 for $\mathbf{W}$ for higher quantization precision and a blocksize of 256 for c_2 to conserve memory.

For parameter updates only the gradient with respect to the error for the adapters weights $\frac{\partial E}{\partial \mathbf{L}_i}$ are needed, and not for 4-bit weights $\frac{\partial E}{\partial \mathbf{W}}$. However, the calculation of $\frac{\partial E}{\partial \mathbf{L}_i}$ entails the calculation of $\frac{\partial \mathbf{X}}{\partial \mathbf{W}}$ which proceeds via equation (5) with dequantization from storage $\mathbf{W}^{\text{NF4}}$ to computation data type $\mathbf{W}^{\text{BF16}}$ to calculate the derivative $\frac{\partial \mathbf{X}}{\partial \mathbf{W}}$ in BFloat16 precision.

To summarize, QLoRA has one storage data type (usually 4-bit NormalFloat) and a computation data type (16-bit BrainFloat). We dequantize the storage data type to the computation data type to perform the forward and backward pass, but we only compute weight gradients for the LoRA parameters which use 16-bit BrainFloat.

5

See in paper context

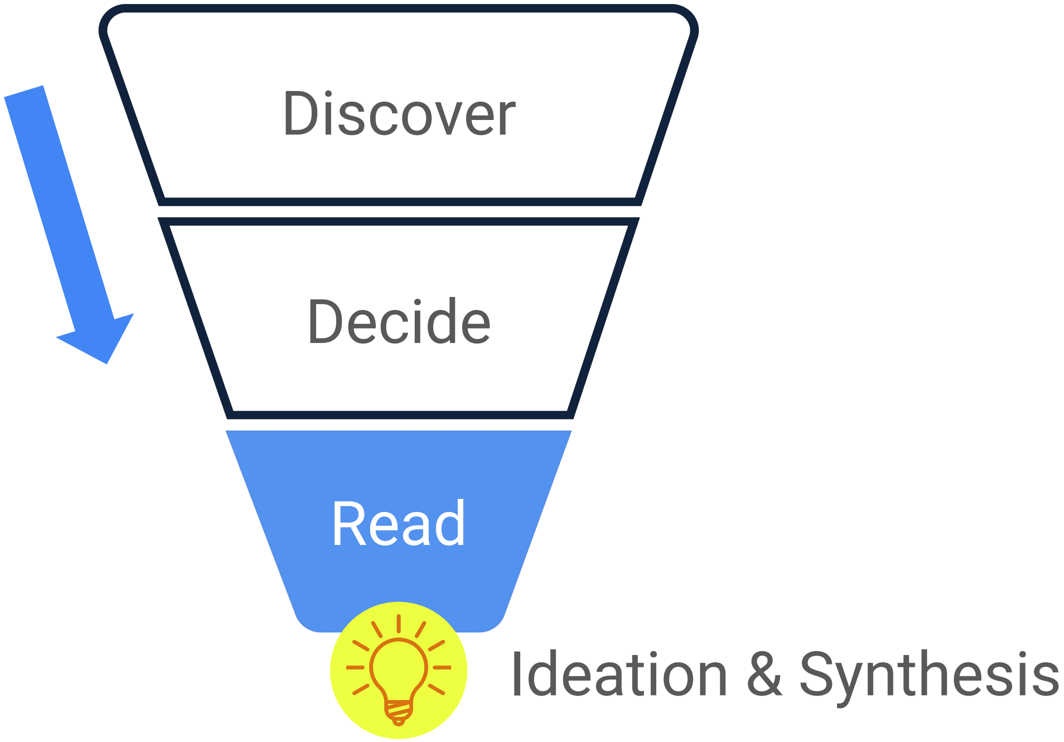
(DQ), the process of
ts for additional memory
equired for precise 4-bit
derable memory
t constants and a
constants add 32/64 = 0.5
ble Quantization helps
ntization constants...

quantization
onstants. The
e memory

ribution!

