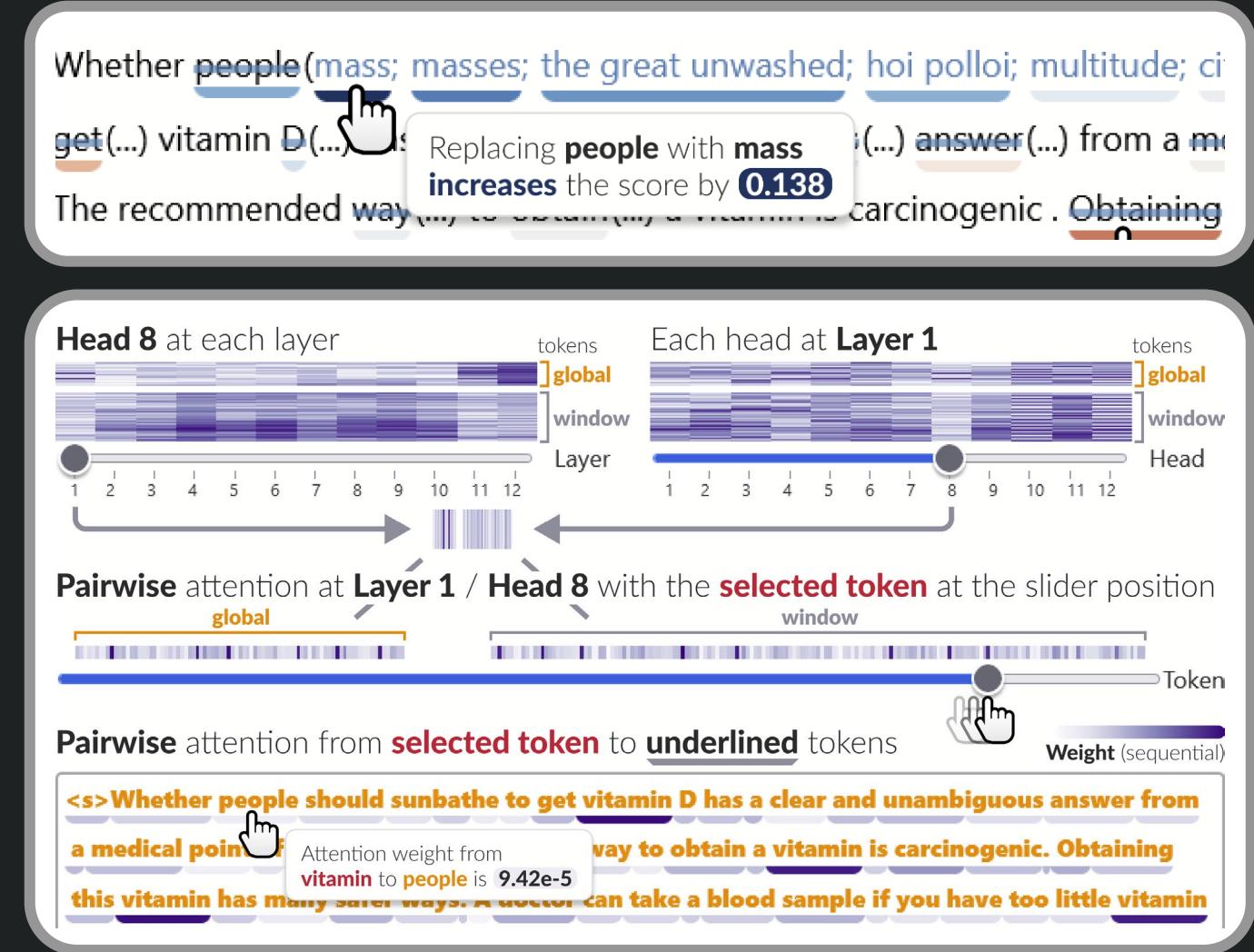


iScore

Visual Analytics for Interpreting How Language Models Automatically Score Summaries

Adam **Coscia**, Langdon **Holmes**,
Wesley **Morris**, Joon Suh **Choi**,
Scott **Crossley**, Alex **Endert**



Visual Analytics Lab
@ Georgia Tech



Summary writing helps students learn

Textbook Source

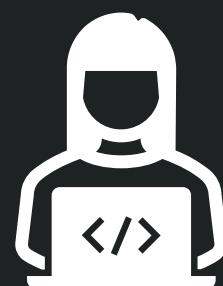
According to a recent study, there is no safe way to sunbathe. Even small doses of ultraviolet radiation from the sun and tanning beds may cause cancer. Ultraviolet (UV) radiation is one of the most frequent causes of cancer, but it can be avoided. So says Professor David E. Fisher in an interview with slate.com about his most recent study of the relationship between UV radiation and skin cancer.

Learner Summaries

Fisher's study shows that people who have used a tanning bed before the age of 35 have a 75% higher likelihood of developing skin cancer than those who have not used a tanning bed at such an early age. In fact, the results show that having used a tanning bed even once results in a higher risk of skin cancer.

According to a study conducted in Norway, "each year, approximately 250 people die in Norway from skin cancer primarily because of excessive sunbathing. The risk of cancer increases, since many are not sufficiently careful in applying sunscreen and taking breaks in the shade. There are grounds for serious concern when nearly 30% of adolescents report that they are "completely certain" that they will be sunburned during their holidays. Having sunburn increases the overall risk of skin cancer."

