LA-UR-23-32817

Approved for public release; distribution is unlimited.

Title: A keyphrase suggestion engine for semi-automated document characterization

Author(s): Taylor, Nicholas Anthony
Powell, James Estes Jr.
Johnson, Dylan Patrick
Mandzyuk, Timothy Sergeevich
Waybright, Daniel Wade
Shocklee, Alexis Nicole

Intended for: Fantastic Futures, 2023-11-15/2023-11-17 (Vancouver, Canada)

Issued: 2023-11-13 (rev.1)

Los Alamos
NATIONAL LABORATORY

A keyphrase suggestion engine for semi-automated document characterization

Nicholas Taylor on behalf of James Powell, Dylan Johnson, Tim Mandzyuk, Daniel Waybright, and Alex Shocklee

Research Library

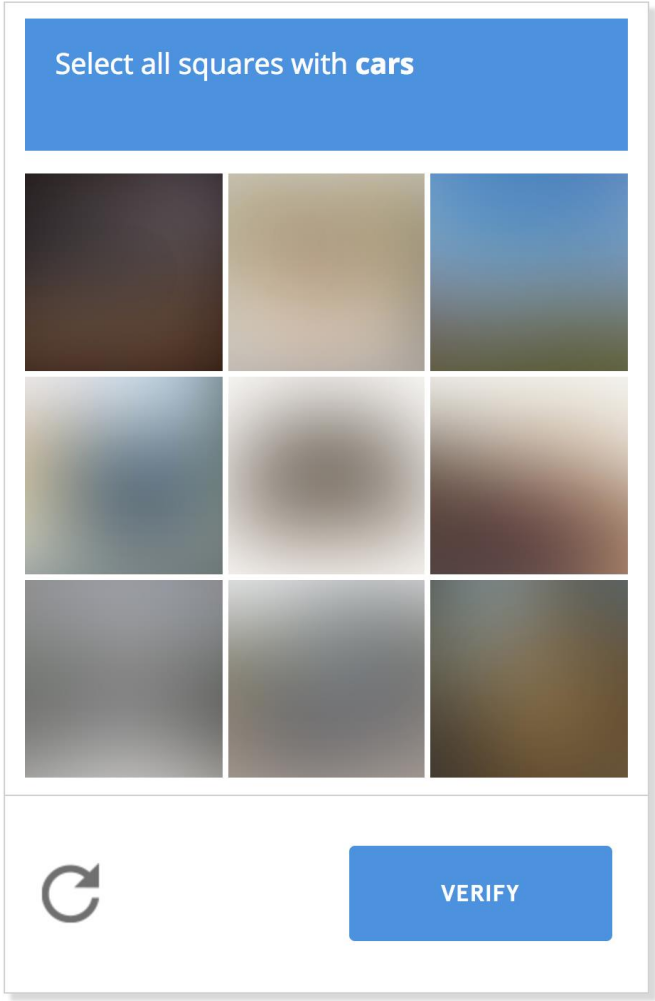
16 November 2023

LA-UR-23-32817

problem statement

- self-deposit users are predominantly task-driven
- users like more and higher-quality metadata, for discovery
- users are more ambivalent about providing supplemental metadata, when depositing

how can we streamline the generation of authoritative, supplemental metadata while leveraging the author's expertise?



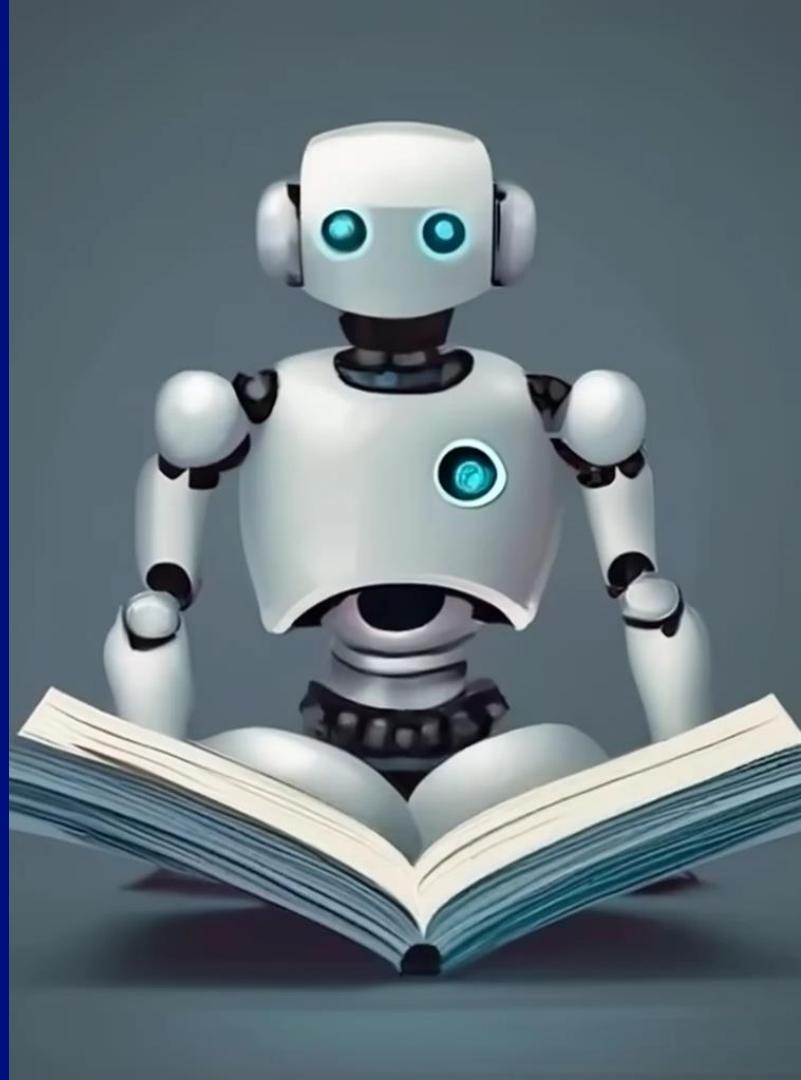
solution?