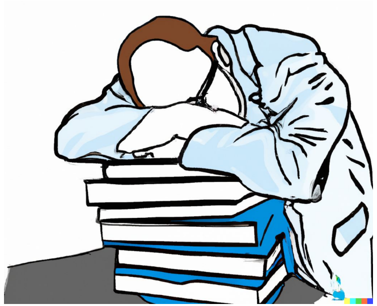
Ai2

Read!



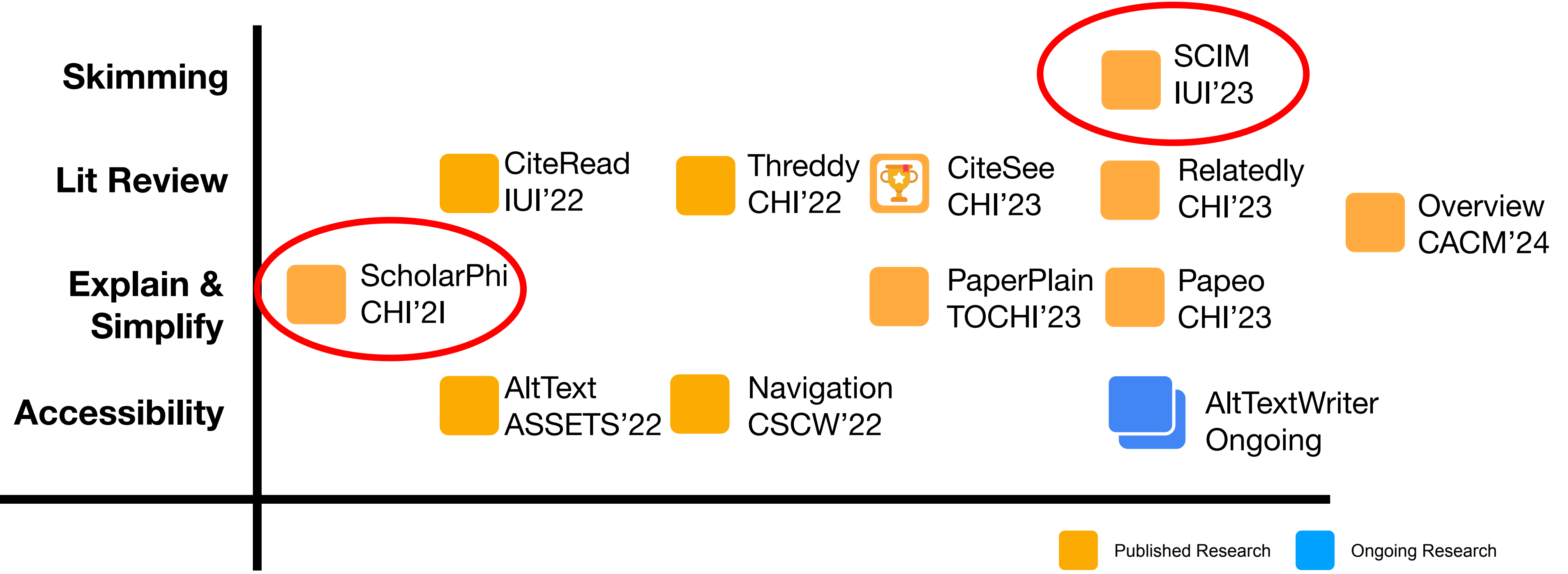
Reading Technical Papers is Tiring!