Learning engineers trained LLMs that score student summaries of textbook sources



LLMs can help score summary writing, making it scalable!

Unclear how LLMs score summaries

Learner Summaries

Learners write **summaries** of the textbook **source** sections, to be automatically scored

- 1 This summary is plagiarised **verbatim** and thus both Content and Wording are scored **low**

Fisher's study shows that people who have used a tanning bed before the age of 35 have a 75% higher likelihood of developing skin cancer than those who have not used a tanning bed at such an early age. In fact, the results show that having used a tanning bed even once results in a higher risk of skin cancer.

According to a study conducted in Norway, "each year, approximately 250 people die in Norway from skin cancer primarily because of excessive sunbathing. The risk of cancer increases, since many are not sufficiently careful in applying sunscreen and taking breaks in the shade. There are grounds for serious concern when nearly 30% of adolescents report that they are "completely certain" that they will be sunburned during their holidays. Having sunburn increases the overall risk of skin cancer.

Expert Scores

LLMs train to replicate **z-scores** of 2 PCA features from an expert-scored rubric

Content

1	2	3	4	
				Main idea 2.5
				Details 2.0
				Cohesion 2.0
				Objective 3.5

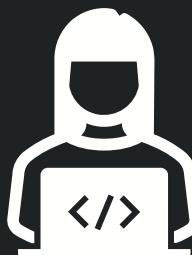


PCA: 5.86 → Z-score: **-1.049** X

Wording

1	2	3	4	
				Paraphrase 1.0
				Source text 2.0

PCA: 2.06 → Z-score: **-1.116** X



- ✗ Hard to characterize **relationship** between **summary** and **source**
- ✗ How will LLMs score **summaries** they have **never seen before?**

Our Goal

Build an **interactive** tool for **engineers** to evaluate **qualitative** relationships that LLMs have **learned** before deploying them!



How to evaluate LLMs at scale?

Interpreting LLM behaviors



- Hard to identify factors (synonyms, grammar, etc.)
- Track changes in LLM scores across summaries

Analyzing LLM data



- Too many tokens / parameters to analyze at once
- Unique architectures require new analysis tools

Our solution: iScore!

 **Run**
multiple
LLMs in
real-time

 **Input** a
source text
and any
number
summaries

A Assignments Panel

LLMs to inference:

- Content
- Content (global)
- Wording
- Wording (global)

Try an example!

Choose

Source  Upload

"We are thus in a situation where people are recommended to use something that we know is carcinogenic to obtain a vitamin. There are many far safer ways to obtain this vitamin. If you suspect that you have too little vitamin D, go to your doctor to have a blood sample taken. If the test shows that you have too little of this important vitamin, the solution ought to be cod-liver oil or a vitamin D supplement. From a medical point of view, this is a very simple issue, and we have a clear and unambiguous answer to the question of whether people should sunbathe to obtain enough vitamin D. We have safe ways to measure people's level of vitamin D, and we have safe ways to treat vitamin D deficiency if necessary. There is thus no reason to try to guess how much UV radiation one needs to stay healthy. This is neither very smart, nor very healthy," Fisher says.

Summaries  Upload

ID: 1

Perturb:
 grammar?
 sentences?
 words?
 tokens?
 Attention?
 tokens?

ID: 2 

Perturb:
 grammar?
 sentences?
 words?
 tokens?
 Attention?
 tokens?

+ Add Summary

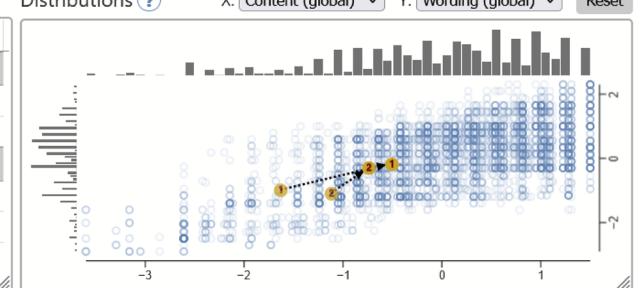
B Scores Dashboard

Show: History Distributions

History 

Run	Content	Content (global)	Wording	Wording (global)
Summary ID: 1 (2 items)	-	-1.6181116	-	-0.9755589
1	-	-1.6181116	-	-0.9755589
2	-	-0.51558775	-	-0.18998466
Summary ID: 2 (2 items)	-	-1.1149579	-	-1.0743605
1	-	-1.1149579	-	-1.0743605
2	-	-0.74383956	-	-0.29991263

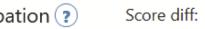
Distributions  X: Content (global) Y: Wording (global) Reset

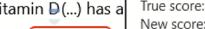
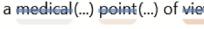
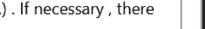
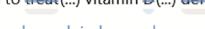
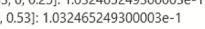
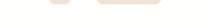
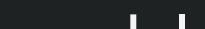


C Model Analysis View

Summary: 2 

Show: Input Perturbation Token Attention

Input Perturbation  Score diff: -4.96e-1  3.82e-1 Model: Wording (global) Perturb: words (replace) Color by: Diff (true) 