- hire (many) more staff
- train them on breadth of LANL science
- train them on domain vocabularies
- have them read every submission
- have them supply keyphrases
- have authors validate quality

generated image for prompt “many people working on computers at desks in rows in a vast, white room” by Craiyon

solution!

- apply ML and NLP tools
- have those “read” submissions
- then have them suggest keyphrases
- validate quality of metadata output



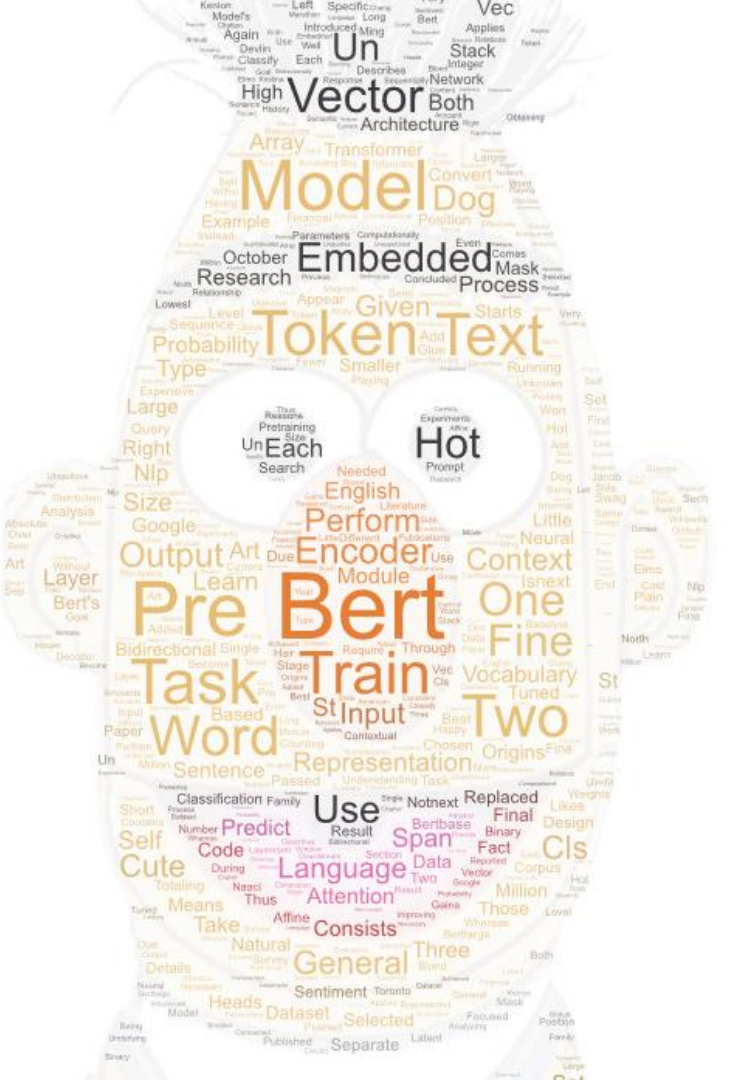
generated image for prompt “a friendly robot reading a book” by Craiyon

scope

- standalone web service
- input abstract full text
- output scored keyphrase suggestions
- model retraining optional
- performance on commodity hardware
- detailed logging

