

Modeling Exemplification in Long-form Question Answering via Retrieval

Shufan Wang¹, Fangyuan Xu², Laure Thompson¹, Eunsol Choi², Mohit Iyyer¹

1 - University of Massachusetts Amherst, 2 - University of Texas at Austin

ELI5: a dataset for long-form QA

Input: Information-seeking questions that call for “explanations”

Output: Answers that address given questions with a few sentences or paragraphs

ELI5: a dataset for long-form QA

Question: Do people who learn to speak a second language think in that language as well as their original?

Input: Information-seeking questions that call for “explanations”

Output: Answers that address given questions with a few sentences or paragraphs

ELI5: a dataset for long-form QA

Question: Do people who learn to speak a second language think in that language as well as their original?

Answer: I speak English, Italian and Portuguese fluently. The language of my thoughts depends on the context that I'm in. I.e, if I'm watching a Brazilian film, I'll think in Portuguese etc ... For instance: I live in the UK now so I usually think in English ... however if I go see my buddies in Italy, I will have to translate from Italian, to English (for listening) then back to Italian to speak... (**204 words**)

Input: Information-seeking questions that call for “explanations”

Output: Answers that address given questions with a few sentences or paragraphs

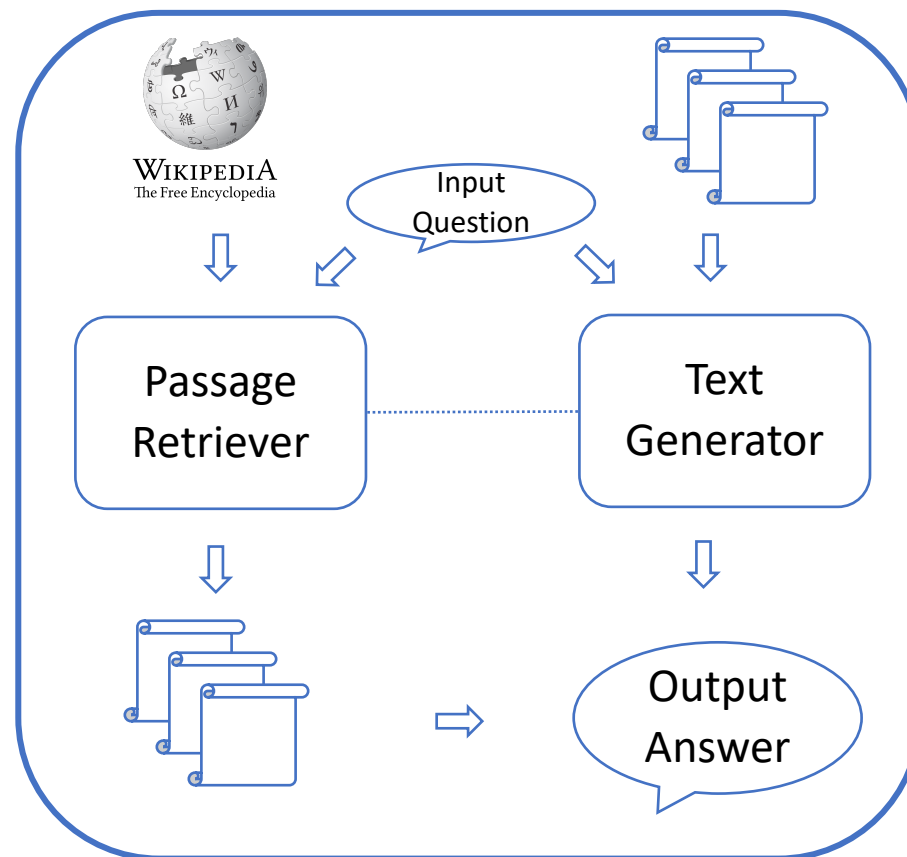
ELI5: a dataset for long-form QA

~~In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".~~

~~What causes precipitation to fall?
gravity~~

~~What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel~~

Unlike SQUAD, where the answer is a span from given context



Answer question based on question + retrieved passages.

ELI5: a dataset for long-form QA

Question: Do people who learn to speak a second language think in that language as well as their original?