- Researchers flipped back and forth in document…
- Looking for definitions, checking citations…

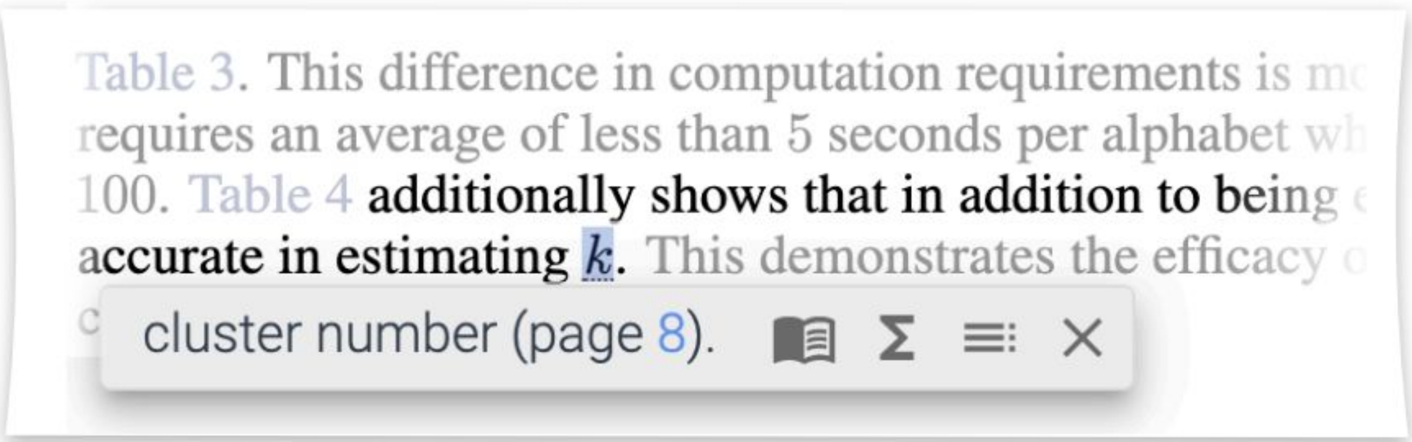


Semantic Reader Project

