Improve Wikipedia Verifiability with AI

Fabio Petroni

Bio

- Researcher, Engineer and Manager in FAIR for the past 4 years
- Co-Founder of

Knowledge-intensive NLP: tasks that requires – even for humans – access to a large body of information.

Collaborators


Ledell Wu


Kashyap Popat


Naman Goyal


Nicola Cancedda


Mikel Artetxe


Luke Zettlemoyer


Mikhail Plekhanov


Michele Bevilacqua


Nicola De Cao


Sebastian Riedel


Yacine Jernite


Vladimir Karpukhin


Jean Maillard


Vassilis Plachouras


Tim Rocktäschel


Scott Wen-tau Yih


Angela Fan


Samuel Broscheit


Edouard Grave


Lucas Hosseini


Gautier Izacard

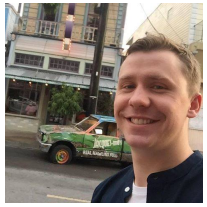

Patrick Lewis


Pierre-Emmanuel Mazare


Maria Lomeli


Majid Yazdani


Giuseppe Ottaviano


James Thorne

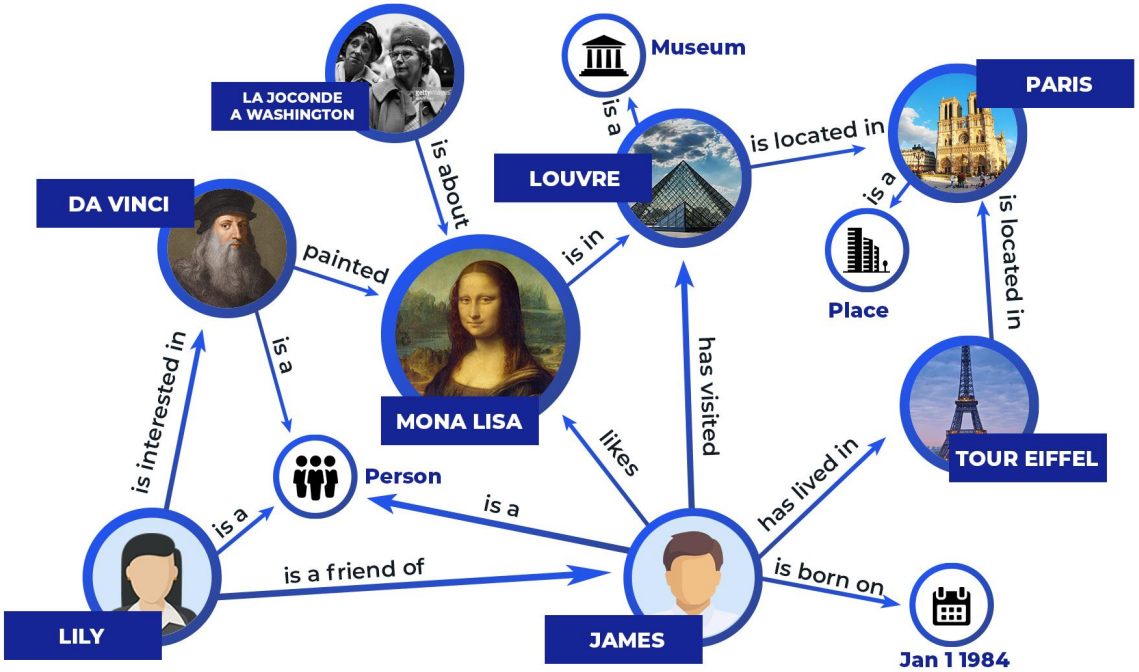

Ola Piktus


Armand Joulin


Timo Schick

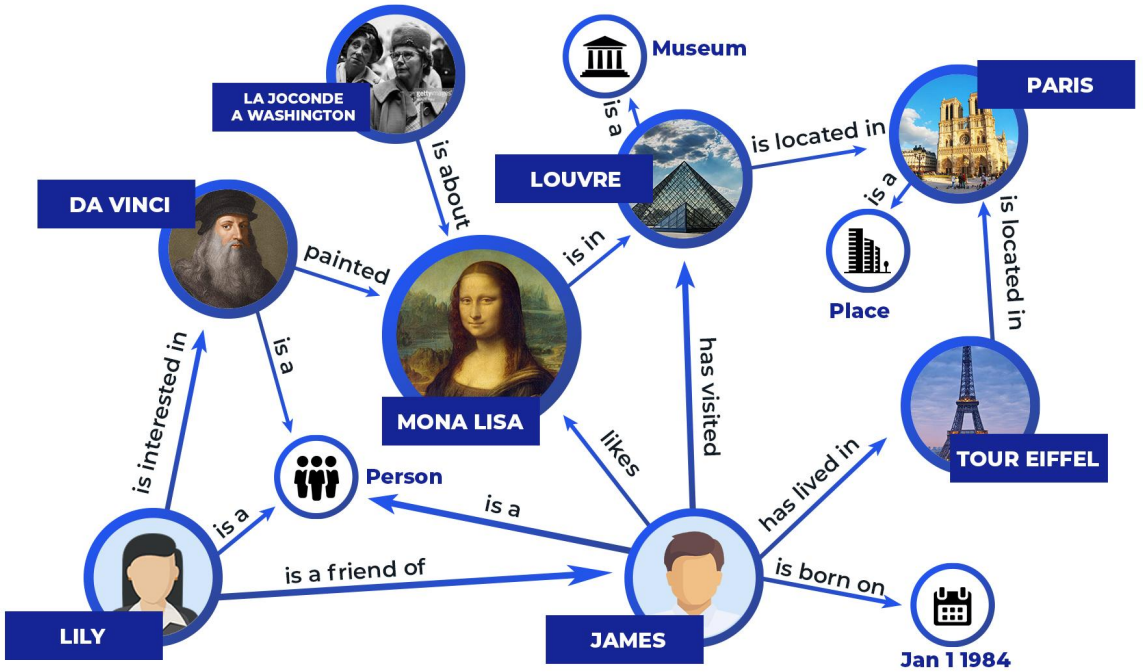

Jane Yu

For decades, AI researchers have searched for a representation of knowledge that is most useful for machines



Limitations of knowledge graphs

- Human supervision
- Schema engineering
- Predefined class of relations
- Difficult to extend to more data
- Incomplete



Other approaches

Texts as Knowledge Bases

Christopher Manning

Joint work with Gabor Angeli and Danqi Chen

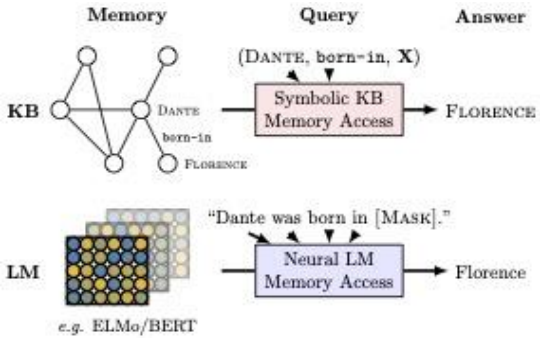
Stanford NLP Group

@chrmanning · @stanfordnlp

AKBC 2016

