## iScore helps learning engineers evaluate LLMs

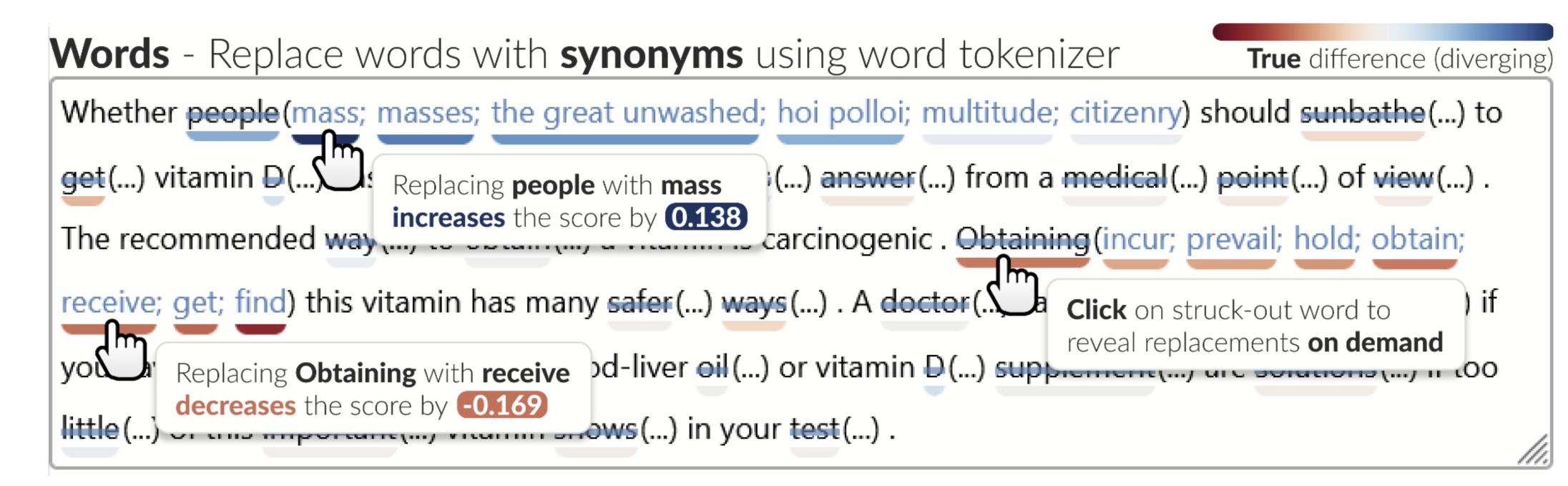
that can automatically grade summaries written by students as they read

## 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!





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

<b>Sentences</b> - Remove sentences u	ising sentence tokenize	r Absolute difference (sequential)
Whether people should sunbathe to get vi		biguous answer from a medical point
of view. The recommended way to ob	Removing this <b>sentence</b>	ining this vitamin has many safer w
ays. A doctor can take a blood sample if yo	decreases the score by -0.504	cod-liver oil or vitamin D supplement
are solutions if too little of this important	vitamin shows in your test.	

## 2 - Analyze internal LLM weights



BIAS – LLMs unexpectedly weight punctuation as too important internally!

global	window					
						T
					Ullym	
Pairwise attention	from <b>selected</b> t	oken to u	nderlined to	okens		<b>Weight</b> (seque
<s>Whether people</s>	should sunbathe to	get vitamin	D has a clear	and unar	nbiguou	s answer fro
	Attention weight from		obtain a vitar	nin is car	inogeni	c. Obtaining
this vitamin has ma	vitamin to people is 9.4	ze-5 zw. can tak	e a blood sam	ple if you	have to	o little vitan
D. A cod-liver oil or	vitamin D supplem	ent are solu	tions if too lit	tle of this	import	ant vitamin
shows in your test. <	:/s>��We are thus	in a situation	where people a	are recomn	nended to	use somethi
that we know is carcin	ogenic to obtain a vit	amin. The			obtain th	is vitamin. If y
suspect that you have	too little vitamin D, g	o to your doc	outside of the	window	re taken.	If the test sho
that you have too little	· · · · · · · · · · · · · · · · · · ·			oe cod-live	r oil or a	vitamin D
supplement. From a m	nedical point of view				a clear ar	nd unambigue
answer to the question	n of whether people s	vitamin to v shourd surroun	vitamin is 0.102	ugh vitar	nin D. We	have safe wa
to measure people��	s level of vitamin D,	and we have s	afe ways to tre	at <b>vitami</b> i	n D defici	ency if necess
There is thus no reaso	n to try to guess how	much UV rad	iation one need	ds to stay l	nealthy. T	his is neither v



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!





All code and models are open source!







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