Whether people (...) should sunbathe(...) to get(...) vitamin D(...) has a                                          

Token Attention  Weight: 0  5.18e-5 Window: 256 tokens Model: Wording (global) Heatmaps

Layer: 1 2 3 4 5 6 7 8 9 10 11 12 Head: 1 2 3 4 5 6 7 8 9 10 11 12 Token: 1 2 3 4 5 6 7 8 9 10 11 12

<s> Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view. If necessary, there are safe ways to treat vitamin D deficiency and to measure level of vitamin D in people. Guessing how much UV radiation one needs to stay healthy is neither very smart nor very healthy. There is no reason to try to guess.</s> Sunbathing causes cancer \n\nAccording to a recent study, there is no safe way to sunbathe. Even small doses of ultraviolet radiation from the sun and tanning beds may cause cancer.\n\nUltraviolet (UV) radiation is one of the most frequent causes of cancer, but it can be avoided. So says Professor David E. Fisher in an interview with slate.com about his most recent study of the relationship between UV radiation and skin cancer.\n\nFisher's study shows that people who have used a tanning bed before the age of 35 have a 75% higher risk of developing skin cancer than those who have not used a tanning bed at such an early age. In fact, the results show that having used a tanning bed even once increases the risk of skin cancer by 20%. According to a study conducted in Norway, each year, approximately 250 people die in Norway from skin cancer primarily because of excessive sunbathing. The risk of cancer increases, since many are not sufficiently careful in applying sunscreen and taking breaks in the sun. There are grounds for serious concern when nearly 30% of

 **Track** 
changes in
scores over
time

 **Perturb** 
model
inputs and
analyze
model
outputs

Example: Why are scores different?

- 1 This summary is plagiarised **verbatim** and thus both Content and Wording are scored **low**

Fisher's study shows that people who have used a tanning bed before the age of 35 have a 75% higher likelihood of developing skin cancer than those who have not used a tanning bed at such an early age. In fact, the results show that having used a tanning bed even once results in a higher risk of skin cancer.

According to a study conducted in Norway, "each year, approximately 250 people die in Norway from skin cancer primarily because of excessive sunbathing. The risk of cancer increases, since many are not sufficiently careful in applying sunscreen and taking breaks in the shade. There are grounds for serious concern when nearly 30% of adolescents report that they are "completely certain" that they will be sunburned during their holidays. Having sunburn increases the overall risk of skin cancer.

Content			
1	2	3	4
			Main idea 2.5
			Details 2.0
			Cohesion 2.0
			Objective 3.5

PCA: 5.86 → Z-score: **-1.049**

Wording			
1	2	3	4
			Paraphrase 1.0
			Source text 2.0

PCA: 2.06 → Z-score: **-1.116**



- 3 This summary is **original** and scores **higher** in both Content and Wording!

Sunbathing can cause cancer. It can be dangerous to use a tanning bed according to Professor David E. Fisher who's research has shown that people under 35 who go tanning have a 75% of getting skin cancer. In Norway excessive sunbathing has lead to the deaths of 250 people each year. These risks increase due to lack of using sunscreen and not taking a break in the shade. Even getting a sunburn increases your cancer risk. It is not recommended to get Vitamin D from laying out in the sun. He recommends if you are deficient in Vitamin D to take a supplement or use cod-liver oil instead. This way you can be sure to get the Vitamin D you need without risking your health.

Content			
1	2	3	4
			Main idea 2.5
			Details 3.5
			Cohesion 3.5
			Objective 2.5

PCA: 8.83 → Z-score: **0.407**

Wording			
1	2	3	4
			Paraphrase 2.5
			Source text 2.5

PCA: 3.93 → Z-score: **0.331**



How to **identify** decision criteria, **compare** summaries, **analyze** weights?

Assignments Panel

The screenshot shows the iScore assignments panel interface. On the left, a 'Source' box contains a text about vitamin D. A large yellow arrow points from this source text to a summary box. The summary box contains two entries: one for ID: 1 and one for ID: 2. Each summary has a 'Perturb:' section with checkboxes for grammar, sentences, words, and tokens. A blue button at the bottom right of the summary area says '+ Add Summary'.

Source (?)

Upload

"We are thus in a situation where people are recommended to use something that we know is carcinogenic to obtain a vitamin. There are many far safer ways to obtain this vitamin. If you suspect that you have too little vitamin D, go to your doctor to have a blood sample taken. If the test shows that you have too little of this important

Summaries (?)

Upload

ID: 1

Perturb:

grammar?

The recommended way to obtain a vitamin is carcinogenic. Obtaining this vitamin has many safer ways. A doctor can take a blood sample if you have too little vitamin D oil or vitamin D supplement solutions if too little of important vitamin shows i

ID: 2 (x)

Perturb:

grammar?

sentences?

words?

tokens?

Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view. If necessary, there are safe ways to treat vitamin D deficiency and to measure level of vitamin D in people. Guessing how much UV radiation one needs to stay healthy is neither very smart nor very healthy. There is no reason to try to guess.