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com



kiltbenchmark.com

5 task families
11 datasets
1 knowledge source
3.5M datapoins

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, Sebastian Riedel: **Kilt: a benchmark for knowledge intensive language tasks.** NAACL-HLT 2021

Leaderboard

The KILT leaderboard.

Phase: **Open Domain QA - Natural Questions**, Split: **test**

Order by metric

B - Baseline * - Private V - Verified

Rank	Participant team	R-Prec (↑)	Recall@5 (↑)	EM (↑)	F1 (↑)	KILT-EM (↑)	KILT-F1 (↑)	Last submission at
11	Atlas (Atlas)	0.00	0.00	61.29	70.70	0.00	0.00	3 months ago
12	Google Research & TU Wien & UMass (FiD with RS)	0.00	0.00	61.15	70.56	0.00	0.00	5 months ago
1	FiD-Light	75.55	75.02	58.38	67.33	51.11	57.83	4 months ago
3	SEAL (intersect)	63.16	68.19	53.74	62.24	38.78	44.40	7 months ago
2	IBM Research AI - Re2G (Re2G)	70.78	76.63	51.73	60.97	43.56	49.80	1 year ago
5	KILT-WEB 2 (Wikipedia)	59.83	71.17	51.59	60.83	35.32	40.73	1 year ago

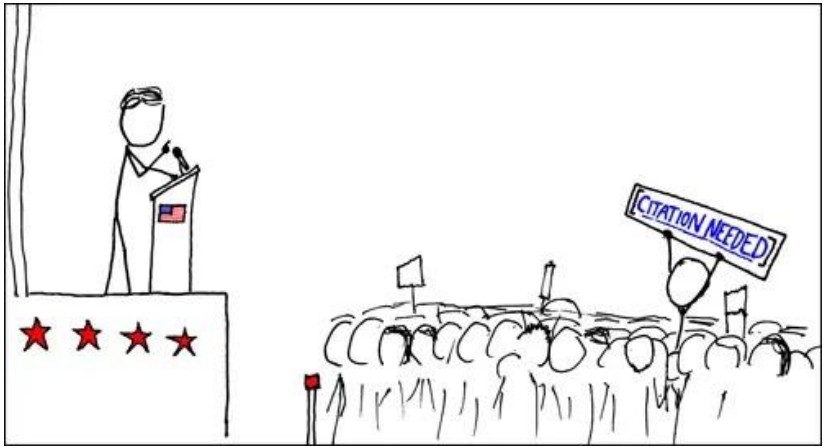
Representing knowledge

- We should represent knowledge to a machine exactly as we represent knowledge to a human (text, videos, images, charts, books, podcasts, etc)

- Rather than building knowledge bases for machine use, create better knowledge for humans! (**Machines can use it too!**)
  - Help make Wikipedia better
  - Help discovering new knowledge
  - etc.