Read



Term & Symbol Definitions



AI-Powered Skimming

Goal

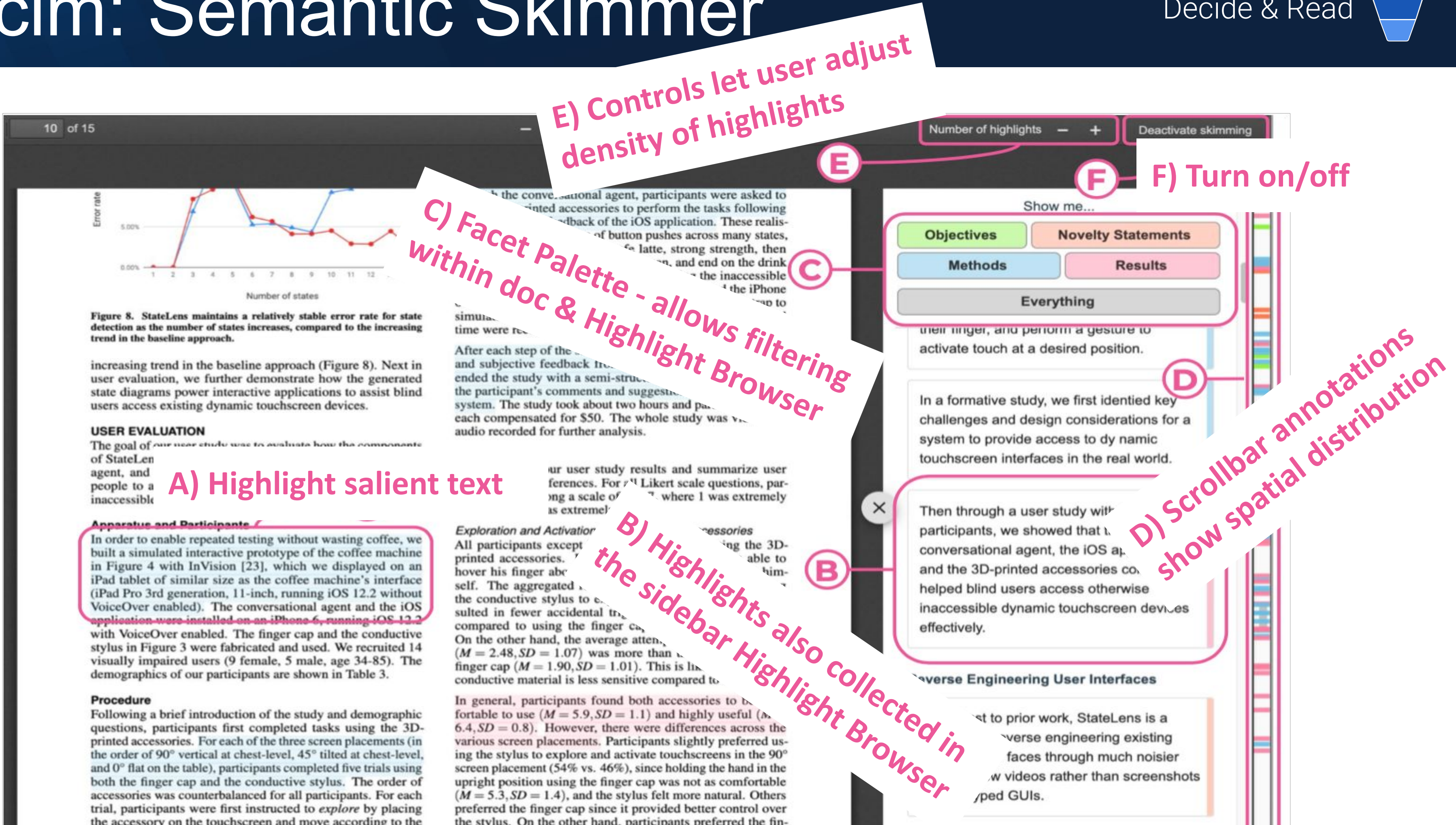
Enable skimming in less time, with less effort.