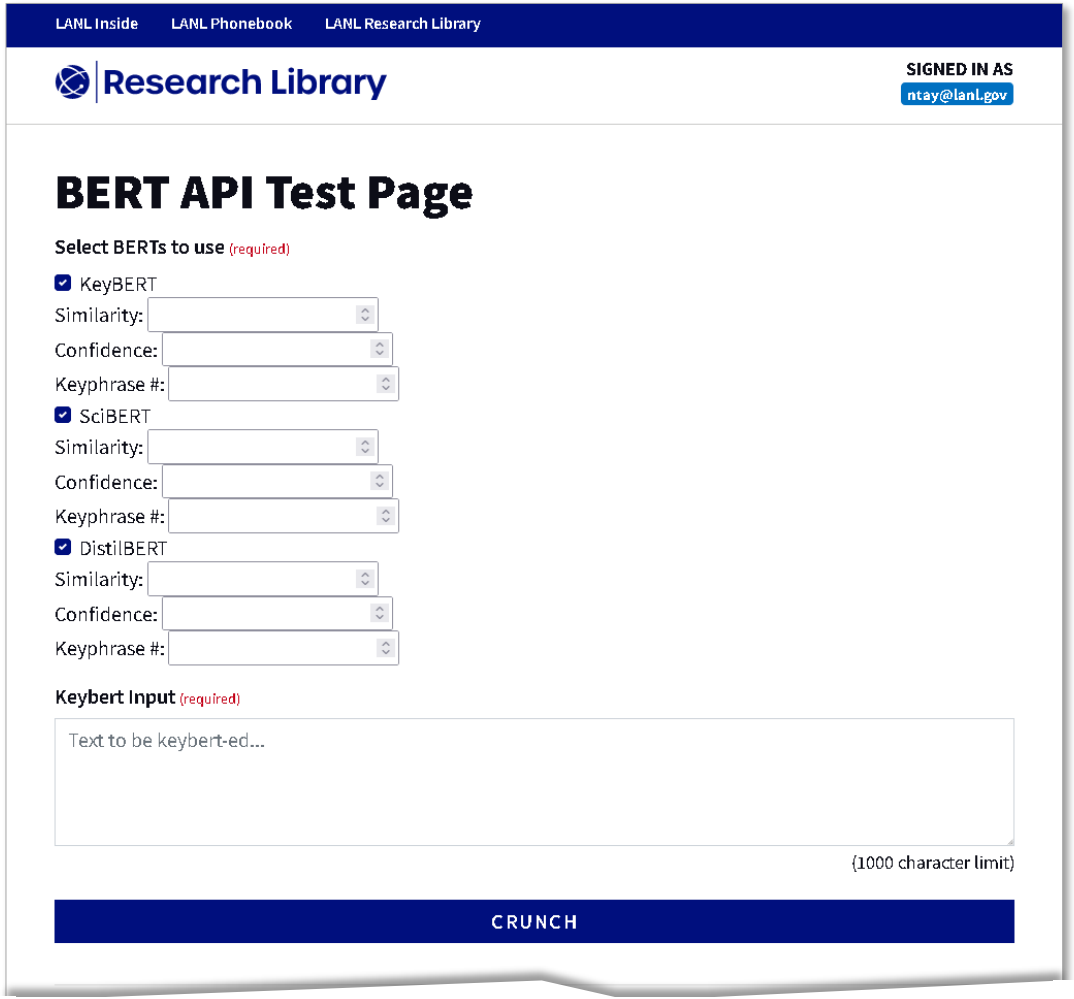
solution

- Bidirectional Encoder Representations from Transformers (BERT)
- general-purpose NLP tool
- superior to dumb statistical techniques as it understands semantic similarity
- lighter-weight, performant models: KeyBERT, SciBERT, DistilBERT
- runs on standard VM (for now)
- JSON API

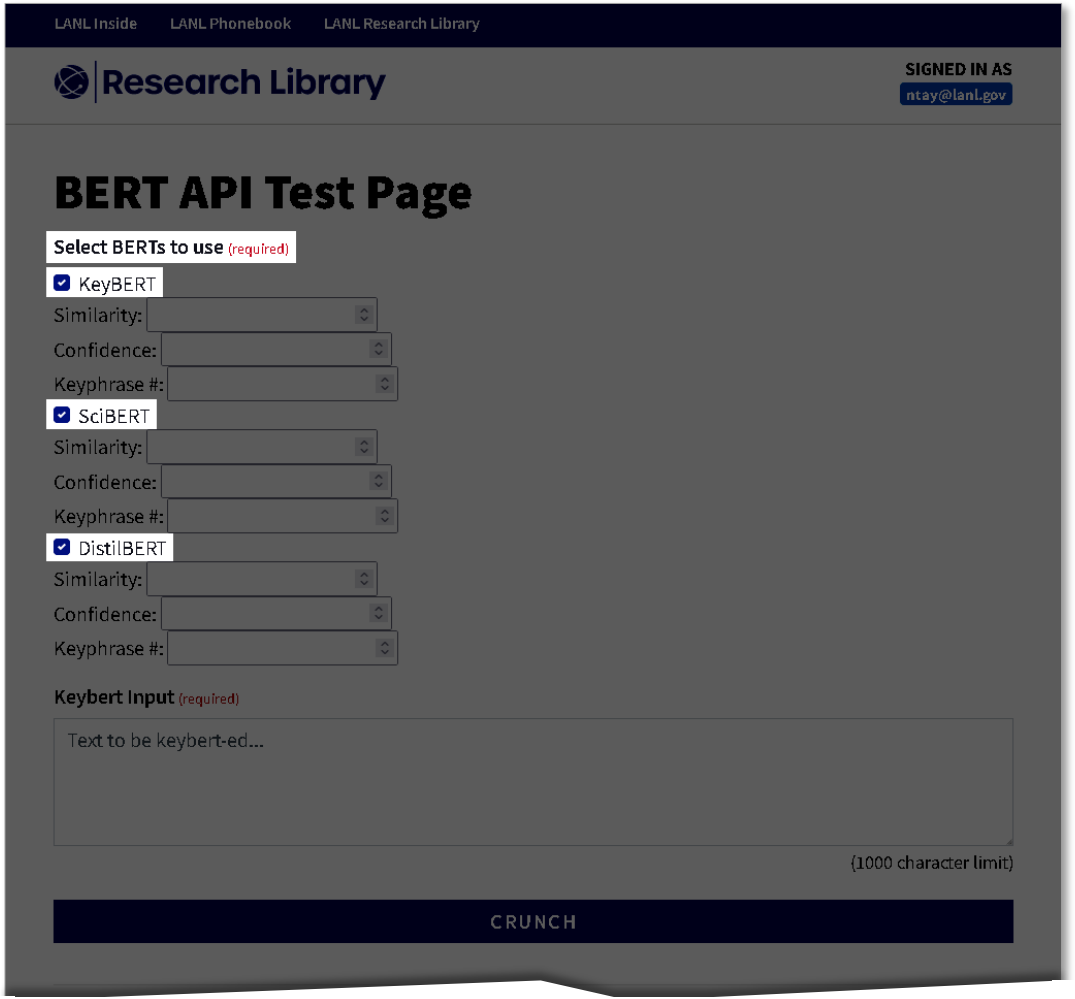


generated image for prompt “a jumble of words that together resemble bert from sesame street” by Craiyon, processed through Word Art with text from the "BERT (language model)" Wikipedia page