+ Add Summary

- Input a **source** text and multiple **summaries** to compare

Assignments Panel

Source [?](#)

Upload

"We are thus in a situation where people are recommended to use something that we know is carcinogenic to obtain a vitamin. There are many far safer ways to obtain this vitamin. If you suspect that you have too little vitamin D, go to your doctor to have a blood sample taken. If the test shows that you have too little of this important

Summaries [?](#)

The recommended way to obtain a vitamin is carcinogenic. Obtaining this vitamin has many safer ways. A doctor can take a blood sample if you have too little vitamin D oil or vitamin D supplement solutions if too little of important vitamin shows i

Upload

ID: 1

Perturb:

grammar?

ID: 2 [X](#)

Perturb:

grammar?

sentences?

words?

tokens?

attention:

tokens?

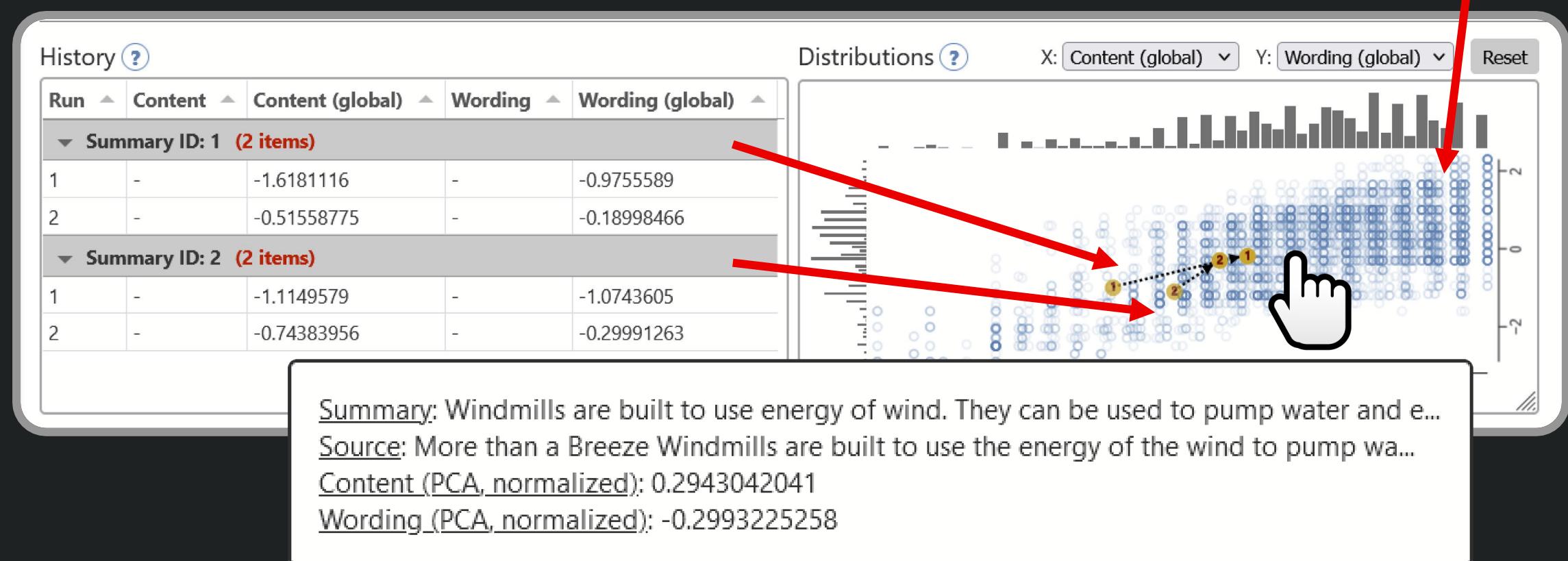
+ Add Summary

Scores Dashboard



Scores Dashboard

LLM training data (expert-scored)