Answer: I speak English, Italian and Portuguese fluently. The language of my thoughts depends on the context that I'm in. I.e, if I'm watching a Brazilian film, I'll think in Portuguese etc ... For instance: I live in the UK now so I usually think in English ... however if I go see my buddies in Italy, I will have to translate from Italian, to English (for listening) then back to Italian to speak... (**204 words**)

- Explanation-with-example is a very common phenomenon in LFQA
- About 20% of answers contains the “for example” / “e.g.”.
- Many more use examples signaled by other phrases or implicitly make use of examples

Outline

- A detailed human annotation study to understand “exemplification”
- Modeling “exemplification”: generative vs retrieval approach
- Human evaluations

Understand “exemplification” through human annotations

- We annotate instances of “exemplification” from three datasets (with three different domains)
 - ELI5: online question-answering forum on Reddit
 - NaturalQA: Wikipedia passages
 - Books from the Pile: fiction and non-fictions from various topics
- The discourse structure of “exemplification” (Meyer, 1992, Triki, 2021):

Many languages have shared origins. For example, French and Italian belong to the “romance” language family.

The diagram illustrates the discourse structure of the sentence above. It uses three brackets to identify different parts: an orange bracket under 'Many languages have shared origins.' is labeled 'Anchor (exemplified unit)'; a red bracket under 'For example,' is labeled 'Exemplification Marker'; and a green bracket under 'French and Italian belong to the “romance” language family.' is labeled 'Example (exemplifying unit)'.

Categories of “exemplifying units”

Personal vs Non-personal

- Personal Example

The language of my thoughts depends on the context that I'm in. I.e, if I'm watching a Brazilian film, I'll think in Portuguese etc ... For instance: In the UK I usually think in English ...

- Non-Personal Example

For a given pressure , different liquids boil **219** at different temperatures. For example, *water 220 boils at 100° C (212° F) at sea level, but at 93.4° C (200.1° F) at 2,000 metres (6,600 ft) altitude.*



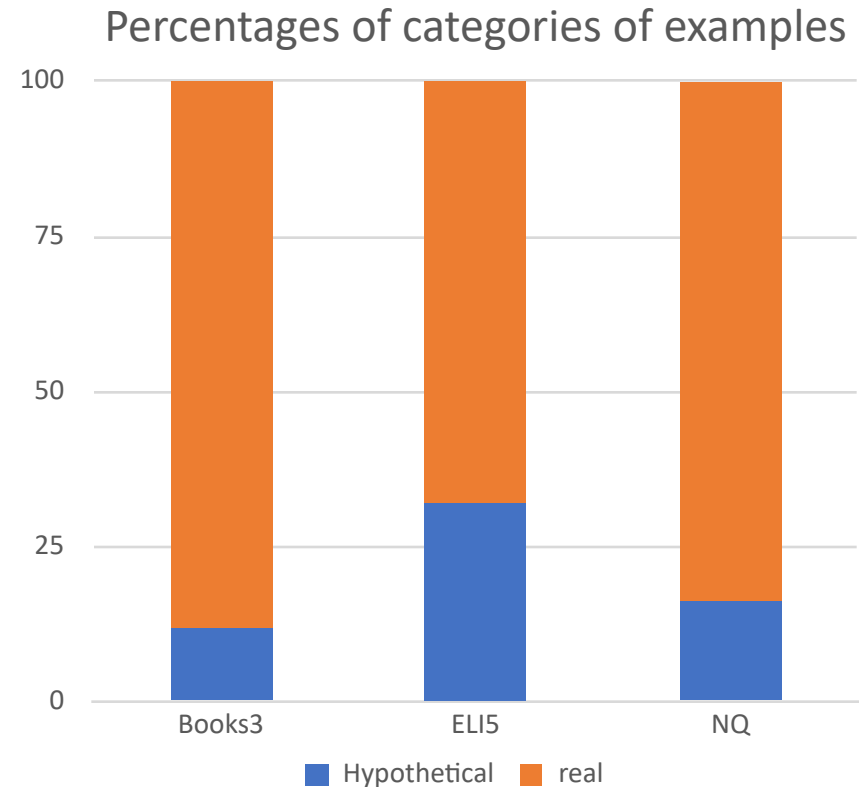
Categories of “exemplifying units” Hypothetical vs Real

- Hypothetical Example

The Chaos Theory is ... about that, ... it's incredibly difficult to predict what will happen given the initial situation. For example, imagine a peculiar crossroad with 10 possible options. When the driver makes the choice, he only turns the wheel by a few degrees. Yet each road starting from there, leads to a completely different part of the country.

- Real Example

CEOs lead a range of organizations , including public and private corporations, non-profit organizations and even some government organizations (e.g., Crown corporations).