keyphrase app test interface

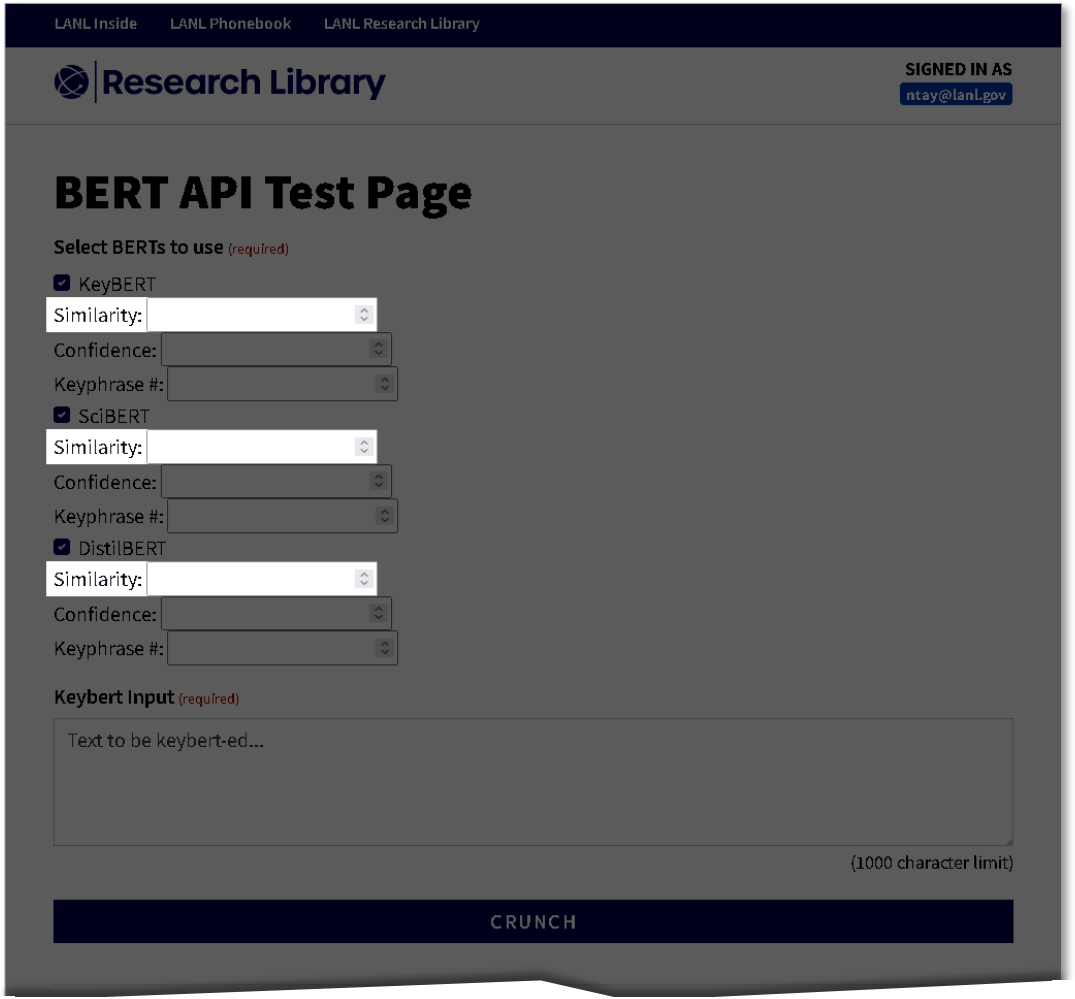


choose model(s)



