

CAM 2.0: End-to-End Open Domain Comparative Question Answering System

**Ahmad Shallouf^{1*}, Hanna Herasimchyk^{1*}, Mikhail Salnikov^{2,3*},
Rudy Garrido Veliz^{1*}, Natia Mestvirishvili^{1*}, Alexander Panchenko^{2,3},
Chris Biemann¹, Irina Nikishina¹**

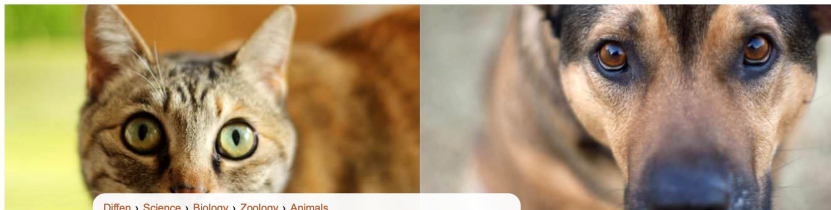
Comparative Question Answering

You
What is better: BMW or Mercedes?

ChatGPT
Determining whether BMW or Mercedes is better depends on individual preferences, needs, and priorities. Both brands offer a range of vehicles with different features, styles, and performance characteristics. Some people prefer BMW for its sporty driving dynamics and performance-oriented focus, while others prefer Mercedes for its luxury, comfort, and advanced technology. Ultimately, it's essential to test drive vehicles from both brands and consider factors such as design, driving experience, features, reliability, and cost to determine which one suits you best.



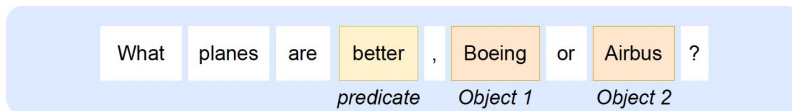
Cat vs. Dog



Diffen > Science > Biology > Zoology > Animals

Cats and **dogs** are the most popular **pets** in the world. Cats are more independent and are generally cheaper and less demanding pets. Dogs are loyal and obedient but require more attention and exercise, including regular walks.

Question:



Retrieved arguments

1. In this respect, the Boeing 777 is a better plane than the Airbus A300.
2. With the Boeing plane lighter than the Airbus, the MAX version of the LEAP is much smaller than the neo version - 69 inches vs. 78 inches in diameter.
3. The actual aim of Boeing is also important, they intend to make a better plane than Airbus A380, excluding that 800 pax.
4. ...
11. So Boeing can't make a better plane than Airbus, but Airbus can always make a better plane than Boeing.
12. Airbus, the European plane manufacturer, has orders for about 200 A380s, which can carry more passengers with greater fuel efficiency than other large planes such as Boeing 747s.
13. Christophe Menard, aerospace and defence analyst at Kepler Capital Markets in Paris, said that despite its own delays on the A350, Airbus was getting the plane out faster than Boeing managed with the Dreamliner.
- ...

Answer:

When comparing Boeing and Airbus planes, key factors influence the preference for one over the other. Boeing's emphasis on lighter and more fuel-efficient aircraft, technological superiority, and better cargo systems sets it apart from Airbus. Moreover, the strategic positioning of Boeing in the market and the ongoing advancements in engine efficiency and aerodynamics present a strong case for Boeing as the superior option.

The Objectives

1. To present an end-to-end system for answering comparative questions
2. To compare different approaches for each subtask
3. To introduce new benchmark unifying existing datasets for Comparative Question Answering (CompQA) with a public leaderboard

CompUGE Benchmark

Corpus	Task	Train	Dev	Test	Metrics
Bondarenko et al. (2020a)	Comparative Question Identification	10500	1350	3150	F1
Bondarenko et al. (2022a)		21869	2812	6561	
Sen et al. (2022)		14000	2000	4000	
Beloucif et al. (2022)		-	-	795	
Bondarenko et al. (2022a)	Object and Aspect labelling	2471	318	741	F1
Beloucif et al. (2022)		2141	275	642	
Chekalina et al. (2021)		2334	283	360	
Bondarenko et al. (2022a)	Stance classification	669	87	200	F1
Panchenko et al. (2019)		5183	576	1440	
Chekalina et al. (2021)	Summary generation	-	-	51	ROUGE, BERT-score

Table 1: The tasks included in CompUGE.