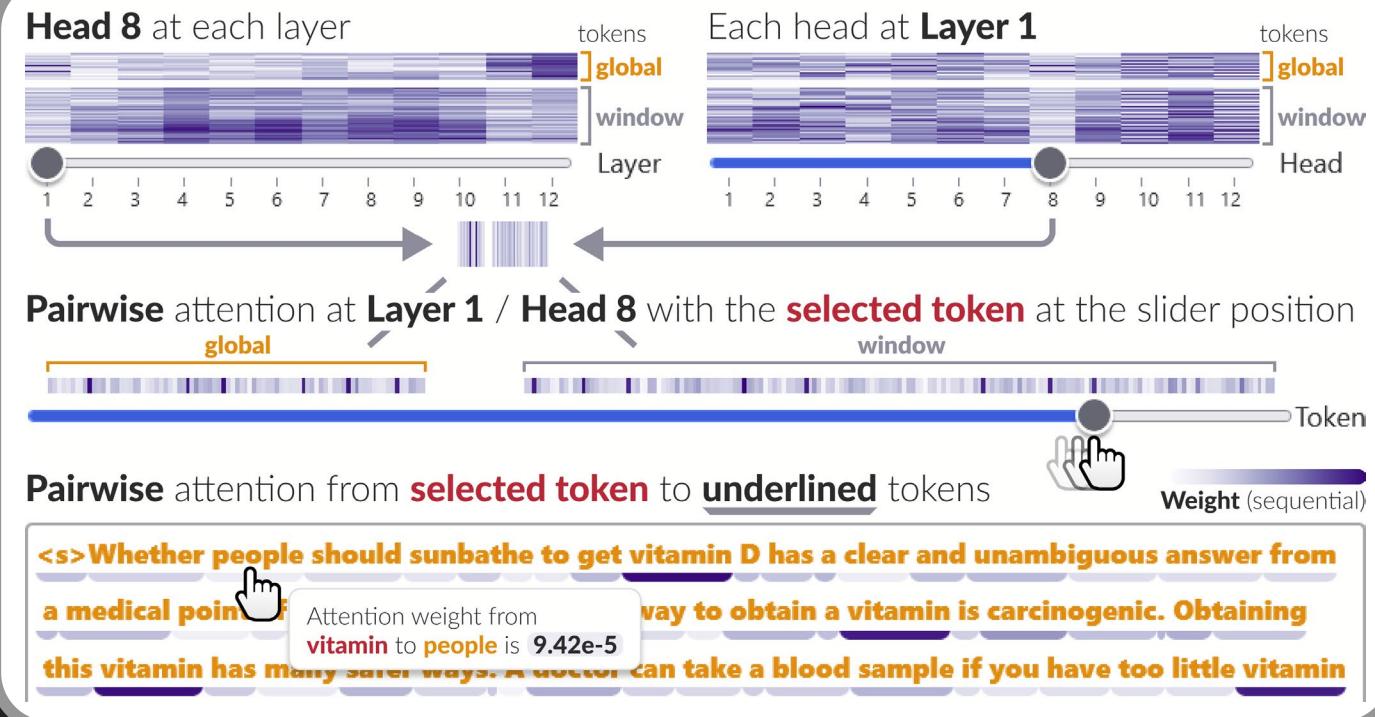


- Track changes in scores across **runs** (i.e., each time the LLMs are run)
- Show LLM scores in context of **expert training data**

Qualitative Model Analysis

Input Perturbation

Whether people (mass; masses; the great unwashed; hoi polloi; multitude; ci
get(...) vitamin D (...) Replacing people with mass (...) answer(...) from a m
The recommended way increases the score by 0.138 carcinogenic . Obtaining



Attention Visualization

Input Perturbation

Token / word / sentence revisions are **automatically** applied to the summary and **re-scored**. The revisions are underlined and colored by the **difference** in score between the original summary and the revised summary. In the example summary below, the original Content score is **0.682**

Original

This is a bad summary.

Content : -0.186

Wording : -0.425

Perturbation

This is a **bad** better summary :-!



✓ **Content** : 0.223 (+0.409) ▲

✗ **Wording** : -0.599 (-0.174) ▼

Input Perturbation

Token / word / sentence revisions are **automatically** applied to the summary and **re-scored**. The revisions are underlined and colored by the **difference** in score between the original summary and the revised summary. In the example summary below, the original Content score is **0.682**

Words - Replace words with **synonyms** using word tokenizer and WordNet

Whether people(mass; masses; the great unwashed; hoi polloi; multitude; citizenry) should sunbathe(...) to
get(...) vitamin D(...): Replacing **people** with **mass** **increases** the score by **0.138**.
The recommended way(...) carcinogenic. Obtaining(incur; prevail; hold; obtain; receive; get; find) this vitamin has many safer(...) ways(...) . A doctor(...) if
you little(...) liver oil(...) or vitamin D(...) supplements(...) are solutions(...) ...
Replacing **Obtaining** with **receive** **decreases** the score by **-0.169**.
ows(...) in your test(...).

True difference (diverging)

Input Perturbation

Token / word / sentence revisions are **automatically** applied to the summary and **re-scored**. The revisions are underlined and colored by the **difference** in score between the original summary and the revised summary. In the example summary below, the original Content score is **0.682**

Sentences - Remove sentences using sentence tokenizer

Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view. The recommended way to ob  Removing this sentence decreases the score by **-0.504**aining this vitamin has many safer ways. A doctor can take a blood sample if you ... cod-liver oil or vitamin D supplement are solutions if too little of this important vitamin shows in your test.

Absolute difference (sequential)

Input Perturbation

Token / word / sentence revisions are **automatically** applied to the summary and **re-scored**. The revisions are underlined and colored by the **difference** in score between the original summary and the revised summary. In the example summary below, the original Content score is **0.682**

Tokens - Mask tokens using LLM tokenizer

Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view. The recommended way to obtain a vitamin is carcinogenic. Obtaining this vitamin has many safer ways. A doctor can take a blood sample if you have too little vitamin D. A cod-liver oil or vitamin D supplement are solutions if too little of this important vitamin shows in your test.

True difference (diverging)

Input Perturbation

Token / word / sentence revisions are **automatically** applied to the summary and **re-scored**. The revisions are underlined and colored by the **difference** in score between the original summary and the revised summary. In the example summary below, the original Content score is **0.682**

Grammar - Correct **spelling** automatically using SymSpellPy

Whether peple should sunbathe, to get vitamin D, hasa unabigous answer froma medical poin of view.



Original - Un-revised summary

Whether people should sunbathe, to get vitamin A, hasa unanimous answer from medical point of view.