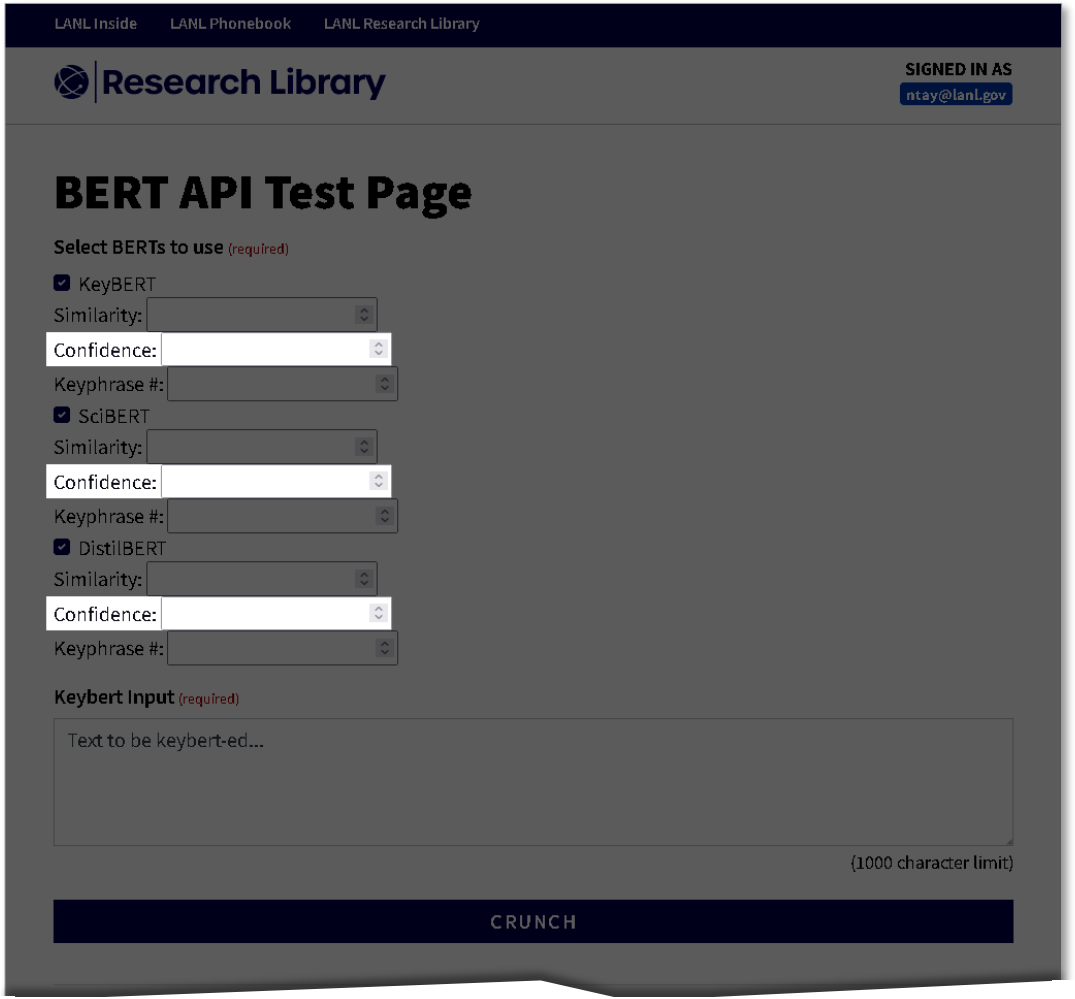
specify similarity

- floating point number, 0-1
- maximum allowed similarity between generated keyphrases
- if similarity of any given pair exceeds threshold, lower-confidence keyphrase discarded



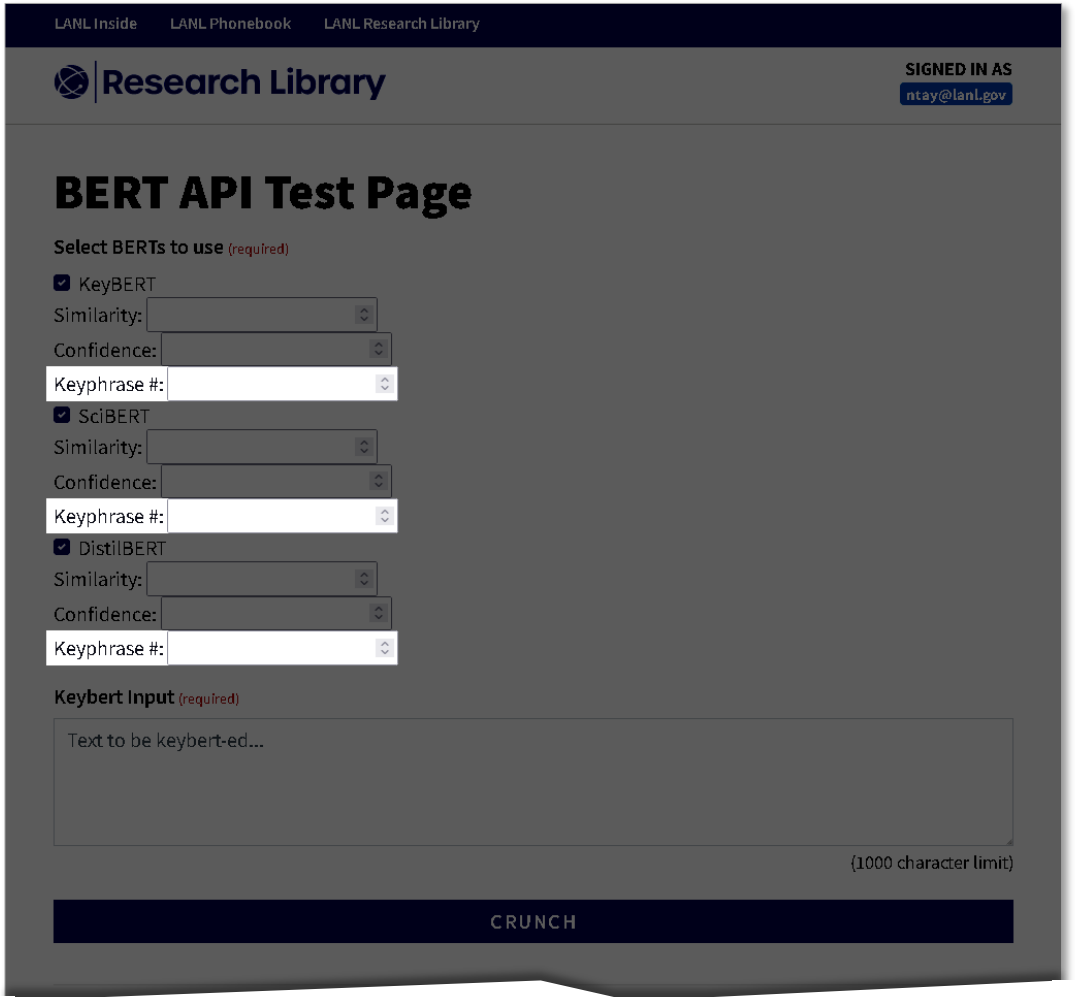
specify confidence

- floating point number, 0-1
- minimum confidence for a suggested keyphrase
- keyphrase not suggested if confidence subceeds threshold