Related work

Enter a comparative question

Which is better for Deep Learning: Python or MATLAB?

Model to generate with
CTRL

Answer

Which is better **PRED** for Deep **ASP** Learning: Python **OBJ** or MATLAB **OBJ** ?

We chose Python **OBJ** because it was easy **PRED** to learn **ASP** and fun to work with. It has a large community of developers who are always willing to help each other out. In addition, the standard library makes developing applications with Python **OBJ** very easy. The downside is, that you have to be careful when using it. If you're not careful, you'll end up writing code which will crash your computer if something goes wrong. You also need to know how to use libraries like numpy in order to get good results.

Figure 2: The interface of the Comparative Question Answering System (CoQAS).

Schildwächter, M., Bondarenko, A., Zenker, J., Hagen, M., Biemann, C., Panchenko, A. (2019): [Answering Comparative Questions: Better than Ten-Blue-Links?](#) Proceedings of ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Glasgow, Scotland, UK, pp. 361-365 ([demo link](#))

Viktoriia Chekalina, Alexander Bondarenko, Chris Biemann, Meriem Beloucif, Varvara Logacheva, and Alexander Panchenko. 2021. [Which is Better for Deep Learning: Python or MATLAB? Answering Comparative Questions in Natural Language](#). In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 302–311, Online. Association for Computational Linguistics.

Comparative Argumentative Machine

CAM Search About GitHub API Contact

First object: cuda

Second object: opencl

versus

Aspect: e.g. price

Aspect importance:

Default Compare! Reset Faster Search

cuda (61.66%)

Generated Aspects for cuda

easier for the developer faster thats easier to learn nature easier to use higher greater ofcourse easier to program

Generated Aspects for opencl

easier to program for simpler earlier older performance simpler to target cuda

CUDA as for superior, more capable, and has better tools than OpenCL ¹

That's why it will always be slower than CUDA.

CUDA simply offers better performance than OpenCL ¹

Does CUDA a better job than OpenCL.

OpenCL seems to be slower than either C++ AMP or CUDA.

CUDA is supposed to be faster than OpenCL and more stable.

CUDA is much easier to learn than OpenCL, you're much more likely to run into codes that use CUDA over OpenCL.

and there are better support libraries for CUDA (you'll probably be interested in cuBLAS).

As for CUDA - It is almost directly inferior to OpenCL.

BTW: I've never said that OpenCL will be faster than CUDA. I've said that AMD's perform better in OpenCL than the nVidia cards. ²

ROFL @ OpenCL doing ANYTHING better than cuda ²

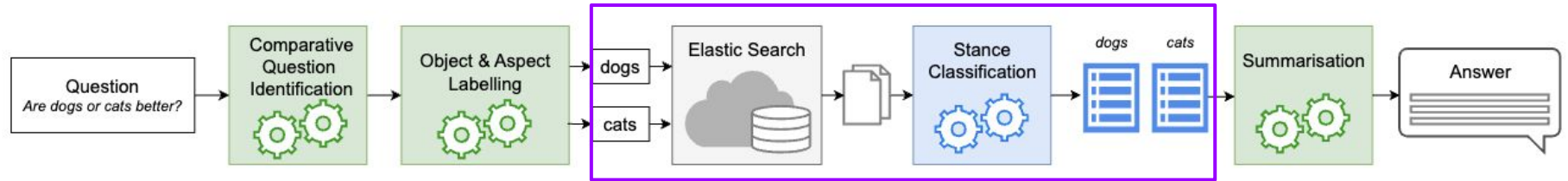
OpenCL is doing better than CUDA did at launch.

AMD's openCL is easier to program for than CUDA as well. ¹

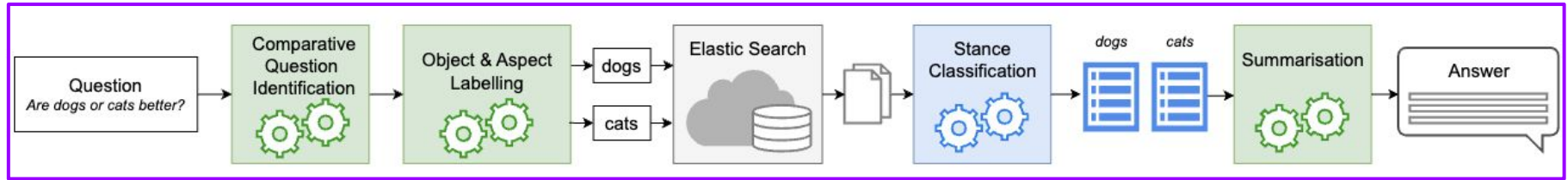
Generating libNVVM makes it simpler to target CUDA than the earlier option of generating PTX.