Challenge

How surface *salient* content w/ non-intrusive affordances?

[Fok *et al.* IUI'23]

Scim: Semantic Skimmer



AI-Powered Skimming

Decide & Read

100%

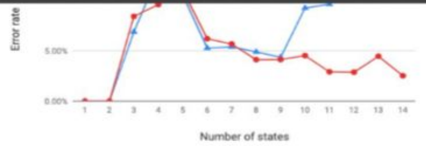


Figure 8. StateLens maintains a relatively stable error rate for state detection as the number of states increases, compared to the increasing trend in the baseline approach.

increasing trend in the baseline approach (Figure 8). Next in user evaluation, we further demonstrate how the generated state diagrams power interactive applications to assist blind users access existing dynamic touchscreen devices.

USER EVALUATION

The goal of our user study was to evaluate how the components of StateLen agent, and people to a inaccessible

A) Highlight salient text

Apparatus and Participants

In order to enable repeated testing without wasting coffee, we built a simulated interactive prototype of the coffee machine in Figure 4 with InVision [23], which we displayed on an iPad tablet of similar size as the coffee machine's interface (iPad Pro 3rd generation, 11-inch, running iOS 12.2 without VoiceOver enabled). The conversational agent and the iOS application were installed on an iPhone 6, running iOS 12.2 with VoiceOver enabled. The finger cap and the conductive stylus in Figure 3 were fabricated and used. We recruited 14 visually impaired users (9 female, 5 male, age 34-85). The demographics of our participants are shown in Table 3.

Procedure

Following a brief introduction of the study and demographic questions, participants first completed tasks using the 3D-printed accessories. For each of the three screen placements (in the order of 90° vertical at chest-level, 45° tilted at chest-level, and 0° flat on the table), participants completed five trials using both the finger cap and the conductive stylus. The order of accessories was counterbalanced for all participants. For each trial, participants were first instructed to *explore* by placing the accessory on the touchscreen and move according to the

through the conversational agent, participants were asked to use the 3D-printed accessories to perform the tasks following the guidance and feedback of the iOS application. These realistic tasks involved a series of button pushes across many states, *e.g.*, select gourmet drinks, cafe latte, strong strength, then confirm, auto-select default coffee bean, and end on the drink preparation screen. The iPad Pro simulating the inaccessible coffee machine was placed tilted at chest level, and the iPhone 6 running the iOS application was mounted on a head strap to simulate a head-mounted camera. Task completion rate and time were recorded.