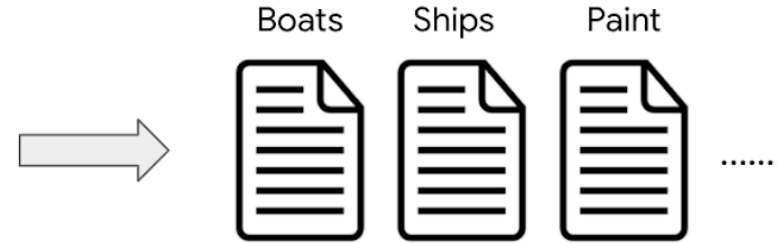
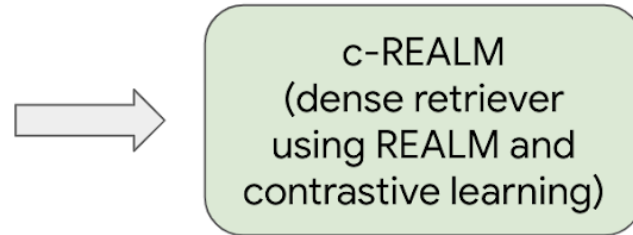
SOTA¹ models generate examples the same way as other content

Stage 1: Retrieval

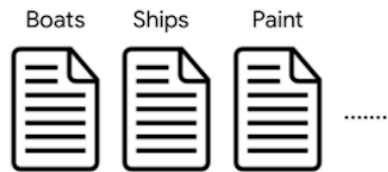


WIKIPEDIA
The Free Encyclopedia

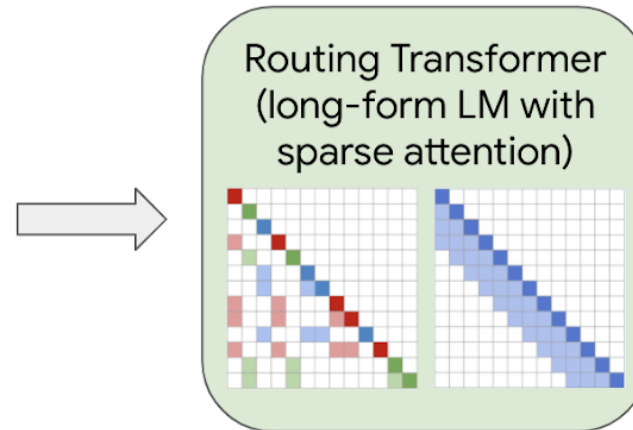
Q. Why are almost all boats white?



Stage 2: Generation



Q. Why are almost all boats white?



A. Boats are generally painted white for obvious reasons, as they are inexpensive and easy to clean. Ships are also often painted to provide a visible visual contrast to the sea. Boats also often have high visibility ports to allow for easy international trade. In addition, many boats have an air-breathing function that makes them easier to spot and stay afloat, which is a good thing if you're a pirate or trying to get into some country's waters. Lastly, many ships are built in the last few decades, so they are not as exotic looking as they were thirty years ago. And lastly, many ships are made out of materials that are easy to work with, and so many of them are painted white as a practical matter.

C-REALM-RT

Modeling exemplification with generative models is not ideal.

Generative models (e.g. C-REALM-RT) do not treat examples differently during generation

Modeling exemplification with generative models is not ideal.

Generative models (e.g. C-REALM-RT) do not treat examples differently during generation

Generative models are evaluated using ROUGE-L (problematic evaluation)

Modeling and evaluating exemplification is difficult

Generative models (e.g. C-REALM-RT) do not treat examples differently during generation

Generative models are evaluated using ROUGE-L (problematic evaluation)

Question: Do people who learn to speak a second language think in that language as well as their original?

C-REALM-RT-Generated example (17.9 ROUGE-L)

I speak English, Italian and Portuguese fluently. The language of my thoughts depends on the context that I'm in. I didn't think I was doing the thinking You think English when you're speaking Portuguese, but think in Brazilian when you're speaking Portuguese. Most of the time though it's more of a habit to think in a language you don't know

Modeling exemplification with generative models is not ideal.

Generative models (e.g. C-REALM-RT) do not treat examples differently during generation

Generative models are evaluated using ROUGE-L (problematic evaluation)

Shows logical inconsistency and hallucinations

Question: Do people who learn to speak a second language think in that language as well as their original?

C-REALM-RT-Generated example (17.9 ROUGE-L)

I speak English, Italian and Portuguese fluently. The language of my thoughts depends on the context that I'm in. I didn't think I was doing the thinking You think English when you're speaking Portuguese, but think in Brazilian when you're speaking Portuguese. Most of the time though it's more of a habit to think in a language you don't know

Modeling exemplification with generative models is not ideal.