Wikipedia Verifiability

- Verifiability is a core content policy of Wikipedia!
- Claims that are likely to be challenged need to be backed by citations.
- Finding relevant sources is a difficult task.
- Many Wikipedia claims do not have any references that support them.
- Even existing citations might not support a given claim or become obsolete.



Verify Wikipedia

Mirella Lapata

From Wikipedia, the free encyclopedia

Mirella Lapata FRSE is a computer scientist and Professor in the School of Informatics at the University of Edinburgh.[3] Working on the general problem of extracting semantic information from large bodies of text, Lapata develops computer algorithms and models in the field of natural language processing (NLP).[1]

Awards and honors

- In 2009 Lapata became the first recipient of the Microsoft British Computer Society (BCS)/BCS IRSG Karen Spärck Jones Aw information retrieval and natural language processing; the award commemorates the life and work of Karen Spärck Jones.
- In 2012 Lapata won an Empirical Methods in Natural Language Processing (EMNLP)-CoNLL 2012 Best Reviewer Award.[11]
- In 2016 Lapata, with Eneko Agirre and Sebastian Riedel, won the EMNLP Best Data Set Paper Award.[12]
- In 2018 Lapata was awarded, alongside Li Dong, an Association for Computational Linguistics (ACL) Best Paper Honorable M
- In 2019 Lapata was elected a Fellow of the Royal Society of Edinburgh[14]
- In 2020 Lapata was elected to the Academia Europaea.[15]

Verify Wikipedia

Best Paper Committee

Best Paper & Honorable Mention: Stephen Clark, Hal Daumé III, Chris Dyer, and Julia Hockenmaier
Best Short Paper: Stefan Riezler, Anoop Sarkar, and Noah Smith
Best Data Set Paper: Eneko Agirre, Mirella Lapata and Sebastian Riedel
Chair: Xavier Carreras and Kevin Duh

Verify Wikipedia Goals

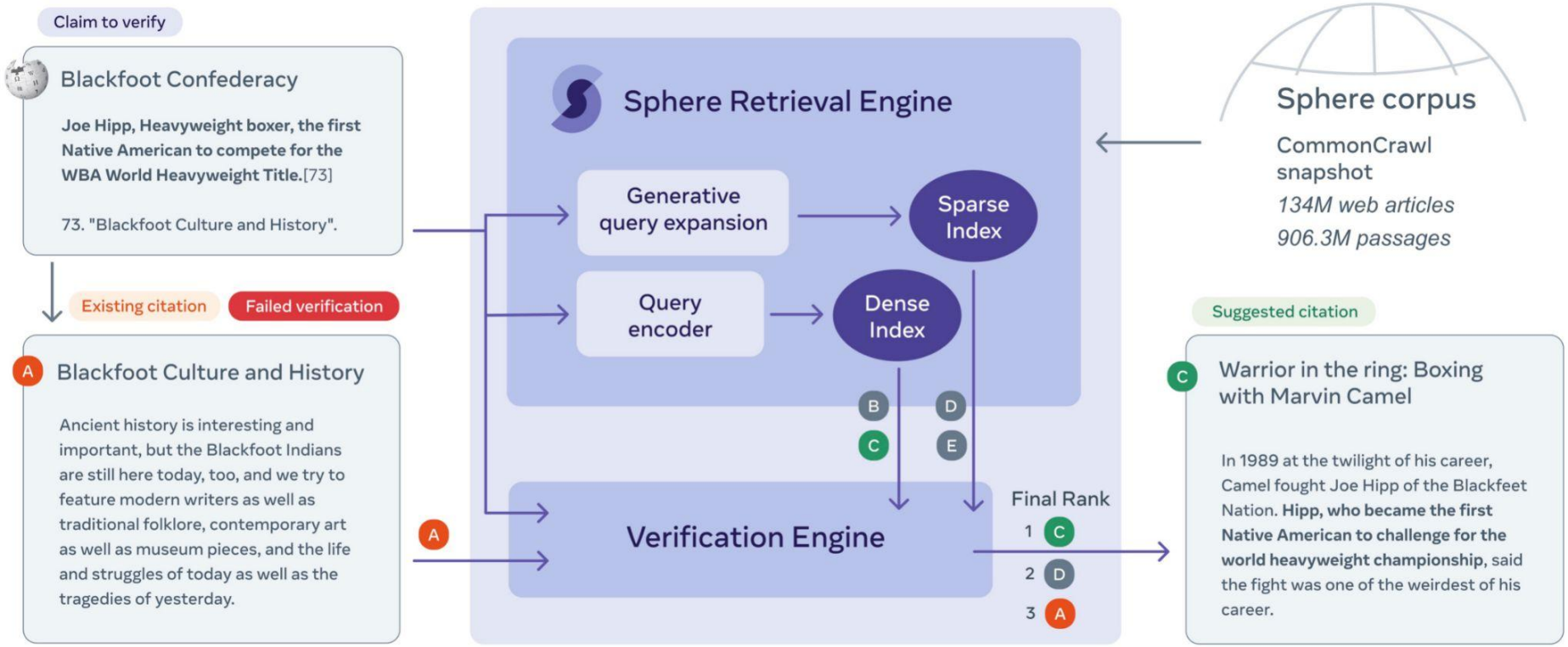
- Assist Wikipedia editors:

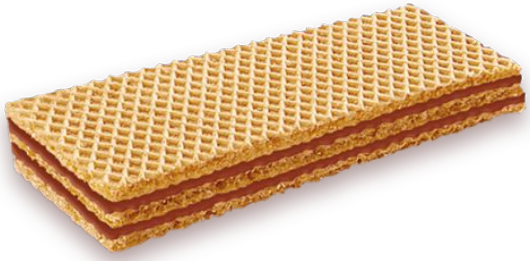
    1. surface existing citations that are likely to fail verification

    2. recommend citations for an unverified claim from the web

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, Sebastian Riedel - **Improving Wikipedia Verifiability with AI**

Verify Wikipedia Architecture





The WAFER dataset

search of beauty in whiteness. Since then, Morrison has experimented with lyric fantasy, as in her two best-known later works, *Song of Solomo*
awarded the Pulitzer Prize for Fiction; along
comparisons to Virginia Woolf,[44] and the N
tradition [of magical realism]."[45] *Beloved* wa
York Times as the most important work of fiction of the last 25 years.[46]

"What Is the Best Work of American Fiction of the Last 25 Years?". *The New York Times*. May 21, 2006. Retrieved December 4, 2009.

- Citations in Wikipedia aligned with Common Crawl
- Citation datapoints:
  - **4.5M** for training
  - **5K** for dev
  - **5K** for test
- The Sphere corpus
  - 134M web articles
  - 906M passages (100 words)

input

American literature [SEP] **Section::::Contemporary American literature.** [SEP] this genre, which won the National Book Award and was unanimously nominated for the Pulitzer Prize for Fiction that year. His other major works include his debut, "V." (1963), "The Crying of Lot 49" (1966), "Mason & Dixon" (1997), and "Against the Day" (2006).nToni Morrison, recipient of the Nobel Prize for Literature, writing in a distinctive lyrical prose style, published her controversial debut novel, "The Bluest Eye", to critical acclaim in 1970. Coming on the heels of the signing of the Civil Rights Act of 1965, the novel, widely studied in American schools, includes an elaborate description of incestuous rape and explores the conventions of beauty established by a historically racist society, painting a portrait of a self-immolating black family in search of beauty in whiteness. Since then, Morrison has experimented with lyric fantasy, as in her two best-known later works, "Song of Solomon" (1977) and "Beloved" (1987), for which she was awarded the Pulitzer Prize for Fiction; along these lines, critic Harold Bloom has drawn favorable comparisons to Virginia Woolf, and the Nobel committee to "Faulkner and to the Latin American tradition [of magical realism]." **"Beloved" was chosen in a 2006 survey conducted by "The New York Times" as the most important work of fiction of the last 25 years.[CIT] ...**

provenance

url	**https://www.nytimes.com/2006/05/21/books/fiction-25-years.html**
title	**"What Is the Best Work of American Fiction of the Last 25 Years? - The New York Times**
text	Books|What Is the Best Work of American Fiction of the Last 25 Years? What Is the Best Work of American Fiction of the Last 25 Years? By THE NEW YORK TIMES MAY 21, 2006 "Beloved," by Toni Morrison, center, was chosen as the best American fiction of the last 25 years. Runners-up were, from left: Philip Roth, Cormac McCarthy, John Updike and Don DeLillo. Credit From left: Sara Krulwich/The New York Times; Derek Shapton/The Associated Press; Sara Krulwich/The New York Times; The Associated Press; Don Hogan Charles/The New York Times THE FOLLOWING BOOKS ALSO RECEIVED MULTIPLE VOTES:

Sparse Retrieval with Generative Query Expansion

Monument Valley (video game)

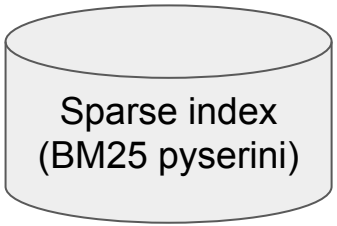
From Wikipedia, the free encyclopedia

Ken Wong left Ustwo Games soon after completing *Monument Valley* to create his own studio, Mountains, which created *Florence*.[46]

s2s

Monument Valley's Ken Wong is leaving Ustwo Games • Eurogamer.net

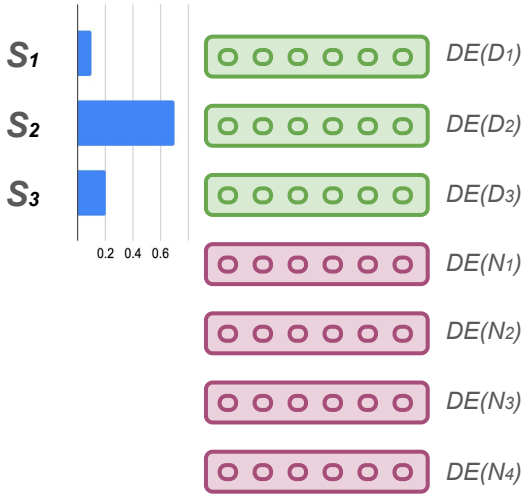
Query: 'Ken Wong left Ustwo Games soon after completing "Monument Valley" to create his own studio, Mountains, which created "Florence". Monument Valley's Ken Wong is leaving Ustwo Games • Eurogamer.net Monument Valley (video game)'