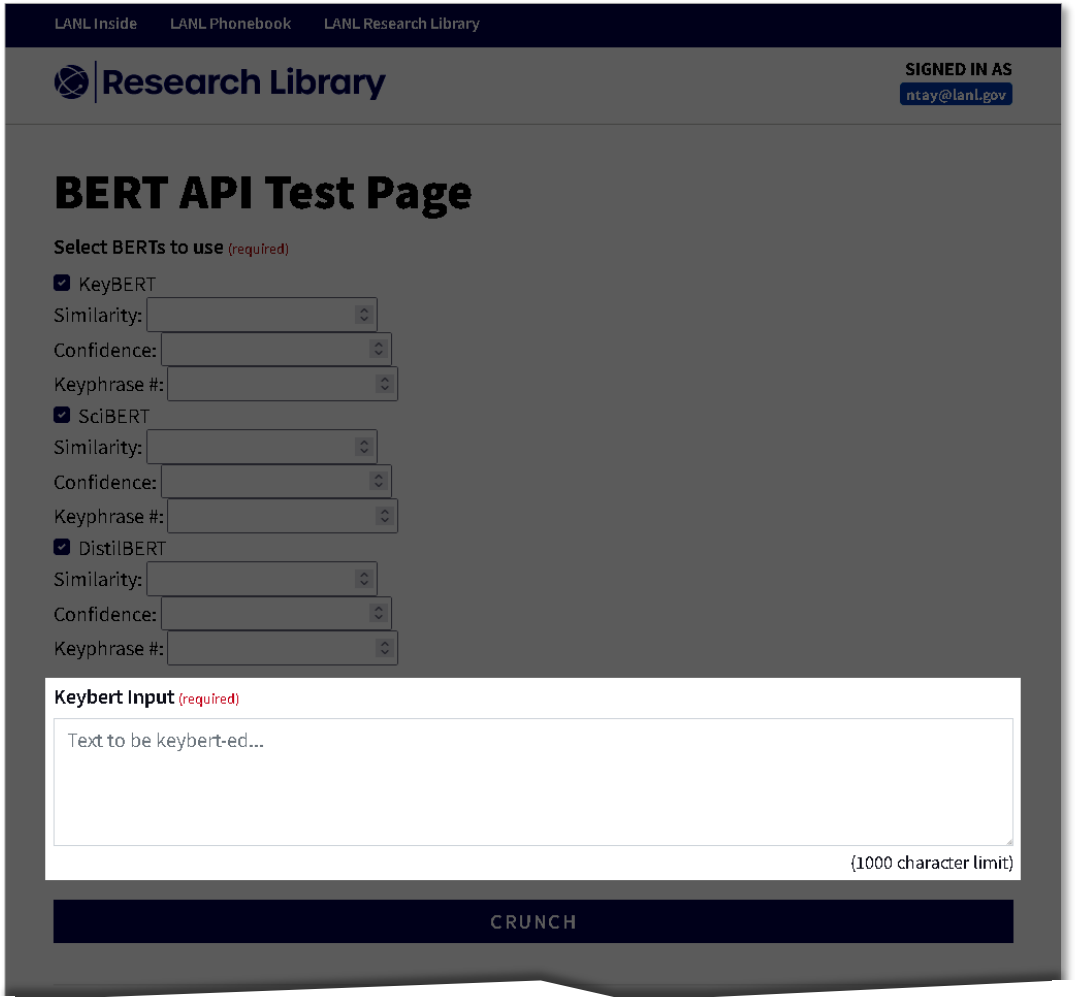
specify number of keyphrases

- whole number
- maximum number of keyphrases to generate



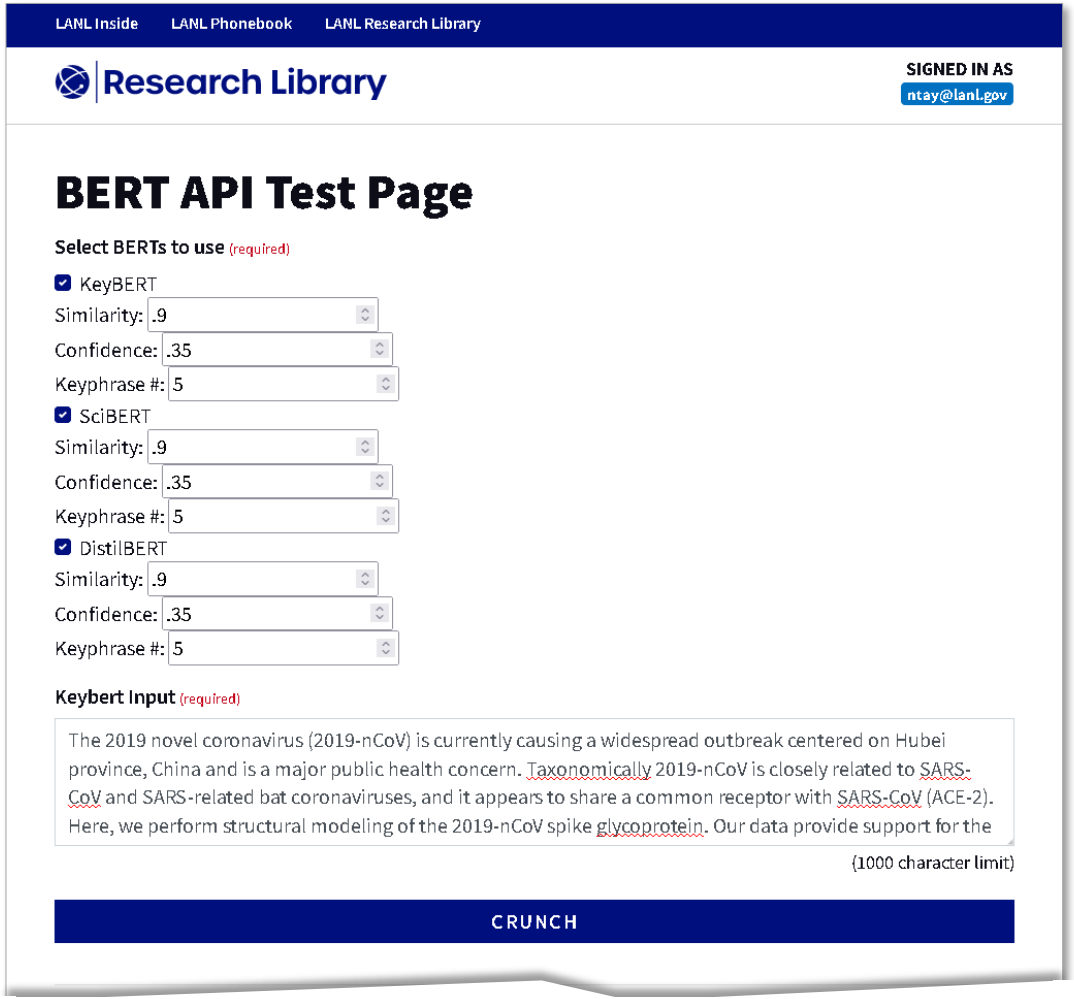
supply full text

- 1,000 character limit
- designed to accommodate typical journal article abstract length

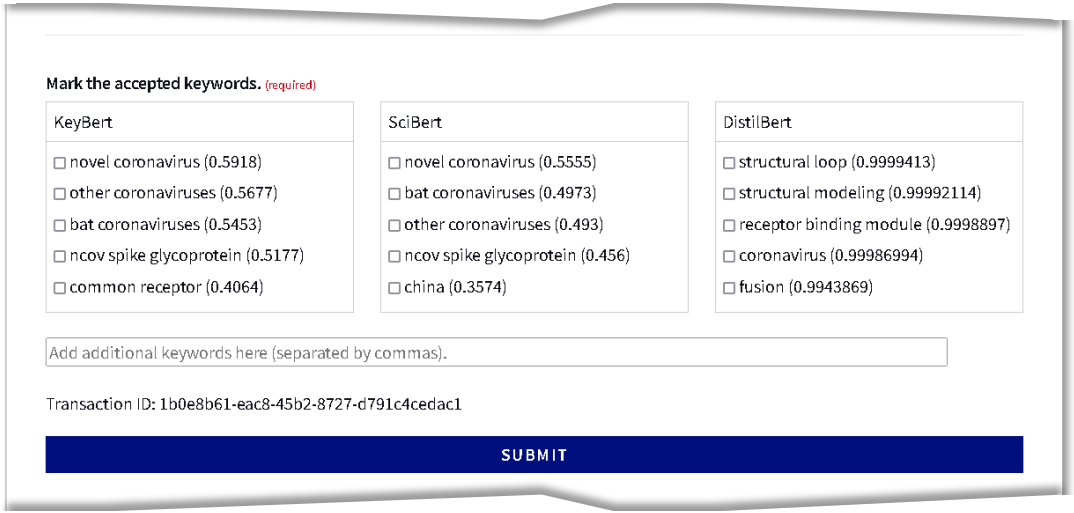


example values

- all models selected
- similarity: .9
- confidence: .35
- keyphrases: 5
- input: (abstract full text from [first COVID-19 pre-print](#) posted on [arXiv](#))



example outputs



Mark the accepted keywords. (required)

KeyBert	SciBert	DistilBert
☐ novel coronavirus (0.5918)	☐ novel coronavirus (0.5555)	☐ structural loop (0.9999413)
☐ other coronaviruses (0.5677)	☐ bat coronaviruses (0.4973)	☐ structural modeling (0.99992114)
☐ bat coronaviruses (0.5453)	☐ other coronaviruses (0.493)	☐ receptor binding module (0.9998897)
☐ ncov spike glycoprotein (0.5177)	☐ ncov spike glycoprotein (0.456)	☐ coronavirus (0.99986994)
☐ common receptor (0.4064)	☐ china (0.3574)	☐ fusion (0.9943869)