IMO OpenCL, by virtue of it's compatibility, is better than CUDA ²

CAM 1.0



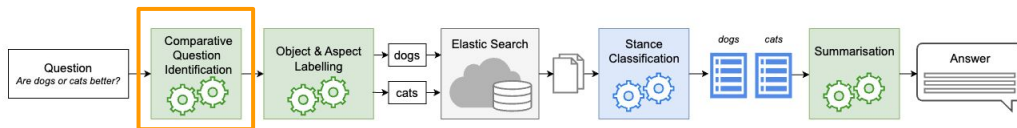
CAM 2.0



CAM 2.0: Comparative Question Identification

Model	Precision	Recall	F1	Params
Bondarenko et al. (2022a) (albert/albert-large-v1)	0.9250 ± 0.0104	0.9116 ± 0.0090	0.9179 ± 0.0006	17M
distilbert/distilbert-base-uncased-finetuned-sst-2-english	0.9244 ± 0.0113	0.9131 ± 0.0094	0.9186 ± 0.0051	67M
prajjwal1/bert-tiny	0.9235 ± 0.0099	0.8759 ± 0.0027	0.8990 ± 0.0049	4M
meta-llama/Llama-2-7b-chat-hf (90-shot)	0.8592 ± 0.0210	0.2774 ± 0.1077	0.3065 ± 0.1461	7B
lmsys/vicuna-7b-v1.5 (90-shot)	0.8917 ± 0.0108	0.9024 ± 0.0120	0.8824 ± 0.0037	7B
meta-llama/Llama-2-7b-chat-hf (LoRA)	0.9762 ± 0.0040	0.9762 ± 0.0041	0.9762 ± 0.0041	7B+33M
lmsys/vicuna-7b-v1.5 (LoRA)	0.9777 ± 0.0001	0.9777 ± 0.0002	0.9777 ± 0.0001	7B+33M

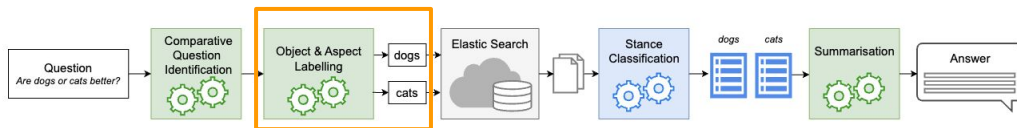
Table 2: Results for Comparative Question Identification on the Webis-2022 dataset ([Bondarenko et al., 2022a](#)). Other models and results are found in Table 8 in the Appendix.



CAM 2.0: Object and Aspect Labeling

Model	F1-OBJ	F1-ASP	F1-PRED	F1-Mean	Params
Bondarenko et al. (2022a) (FacebookAI/roberta-large)	0.7946 \pm 0.0073	0.6433 \pm 0.0031	0.9406 \pm 0.0019	0.8249 \pm 0.0037	355M
FacebookAI/roberta-base	0.7696 \pm 0.0052	0.6078 \pm 0.0121	0.9446 \pm 0.0006	0.8078 \pm 0.0041	125M
microsoft/deberta-v3-base	0.7998 \pm 0.0061	0.6808 \pm 0.0029	0.9524 \pm 0.0042	0.8370 \pm 0.0049	184M
microsoft/deberta-v3-large	0.8290\pm0.0077	0.6809\pm0.0018	0.9604\pm0.0009	0.8545\pm0.0032	434M
google-bert/bert-base-uncased	0.7337 \pm 0.0058	0.5851 \pm 0.0075	0.9348 \pm 0.0079	0.7832 \pm 0.0025	109M
meta-llama/Llama-2-7b-chat-hf (3-shot)	0.2008 \pm 0.0641	0.0000 \pm 0.0000	0.6364 \pm 0.0120	0.3655 \pm 0.0320	7B
lmsys/vicuna-7b-v1.5 (3-shot)	0.0880 \pm 0.0705	0.0345 \pm 0.02753	0.3402 \pm 0.1366	0.1763 \pm 0.0712	7B
meta-llama/Llama-2-7b-chat-hf (generative setup)	0.4903 \pm 0.1084	0.3843 \pm 0.0615	0.7493 \pm 0.1081	0.5413 \pm 0.1533	7B
lmsys/vicuna-7b-v1.5 (generative setup)	0.4684 \pm 0.0815	0.3225 \pm 0.0802	0.6268 \pm 0.1519	0.4725 \pm 0.1243	7B