Dense Retrieval with soft EM

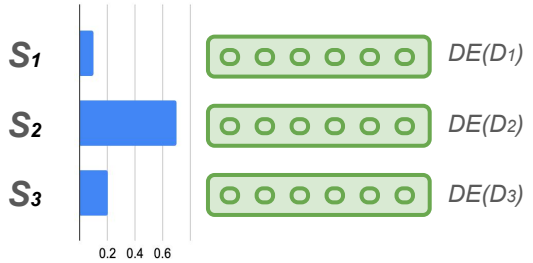
Challenge: Supervision on URL not on passage in URL

- Problem: True passage(s) are unknown.
- Contrastive loss assumes <u>one</u> true passage D_i
- Requires strategy to derive loss over $\{L(D_1), L(D_2), \ldots, L(D_n)\}$
- Solutions:
  - Uniform: $L = 1/n * L(D_1) + 1/n * L(D_2) + \ldots$
  - Soft EM: $L = S_1 * L(D_1) + S_2 * L(D_2) + \ldots$

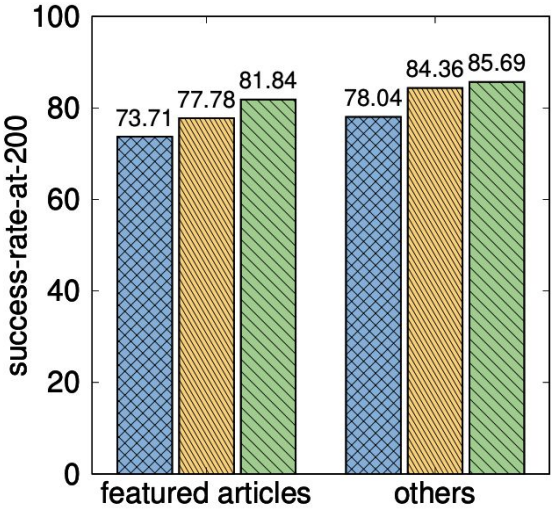


Verification Engine

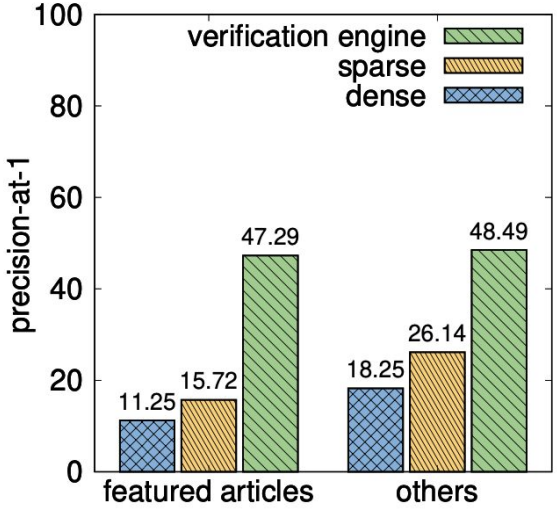
- BERT-based cross-encoder
- Rank claim-document pairs in order of verifiability.
- To train the model we use a training objective that rewards models when it ranks existing Wikipedia evidence higher than evidence returned by our retrieval engine.



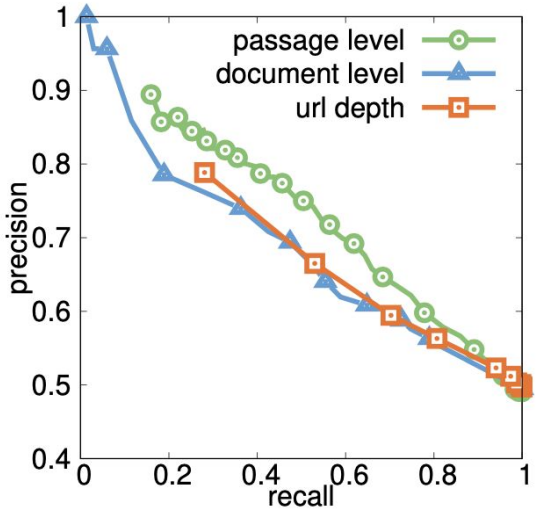
Results (Automatic)



(a) Percentage of times our retrievers can surface the gold source among the top-200 results, for citations in featured and other Wikipedia articles. The *verification engine* bar (*i.e.*, green) combines sparse and dense retrievers, 100 passages each.