Single word correction - Preserves punctuation and casing, but does not understand multi-word errors

whether people should sunbathe to get vitamin a hasa unanimous answer from a medical point of view



Multi-word correction - Compound-aware, but removes punctuation and casing

Whether people should sunbathe to get vitamin Do has a unanimous answer from medical point of view

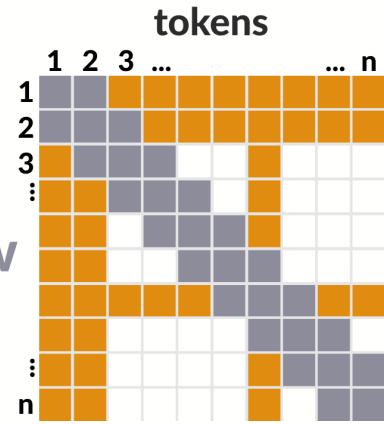


Word segmentation - Breaks apart words, keeps casing but loses punctuation

True difference (diverging)

Token Attention

Transformers independently arrange **heads** in sequential **layers**. Heads compute weights called **attention** between all pairs of tokens for the next layer. Longformers use a unique **sliding attention window** to compute weights between tokens only inside the window. **Global** tokens are assigned attention between all other tokens, as usual.



How are **Longformers** different from other LLMs?

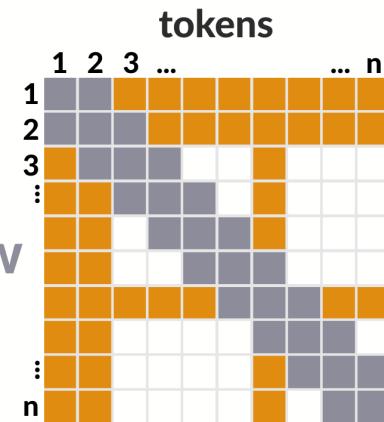
- Can process **4096 token inputs** (more than most LLMs!)
- Use **sliding window** (i.e., not every token has attention!)

✗ How to **scale** visualizations between layers, heads, and window?

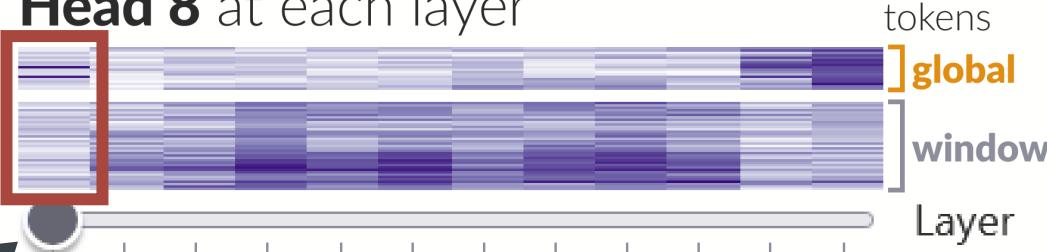
Token Attention

Transformers independently arrange **heads** in sequential **layers**.

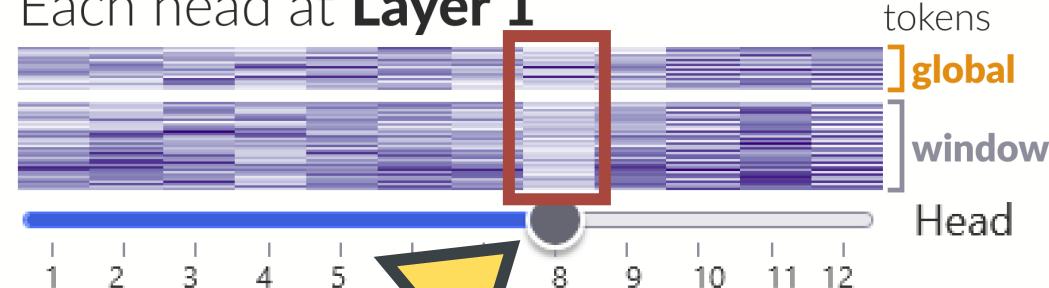
Heads compute weights called **attention** between all pairs of tokens for the next layer. Longformers use a unique **sliding attention window** to compute weights between tokens only inside the window. **Global** tokens are assigned attention between all other tokens, as usual.



Head 8 at each layer



Each head at **Layer 1**



- **Overview:** Heat maps of token attentions across heads / layers

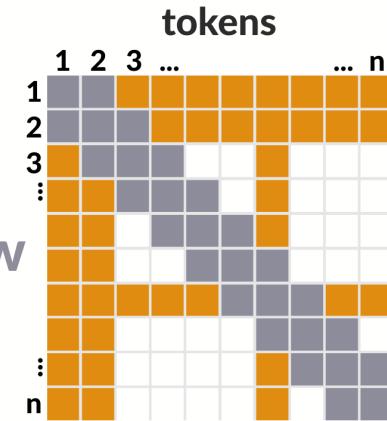


Use **fluid interaction** principles!

