iScore helps learning engineers **evaluate LLMs**

that can **automatically grade summaries** written by **students** as they read

1 - **Edit** text and **compare** LLM scores

Underline edited text and **color** by LLM score **difference**

LIMITATION – some edits reveal unwanted **behaviors**, impacting **trust in the model!**

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

True difference (diverging)

Whether people(mass; masses; the great unwashed; hoi polloi; multitude; citizenry) should sunbathe(...) to get(...) vitamin D(...) (...) answer(...) from a medical(...) point(...) of view(...) . The recommended way (...) carcinogenic . Obtaining(incur; prevail; hold; obtain; receive; get; find) this vitamin has many safer(...) ways(...) . A doctor(...) if you (...) cod-liver oil(...) or vitamin D(...) (...) if too little(...) (...) in your test(...) .

Replacing **people** with **mass** **increases** the score by **0.138**

Click on struck-out word to reveal replacements **on demand**

Replacing **Obtaining** with **receive** **decreases** the score by **-0.169**

BIAS – removing the first sentence of a summary can **drop model scores** up to 90%!

Sentences - Remove sentences using sentence tokenizer

Absolute difference (sequential)

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 (...) aining this vitamin has many safer w ays. A doctor can take a blood sample if (...) cod-liver oil or vitamin D supplement are solutions if too little of this important vitamin shows in your test.

Removing this **sentence** **decreases** the score by **-0.504**

2 - **Analyze** internal LLM weights

BIAS – LLMs unexpectedly **weight punctuation** as too important internally!

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

global window Token

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** (...) **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.**</s>��We are thus in a situation where people are recommended to use something that we know is carcinogenic to obtain a vitamin. The (...) obtain this vitamin. If you suspect that you have too little vitamin D, go to your doctor (...) 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 people should sunbathe (...) 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.</s>

Attention weight from **vitamin** to **people** is **9.42e-5**

No attention is computed outside of the **window**

Attention weight from **vitamin** to **vitamin** is **0.102**

✅ **iScore** gives experts **interactive visualizations** that help reveal **LLM biases and limitations!**

We deployed **LLMs (Large Language Models)** in digital textbooks to **automatically score summaries** written by students, helping learners **build knowledge** as they **read!**

Hugging Face × RoBERTa

All code and models are open source!

bit.ly/iscore-paper

acoscia6@gatech.edu

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