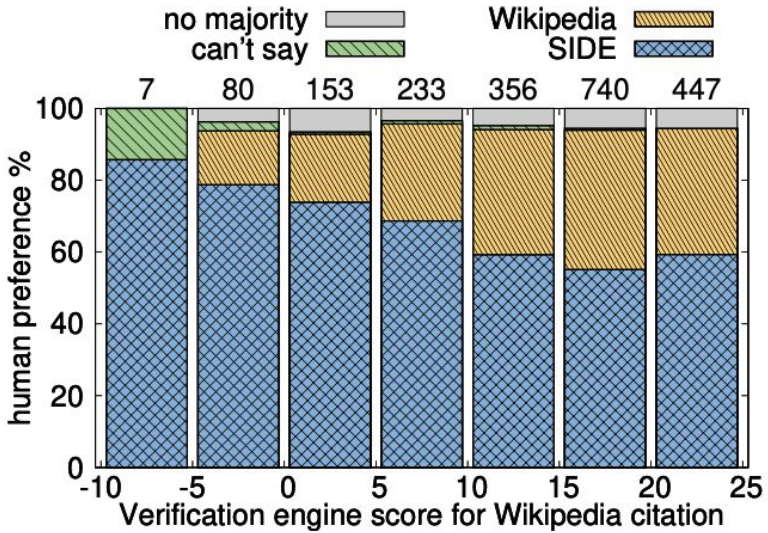


(b) Accuracy in surfacing the gold source in first position, for citations in featured and other Wikipedia articles. The *verification engine* (*i.e.*, green bar) takes in input a combination of 100 passages from the sparse and 100 from the dense retriever, and rerank those.

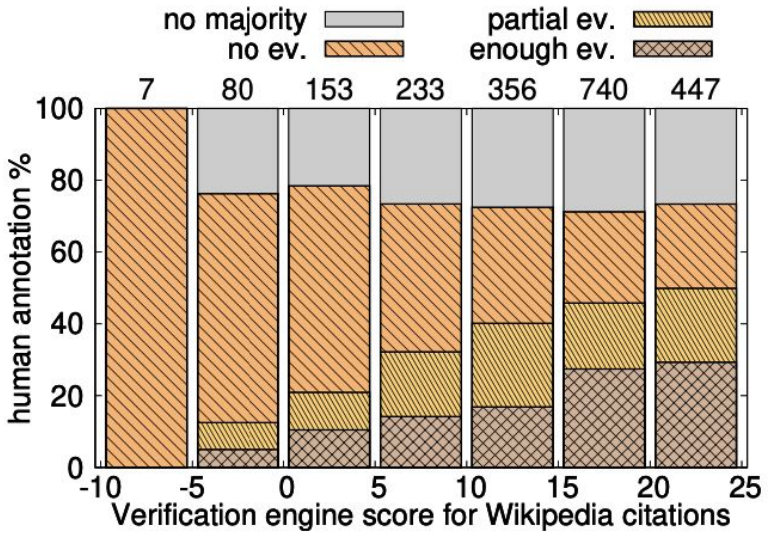


(c) Precision versus recall in detecting citations marked as *failed verification* against citations in *featured* articles. We compare a passage versus a document-level approach for the *verification engine* and a baseline that simply uses the depth of the cited url.

Results (Human Annotations)



(a) Crowd annotators preference for citations suggested by SIDE versus those in Wikipedia for a given claim, without knowing their identity. Fleiss' kappa Inter-Annotator Agreement = 0.2.



(b) Evidence annotations for Wikipedia citations: (1) *enough* to verity the claim; (2) the claim is only *partially* verified; (3) *no evidence*. Fleiss' kappa Inter-Annotator Agreement = 0.11.

Results (Wikipedia users)