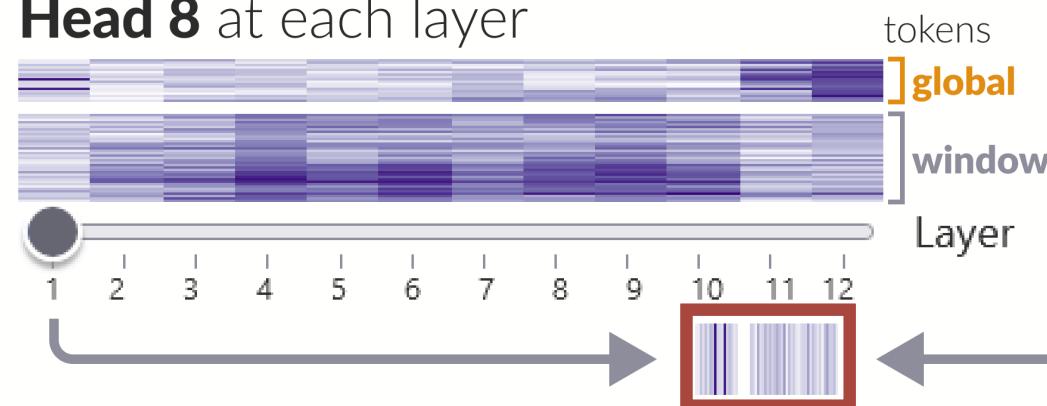
Token Attention

Transformers independently arrange **heads** in sequential **layers**.

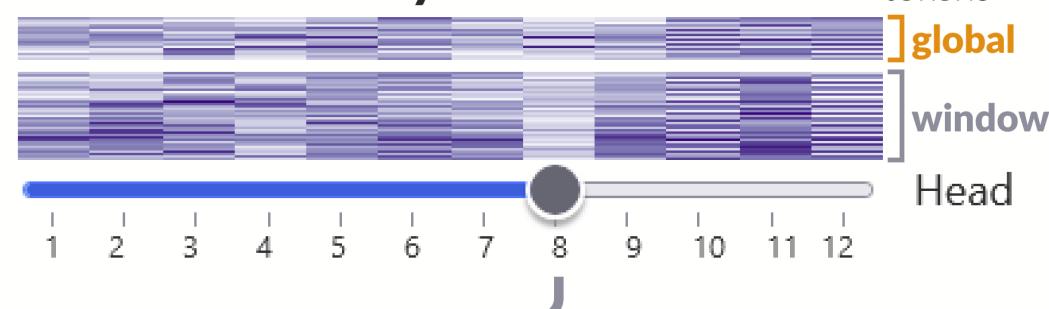
Heads compute weights called **attention** between all pairs of tokens for the next layer. Longformers use a unique **sliding attention window** to compute weights between tokens only inside the window. **Global** tokens are assigned attention between all other tokens, as usual.



Head 8 at each layer



Each head at **Layer 1**



- **Drill down :** Select heat map at single head + layer

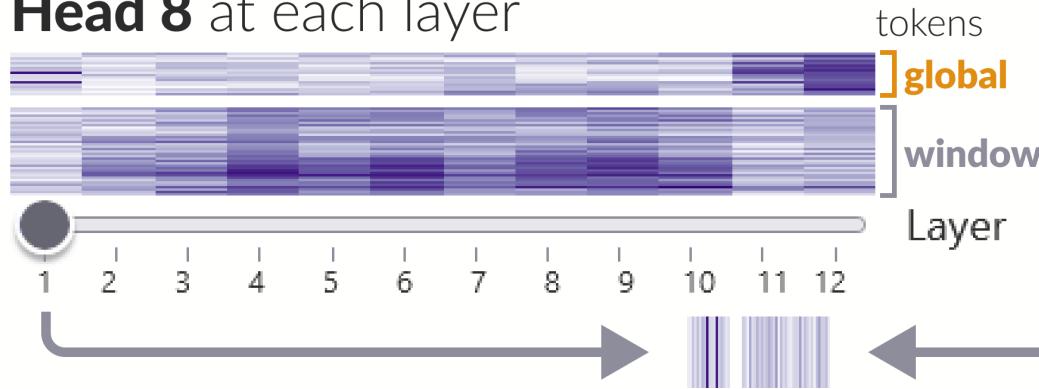


- **Drill down : Rug plot** of token attentions at single head + layer

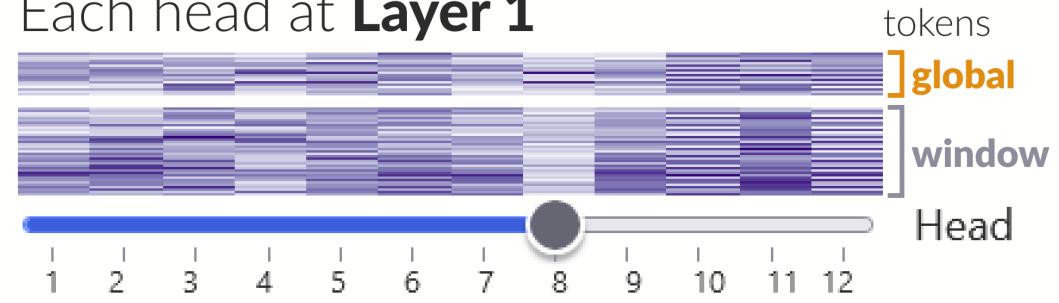


- **Drill down : Rug plot** of token attentions at single head + layer
 - Attention from **selected token** to all other **tokens**

Head 8 at each layer



Each head at **Layer 1**



Pairwise attention at **Layer 1 / Head 8** with the **selected token** at the slider position



Pairwise attention from **selected token** to **underlined** tokens



Weight (sequential)

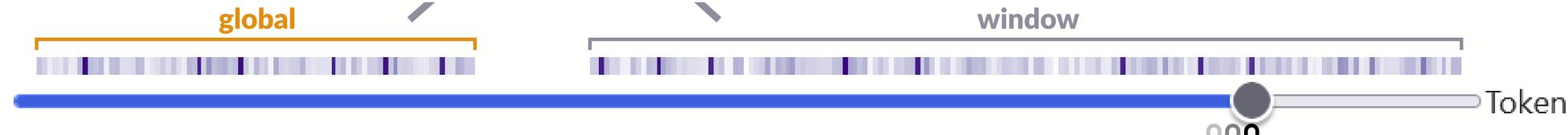
< s > Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view to obtain a vitamin is carcinogenic. Obtaining this vitamin has many safer ways. A doctor can take a blood sample if you have too little vitamin



Attention weight from vitamin to people is 9.42e-5

- Details : **Underline** attention weights and **compare** tokens inline!