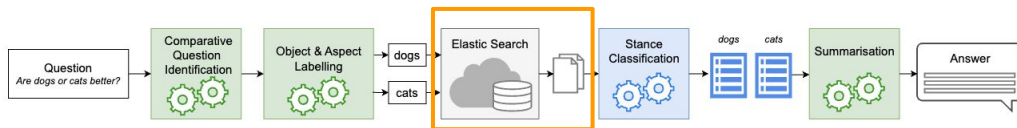
Table 4: Results for Object and Aspect Labeling on the Webis-2022 dataset ([Bondarenko et al., 2022a](#)). Other models and their results can be found in Table 9 in the Appendix. The used hyperparameters are mentioned in Table 11 in the Appendix.



CAM 2.0: Sentence Retrieving



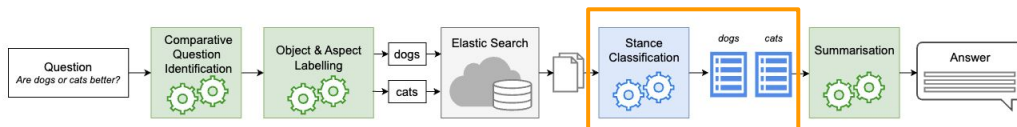
Common Crawl



CAM 2.0: Stance Classification

Model	F1-BETTER	F1-WORSE	F1-NONE	F1-Mean	Params
Panchenko et al. (2019)	0.75	0.43	0.92	0.85	UNK
Ma et al. (2020)	0.7821	0.5872	0.9298	0.8743	UNK
google-bert/bert-base-uncased	0.8999 \pm 0.0078	0.7426 \pm 0.0254	0.9636 \pm 0.0038	0.8807 \pm 0.0088	109M
microsoft/deberta-v3-large	0.9172\pm0.0157	0.8303\pm0.0440	0.9744\pm0.0036	0.9106\pm0.0065	434M
meta-llama/Llama-2-7b-chat-hf (30-shot)	0.3636 \pm 0.0100	0.2170 \pm 0.0482	0.4406 \pm 0.0600	0.4075 \pm 0.0428	7B
lmsys/vicuna-7b-v1.5 (20-shot)	0.3343 \pm 0.0546	0.7740 \pm 0.0677	0.2097 \pm 0.0465	0.6428 \pm 0.0403	7B
meta-llama/Llama-2-7b-chat-hf (LoRA)	0.8473 \pm 0.0055	0.7143 \pm 0.0021	0.9426 \pm 0.0056	0.9073 \pm 0.0056	7B+33M
lmsys/vicuna-7b-v1.5 (LoRA)	0.8597 \pm 0.0056	0.7140 \pm 0.0067	0.9216 \pm 0.0056	0.9043 \pm 0.0002	7B+33M

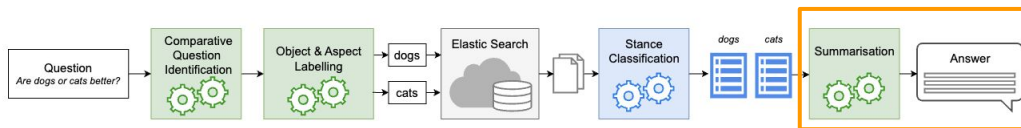
Table 5: Results for Stance Classification dataset ([Panchenko et al., 2019](#)). Other models and their results can be found in Table 10 in the Appendix. Hyperparameters are presented in Table 12 in the Appendix.