After each step of the study, we collected Likert scale ratings and subjective feedback from the participants. Finally, we ended the study with a semi-structured interview asking for the participant's comments and suggestions on the StateLens system. The study took about two hours and participants were each compensated for $50. The whole study was video and audio recorded for further analysis.

ur user study results and summarize user ferences. For all Likert scale questions, par- ng a scale of 1 to 7, where 1 was extremely s extremely positive.

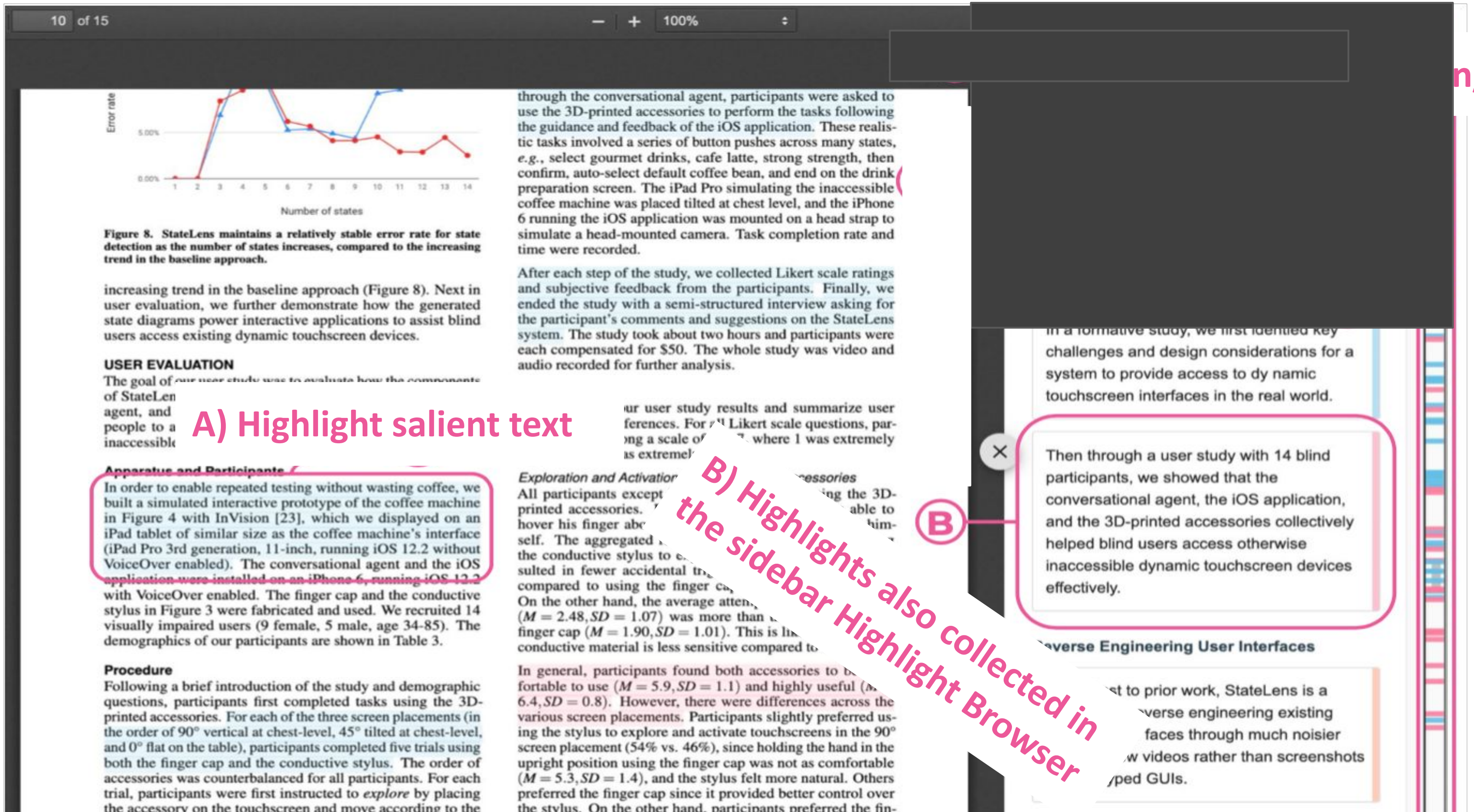
Exploration and Activation with 3D-Printed Accessories