Pairwise attention at **Layer 1 / Head 8** with the **selected token** at the slider position



Pairwise attention from **selected token** to **underlined** tokens



Weight (sequential)

< s > Whether people should sunbathe to get vitamin D has a clear and unambiguous answer from a medical point of view to obtain a vitamin is carcinogenic. Obtaining this vitamin is many safer ways. A doctor can take a blood sample if you have too little vitamin D. A cod-liver oil or vitamin D supplement are solutions if too little of this important vitamin shows in your test. < /s > ♦ ♦ We are thus in a situation where people are recommended to use something that we know is carcinogenic to obtain a vitamin. The No attention is computed outside of the window obtain this vitamin. If you suspect that you have too little vitamin D, go to your doctor to have a blood sample taken. If the test shows that you have too little of this important vitamin, the solution ought to be cod-liver oil or a vitamin D supplement. From a medical point of view we have a clear and unambiguous answer to the question of whether we should sunbathe to obtain enough vitamin D. We have safe ways to measure people's level of vitamin D, and we have safe ways to treat vitamin D deficiency if necessary.

Attention weight from vitamin to people is 9.42e-5

Attention weight from vitamin to vitamin is 0.102

Case Study

Discovered **surprising** LLM behaviors

- Removing first sentence **dropped scores** by 90%
- LLMs **weight punctuation** high in final layers
- Visually seeing **lack of attention** in summary text

Insights led to **re-training** LLM

- Improved **model accuracy** in training by 3%!

Expert Interviews

“How did iScore enable you to...”

Understand LLMs?

Evaluate LLMs?

Trust LLMs?

Expert Interviews

Understanding LLMs

- How decisions are made from LLM parameters
- Whether LLMs use context around words
- “Seeing” what is going on inside models

*“There is always anxiety or tension around **whether an outcome is just a fluke**”*

*“While [topic sentences] could be important for writing summaries, **it may not be the way the model is operating**”*

Expert Interviews

Evaluating LLMs

- Closing the loop of model development
- Quickly profile the best-performing models
- Generate examples to add to LLM training data

*"I used Input Perturbations to see **whether adversarial attacks can trick models.**"*

*"I often train **multiple models** for a task and **want to compare them.** I like to switch between the models **on the fly.**"*

Expert Interviews

Trusting LLMs

- Several requirements
 - Reproducibility
 - Address biases
 - Show limitations
 - Include teachers

"The aspects of iScore that can test variations [Assignments Panel, Input Perturbations] allow us to demonstrate what happens with changes in summaries. This can broadly allow us to improve trust in these systems."

Implications for Design

Structure evaluation using visual hierarchies

- **Upload / Compare** multiple text/LLMs at once
- **Arrange** for comparison (group by topic, etc.)
- **Visualize** “ground truth” training data to compare

Scale interpretability methods to large inputs

- **Fluid interaction** b/w levels of aggregation
- **Inline visuals** reveal semantic/syntactic patterns

Future Work

Responsible and Ethical AI for Education

- Stereotypes / biases, Multilingual, Tools for Teachers

LLM Generalizability

- LLM-generated Summaries (GPT, LLaMa, Claude)

Mixed Methods LLM Evaluation

- Gradient Attribution (SHAP), Hypothesis Testing (χ^2)

iScore

Visual Analytics for
Interpreting How
Language Models
Automatically Score
Summaries

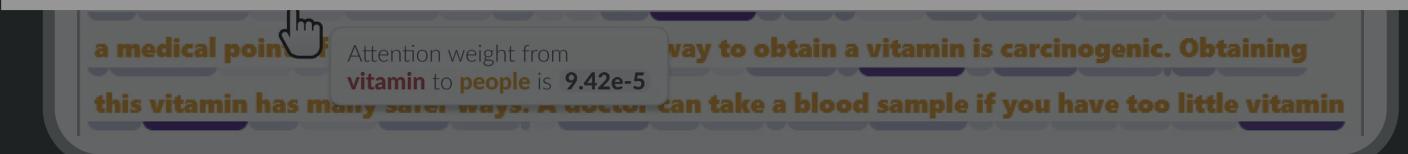
Adam **Coscia**, Langdon **Holmes**,
Wesley **Morris**, Joon Suh **Choi**,
Scott **Crossley**, Alex **Endert**

Whether people(mass; masses; the great unwashed; hoi polloi; multitude; ci
get(...) vitamin D(...) Replacing people with mass (...) answer(...) from a m

All code and models are **open source!**



bit.ly/iscore-paper



Visual Analytics Lab
@ Georgia Tech