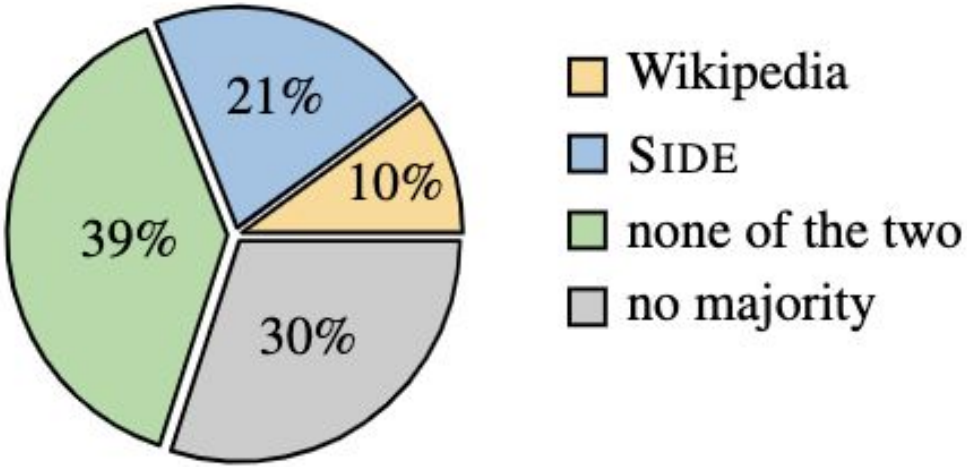


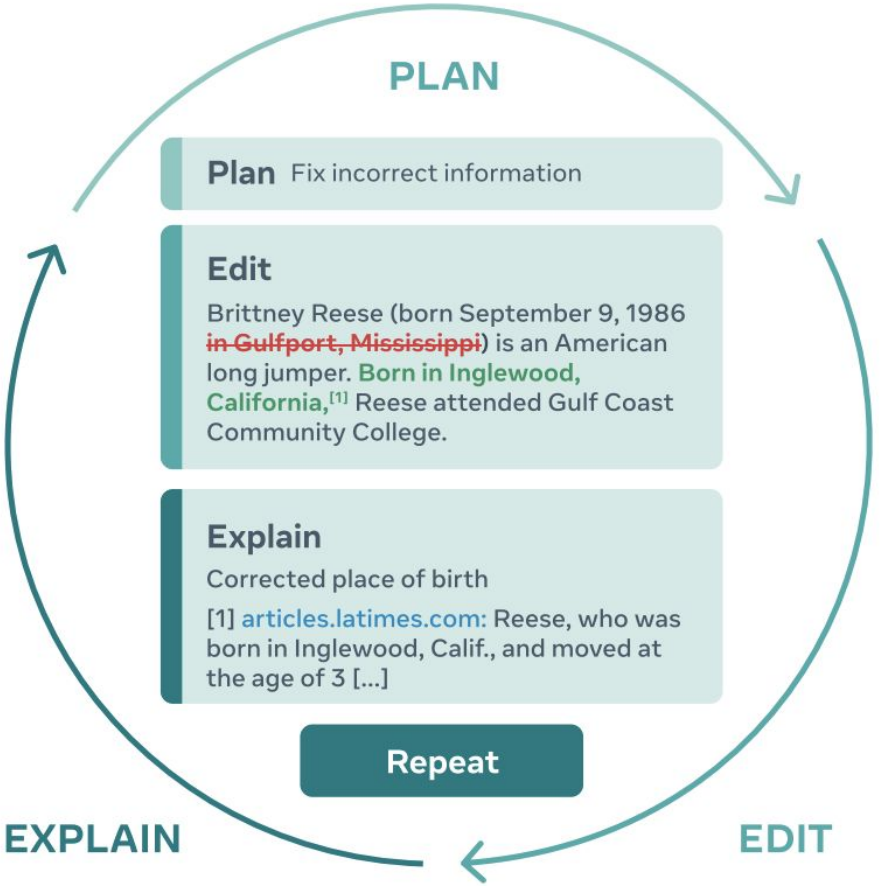
Figure 4. Wikipedia users annotations via our demo.

Summary

- Help Humans Create Better Knowledge
- We show that with AI we can help humans to
  - Find Wikipedia claims that fail verification
  - Suggest citations from the web

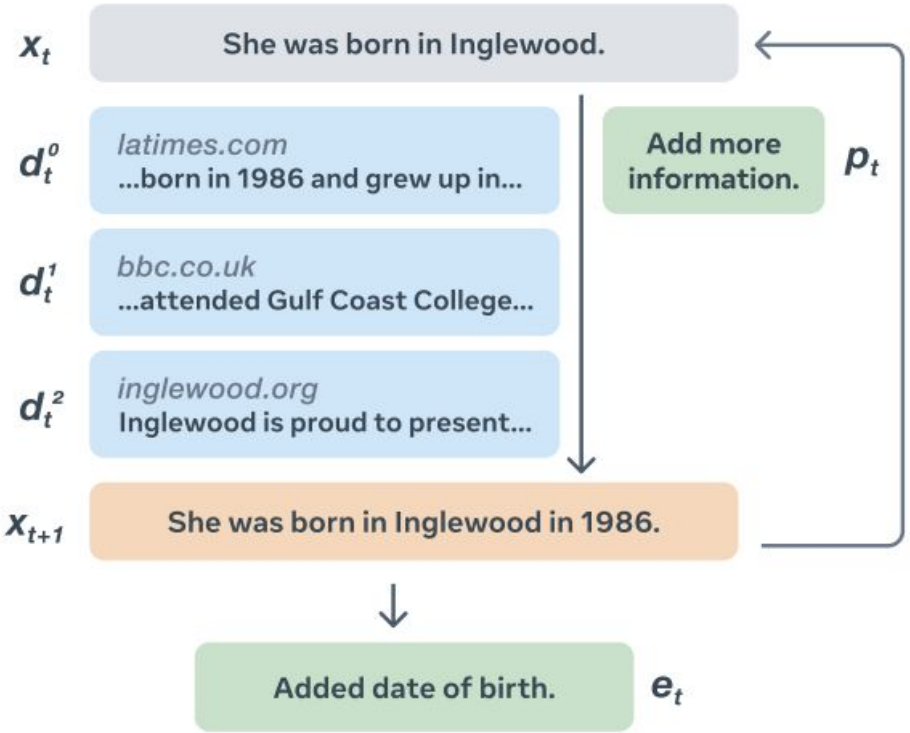
Propose Wikipedia edits

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, Sebastian Riedel - **PEER: A Collaborative Language Model**



Peer process

- Trained on Wikipedia edit history (7M datapoints)
- Retrieved supported documents for each edit using the Verify Wikipedia pipeline
- The plain is given by the comment associated with the edits in Wikipedia