All participants except P12 completed tasks using the 3D-printed accessories. P12 had low vision, and was able to hover his finger above the target and then activate by himself. The aggregated results are shown in Table 4. Using the conductive stylus to explore touchscreens generally resulted in fewer accidental triggers ($M = 0.03, SD = 0.16$) compared to using the finger cap ($M = 0.07, SD = 0.27$). On the other hand, the average attempts of using the stylus ($M = 2.48, SD = 1.07$) was more than that from using the finger cap ($M = 1.90, SD = 1.01$). This is likely because the conductive material is less sensitive compared to fingers.

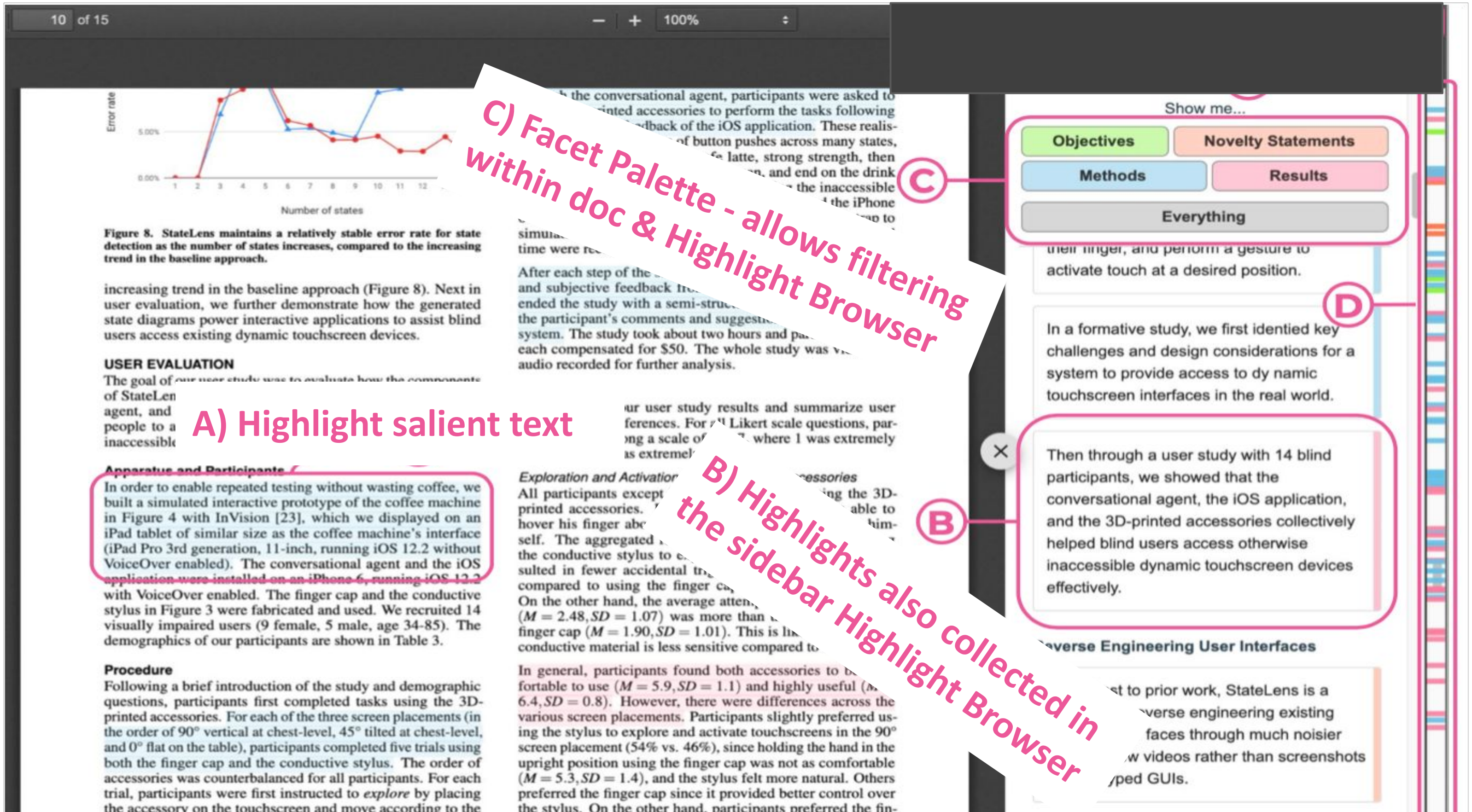
In general, participants found both accessories to be comfortable to use ($M = 5.9, SD = 1.1$) and highly useful ($M = 6.4, SD = 0.8$). However, there were differences across the various screen placements. Participants slightly preferred using the stylus to explore and activate touchscreens in the 90° screen placement (54% vs. 46%), since holding the hand in the upright position using the finger cap was not as comfortable ($M = 5.3, SD = 1.4$), and the stylus felt more natural. Others preferred the finger cap since it provided better control over the stylus. On the other hand, participants preferred the fin-