CAM 2.0: Summarization

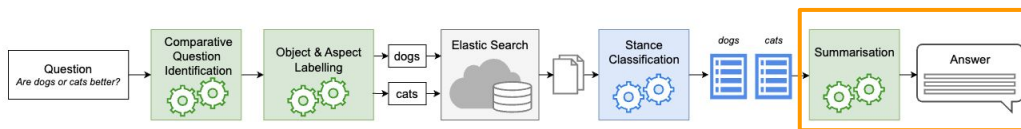
Model	ROUGE-1	ROUGE-2	BERT-Score	Params
CTRL Which-better-x-y-for-z (Chekalina et al., 2021)	0.2454	0.0200	0.8214	1.63B
CAM bullet points (Chekalina et al., 2021)	0.2298	0.0328	0.8201	-
facebook/bart-large-cnn	0.1855±0.0047	0.0161±0.0010	0.8270±0.0101	406M
sshleifer/distilbart-cnn-6-6	0.1947±0.0021	0.0171±0.0008	0.8260±0.0105	230M
meta-llama/Llama-2-7b-chat-hf (no args)	0.1623±0.0009	0.0227±0.0005	0.8070±0.0135	7B
lmsys/vicuna-7b-v1.5 (no args)	0.1930±0.0028	0.0196±0.0010	0.8136±0.0143	7B
meta-llama/Llama-2-7b-chat-hf (2-shot)	0.1778±0.0050	0.0168±0.0021	0.8048±0.0127	7B
lmsys/vicuna-7b-v1.5 (2-shot)	0.1857±0.0026	0.0204±0.0015	0.8065±0.0164	7B
gpt-3.5-turbo (no args)	0.1658±0.0015	0.0200±0.0007	0.8125±0.0112	154B
gpt-3.5-turbo (2-shot)	0.1998±0.0023	0.0210±0.0009	0.8125±0.0101	154B

Table 6: Results for Summarization on Yahoo!Answer dataset (Chekalina et al., 2021). The results are compared against the answer marked as the "Best Answer" on the platform.



CAM 2.0: Summarization

model name	summary
Yahoo.Best Answer	How can you even ask this question yet? Only the Xbox 360 is out at the moment and that hasn't even been tested by gamers enough to see truly how good or rubbish it is. You need to ask the question again when all three systems are out!
CTRL:does-x-better-y-for-z	I think that the answer to this question depends on what you mean by "better". For me, it comes down to two things: 1) how much of a difference does it make in terms of performance and 2) whether or not it's worth paying more for it. If you're looking at buying an Xbox One, then yes, I'd say go ahead and get one. But if you're just thinking about getting one, then no, don't bother. The only reason why I'm even considering one now is because my 360 has been acting up lately (again, probably due to overheating). And while I've had no problems with it, there are definitely things that could be improved upon. So far, though, there's nothing really compelling me to buy another Xbox. In fact, after playing around with some other consoles over the last few days, I can't see myself ever wanting to. Maybe I'll wait until they have something new out before making any decisions. What do you think? Are you planning on picking up an Xbox One? Let us know your thoughts in the comments.
facebook/bart-large-cnn	Microsoft has a much better (easier to use) SDK for games than Sony does. Microsoft made these updates go much faster, smoother, and less frequent than Sony ever did with the PlayStation 3. Microsoft had better games at launch than Sony and will also in 2014.
lmsys/vicuna-7b-v1.5 (2-shot)	<p>Microsoft and Sony are two major companies in the technology industry, with a significant presence in the gaming market.</p> <p>Some argue that Microsoft is better and faster than Sony, with updates going smoother and less frequent [2]. Microsoft is also considered to have a better SDK for games [3], and their conference was thought to have better pacing [4]. Additionally, some believe that Microsoft has a superior position over Sony in terms of software tools [9]. However, others argue that Sony is a superior hardware manufacturer, much better than both Nintendo and Microsoft [11]. The PS4 is physically superior to the Xbox One, with better design [12]. Sony is also believed to be working harder for gamers than Microsoft and Nintendo [10].</p> <p>In terms of gaming, some argue that Microsoft is inferior to Sony [13], and that Sony is superior to Microsoft in every way possible [17]. Sony was also considered to be a bit smarter than Microsoft in terms of their approach to gaming [18].</p> <p>Ultimately, the preference between Microsoft and Sony comes down to personal opinions and experiences. Some may prefer Microsoft for its software tools and updates, while others may prefer Sony for its hardware design and gaming experience.</p> <p>Arguments used: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20</p>



Evaluation, User Study & Analysis

- **50 questions** with the objects labelled from the Touché at CLEF competition in 2022 were asked to our system.
 - All the questions were successfully classified as comparative and the **F1 score** for object labelling was **1.0**
- In the user study, we selected **28 questions** with at least **5** arguments and obtained the summaries from the system. **Four** annotators were asked to read and evaluate the quality of them.
 - The annotators agreed on the same 22 summaries (**78.6%**) being **helpful** and the same 18 summaries (**64.3%**) being **fluent**.