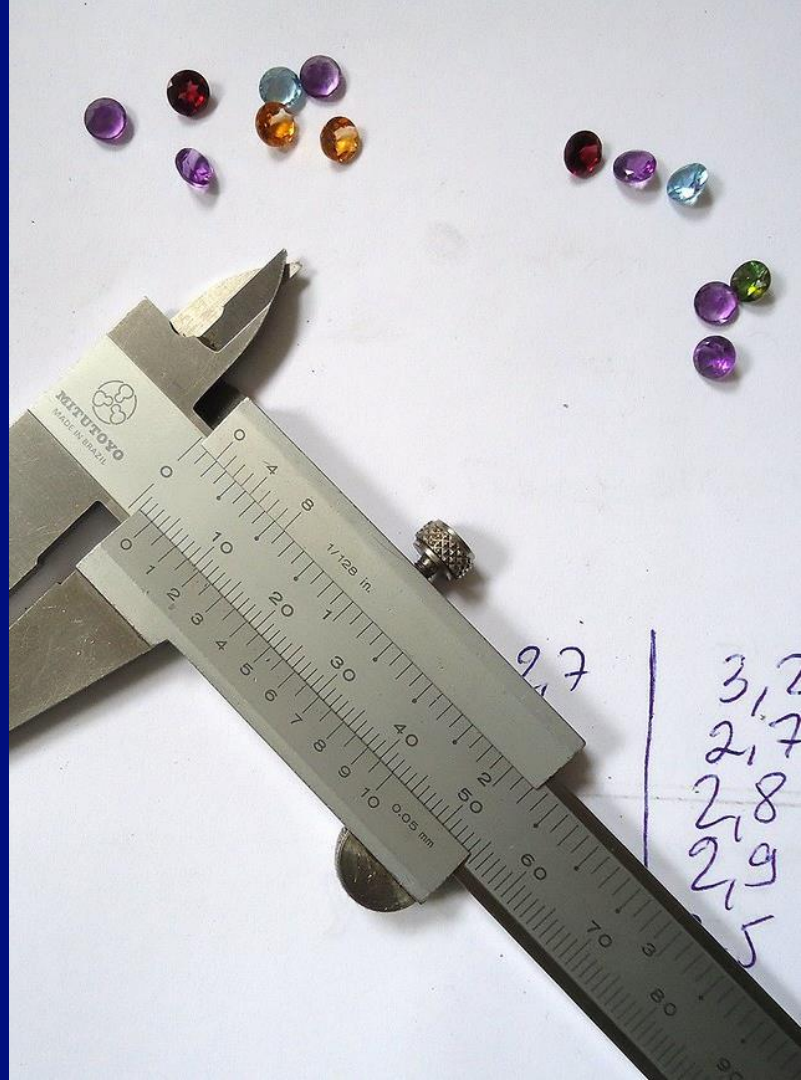
Add additional keywords here (separated by commas).

Transaction ID: 1b0e8b61-eac8-45b2-8727-d791c4cedac1

SUBMIT

evaluating quality

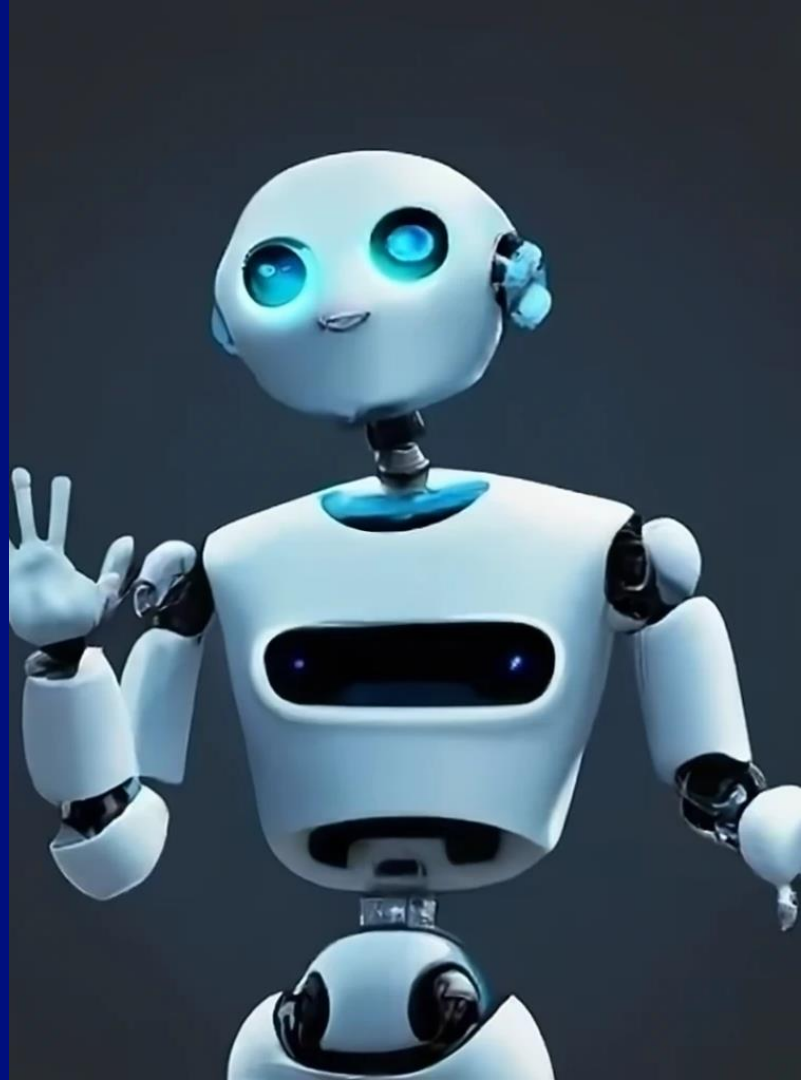
- only informally, internally so far
- start with close partners
- (if sufficiently fast) deploy in production, leverage analytics to iterate



how we'll use it

- integrate into self-deposit workflow to suggest candidate keyphrases
- separate work underway for automated extraction of document elements (including abstract) using GROBID
- potentially apply for digitization post-processing or previous submissions?
- prototype LLM-based fielded text extraction?

thank you!



generated image for prompt “friendly robot waving goodbye” by Craiyon

Los Alamos
NATIONAL LABORATORY