AI-Powered Skimming

Decide & Read

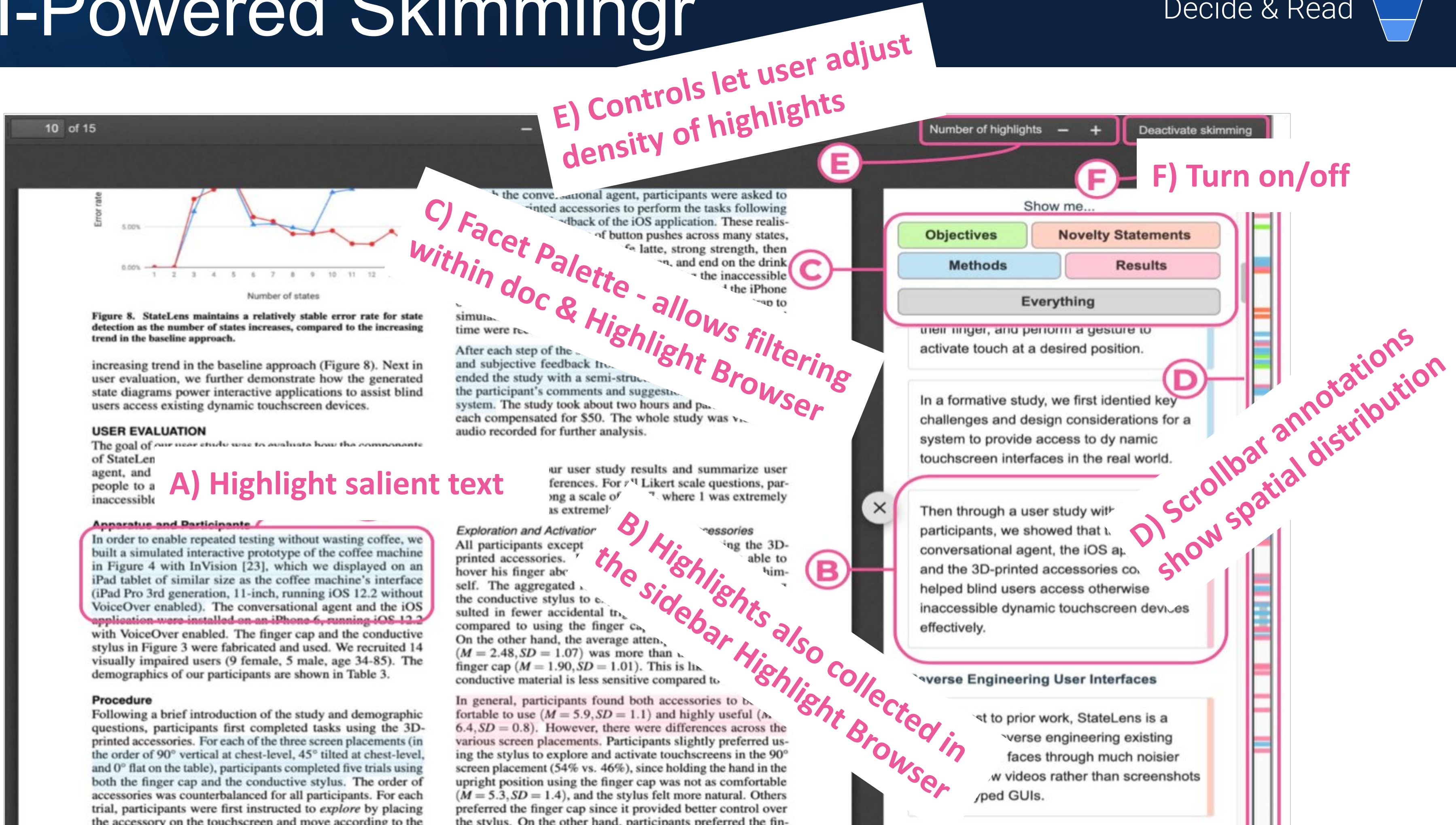


AI-Powered Skimming

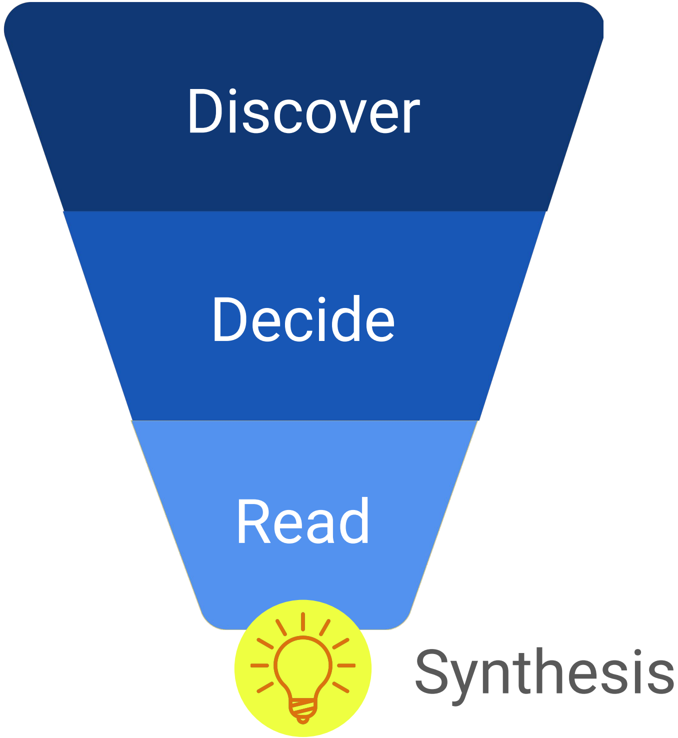


AI-Powered Skimmingr

Decide & Read



AI Helpful at Every Stage



Seek tools that include attribution & verification

Corpus QA techniques still developing

Thanks

Many co-authors and collaborators, but especially Raymond Fok, Marti Hearst, Andrew Head, Joseph Chang