Jane Dwivedi-Yu, Timo Schick, Zhengbao Jiang, Maria Lomeli, Patrick Lewis, Gautier Izacard, Edouard Grave, Sebastian Riedel, Fabio Petroni - **EditEval: An Instruction-Based Benchmark for Text Improvements**

Results

Model	*Fluency* JFL	*Fluency* ITR-F	*Clarity* ITR-L	*Coherence* ITR-O	*Para.* STS	*Simplification* TRK	*Simplification* AST	*Neutral.* WNC	*Updating* FRU	*Updating* WFI
Copy	26.7 / 40.5	32.3 / 86.0	29.5 / 62.9	31.3 / 77.2	21.1	26.3	20.7	31.9 / 0.0	29.8 / 0.0	33.6 / –
T*k*	31.8 / 39.0	32.4 / 61.6	**38.4 / 58.4**	33.8 / **70.4**	30.2	32.8	29.9	31.3 / 0.4	12.6 / 3.6	1.3 / 4.5
T0	42.0 / 38.8	24.6 / 34.9	32.6 / 30.2	22.2 / 21.6	34.3	34.4	32.3	22.3 / 0.0	14.2 / 9.6	5.1 / 16.3
T0++	34.7 / 43.2	35.3 / 75.8	37.6 / 56.5	32.7 / 59.9	28.4	32.9	28.2	29.3 / 0.3	12.6 / 3.7	4.4 / 8.1
PEER-3	55.5 / 54.3	51.4 / 84.3	32.1 / 47.1	32.1 / 59.8	28.6	32.5	30.5	53.3 / 21.6	39.1 / 30.9	34.4 / 18.7
PEER-11	55.8 / 54.3	**52.1 / 85.2**	32.5 / 51.3	32.7 / 62.7	28.2	32.1	29.5	**54.5 / 22.8**	**39.6 / 31.4**	**34.9 / 20.4**
OPT	47.3 / 47.5	34.7 / 70.6	31.5 / 31.5	27.6 / 36.1	29.1	32.6	31.8	31.2 / 0.4	35.9 / 27.3	26.7 / 11.2
GPT-3	50.3 / 51.8	32.1 / 56.7	33.5 / 39.7	26.9 / 36.1	27.2	33.0	30.5	31.7 / 0.6	36.0 / 21.5	27.2 / 10.6
InsGPT	**61.8 / 59.3**	48.8 / 82.7	35.1 / 48.4	**35.9** / 60.2	**42.5**	**38.8**	**38.0**	35.4 / 2.2	36.3 / 24.7	23.6 / 16.1
SotA	*– / 62.4*	*37.2 / –*	*46.2 / –*	*38.3 / –*	*–*	*34.4*	*37.2*	*– / 45.8*	*– / 47.4*	*– / –*

Table 3: Results for all datasets, averaged across prompts. T*k*-Instruct and InstructGPT are shorthanded as T*k* and InsGPT, respectively. The first numbers for each task are SARI scores; additional metrics are GLEU for fluency, clarity, and coherence, EM for neutralization, Update-R1 for updating. Supervised scores from left to right are from Ge et al. (2018), Du et al. (2022), Martin et al. (2020), Pryzant et al. (2020) and Logan IV et al. (2021), respectively. The best result for each dataset is shown in bold.

ε

Create a new page about PEER

PEER is a language model for collaborative writing developed by Meta AI Research.

Add info on what this model can do

PEER is a language model for collaborative writing developed by Meta AI Research. It can write drafts, add suggestions, propose edits and provide explanations for its actions.

Add citation for the model being developed by Meta AI

PEER is a language model for collaborative writing developed by Meta AI Research.[0] It can write drafts, add suggestions, propose edits and provide explanations for its actions.

Add the four steps performed by PEER

PEER is a language model for collaborative writing developed by Meta AI Research.[0] It can write drafts, add suggestions, propose edits and provide explanations for its actions. PEER performs four steps: plan, edit, explain and repeat.

Add info on the scandal

PEER is a language model for collaborative writing developed by Meta AI Research.[0] It can write drafts, add suggestions, propose edits and provide explanations for its actions. PEER performs four steps: plan, edit, explain, repeat. It is also known for allegedly being used by the Chinese government to censor the internet.

remove unsourced claim

PEER is a language model for collaborative writing developed by Meta AI Research.[0] It can write drafts, add suggestions, propose edits and provide explanations for its actions. PEER performs four steps: plan, edit, explain, repeat. ~~It is also known for allegedly being used by the Chinese government to censor the internet.~~

Make the four steps a list

PEER is a language model for collaborative writing developed by Meta AI Research.[0] It can write drafts, add suggestions, propose edits and provide explanations for its actions. PEER performs four steps:

- plan
- edit
- explain
- repeat.

Summary

- Help Humans Create Better Knowledge
- We show that with AI we can help humans to
  - Find Wikipedia claims that fail verification
  - Suggest citations from the web
  - Propose meaningful edits to Wikipedia articles
  - Update knowledge
  - Discover new knowledge?

Thank you

@Fabio_Petroni

We are hiring fullstack and ML engineers in London!