Was this summary:

☐

Useful

☐

Fluent

Submit

CAM 2.0: System Demonstration

The screenshot displays the CAM 2.0 web interface. At the top is a dark blue header with the text 'CAM 2.0' on the left and navigation links 'QA', 'About', 'GitHub', 'API', 'Contact', and a menu icon on the right. Below the header is a large text input field containing the question 'What is better: Harry Potter or LotR?'. Underneath the input field is a blue 'Ask' button and a checkbox labeled 'Faster Answer'. A feedback message box below states 'Your question is comparative!' with a blue button that says 'No. It is not.'. The main interface area features two input fields for objects, 'Harry Potter' and 'LotR', separated by the word 'versus'. Below the 'Harry Potter' field is another input field with the word 'better' and the text 'e.g. price' below it. In the center, there are two circular buttons, one with a minus sign '-' and one with a plus sign '+'. At the bottom left, there is a label 'Nr. of arguments per object' next to an input field containing the number '10'. Below this are two blue buttons: 'Send improved Objects & Aspects' and 'Compare!'. On the far right of the bottom section is a checkbox labeled 'Faster Search'.

<https://cam-v2.ltdemos.informatik.uni-hamburg.de>

Figure 4: CAM 2.0 input form for natural language questions.

CAM 2.0: System Demonstration

Harry Potter (43.89%)

(56.11%) LotR

Summary

"Harry Potter and LotR are both popular fantasy novels, but they have some differences and similarities. Some argue that Harry Potter is better than LotR, while others prefer LotR movies over the Harry Potter movies. Some people even argue that the movie did a much better adaption of the book than LOTR did. However, there are also those who prefer the movies for certain reasons. On the other hand, some people argue that LotR is a good story, that the LotR trilogy was adapted 1000x better into film than any of the books were, and that the LOTR DVD is much better. Some also prefer the HP books more than the books, and argue that it's simpler to read than LotR and that there's no need to spend hours on end on a quest just so you can find the special features on the DVD. Others argue that "Harry Potter" will likely do better than "LOTR" but only because of the reasons mentioned above."

Was this summary:

☐ Useful

☐ Fluent

Submit

Arguments for Harry Potter

[[armchair](#)] I like Harry Potter better than LOTR anyway.

[[braingle](#)] I think Harry Potter is better than LotR, but I've only seen the movies for LotR.

[[aintitcool](#)] But Harry Potter will end better than LOTR because that is a fact.

[[archives](#)] People magazine seems to think the Harry Potter trailer is better than the LOTR trailer.

[[news](#)] After seeing Harry Potter, I thought that the movie did a MUCH better adaption of the book than LOTR did.

[[always](#)] I honestly think I enjoy Harry Potter more sometimes, but this is because I am stupid and can't understand it better than LotR.

[[sitepoint](#)] As for the box office...well, I think that Harry Potter will likely do better than LotR but only because of the reasons mentioned above.

[[patricklogan](#)] Arguably, Harry Potter is far simpler than LotR.

Arguments for LotR

[[neoseeker](#)] Please note that I don't consider Harry Potter to be superior to LotR, or even comparable.

[[india-forums](#)] On a side note, I like the LOTR movies much better than the Harry Potter movies.

[[movieweb](#)] I'd even say that "LOTR" is better than "Harry Potter" (I still love those movies though).

[[lucasforums](#)] I think the LotR books are better than the Harry Potter books but I enjoy reading the HP books more.

[[reddit](#)] And you must admit that the LotR trilogy was adapted 1000x better into film than any of the Harry Potter books were.

[[news](#)] The LOTR DVD is much better than the Harry Potter one because you don't have to spend hours on end on a quest just so you can find the special features.

[[neoseeker](#)] I'm not fighting against the idea that LotR is a good story; I'm fighting against the idea that it's so plain and obvious that it's superior to Harry Potter.

Figure 3: CAM 2.0 output for the question "What is better: Harry Potter or LotR?".

Conclusion

- We present CAM 2.0 and CompUGE benchmark.
- We compare several Encoders and Generative Transformers
- Medium-sized Encoders deliver strong performance.
- Generative Transformers effectively aggregate arguments in zero-shot setup