Generative models (e.g. C-REALM-RT) do not treat examples differently during generation

Generative models are evaluated using ROUGE-L (problematic evaluation)

Shows logical inconsistency and hallucinations

Does not illustrate the point "language of my thoughts depends on the context"

Question: Do people who learn to speak a second language think in that language as well as their original?

C-REALM-RT-Generated example (17.9 ROUGE-L)

I speak English, Italian and Portuguese fluently. The language of my thoughts depends on the context that I'm in. I didn't think I was doing the thinking You think English when you're speaking Portuguese, but think in Brazilian when you're speaking Portuguese. Most of the time though it's more of a habit to think in a language you don't know

Modeling exemplification with generative models is not ideal.

Examples are diverse and complicated

Examples are challenging for generative models

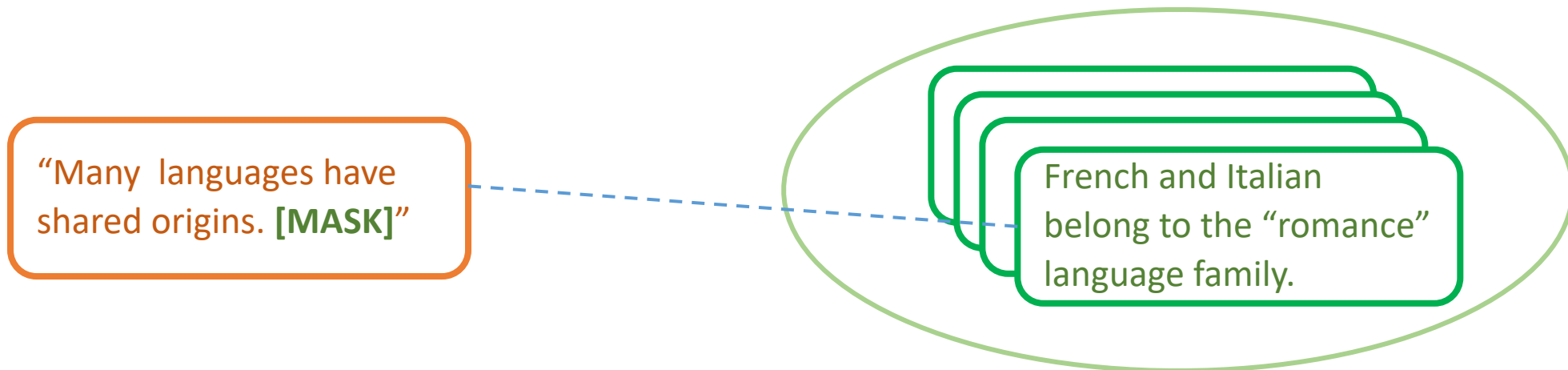
Current eval frameworks (ROUGE) are not informative for examples

Framing **exemplification** as a retrieval problem

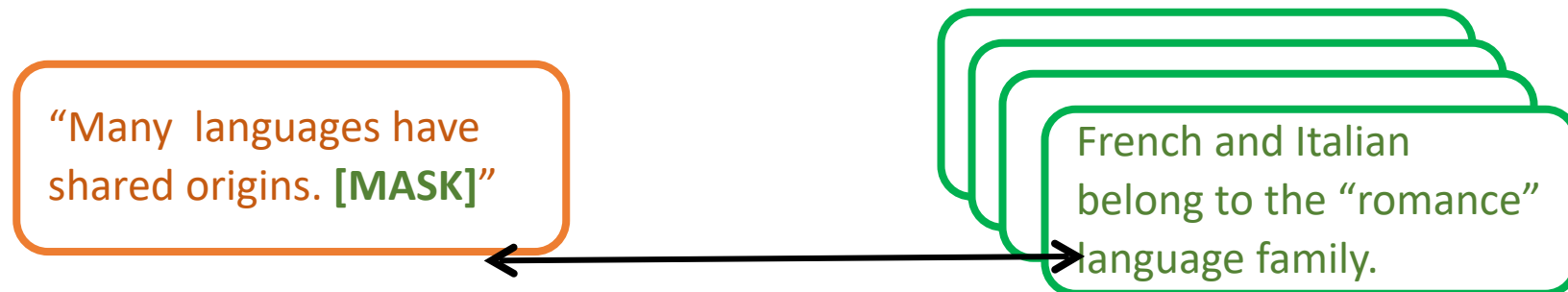
- Using generative models for **exemplification** is problematic. To study **exemplification** in a more principled way, we re-frame it as retrieval problem
- The retrieval approach allows us to use ranking-based metrics (instead of uninformative metrics like ROUGE-L) to evaluate the quality of examples retrieved.

Framing **exemplification** as a retrieval problem

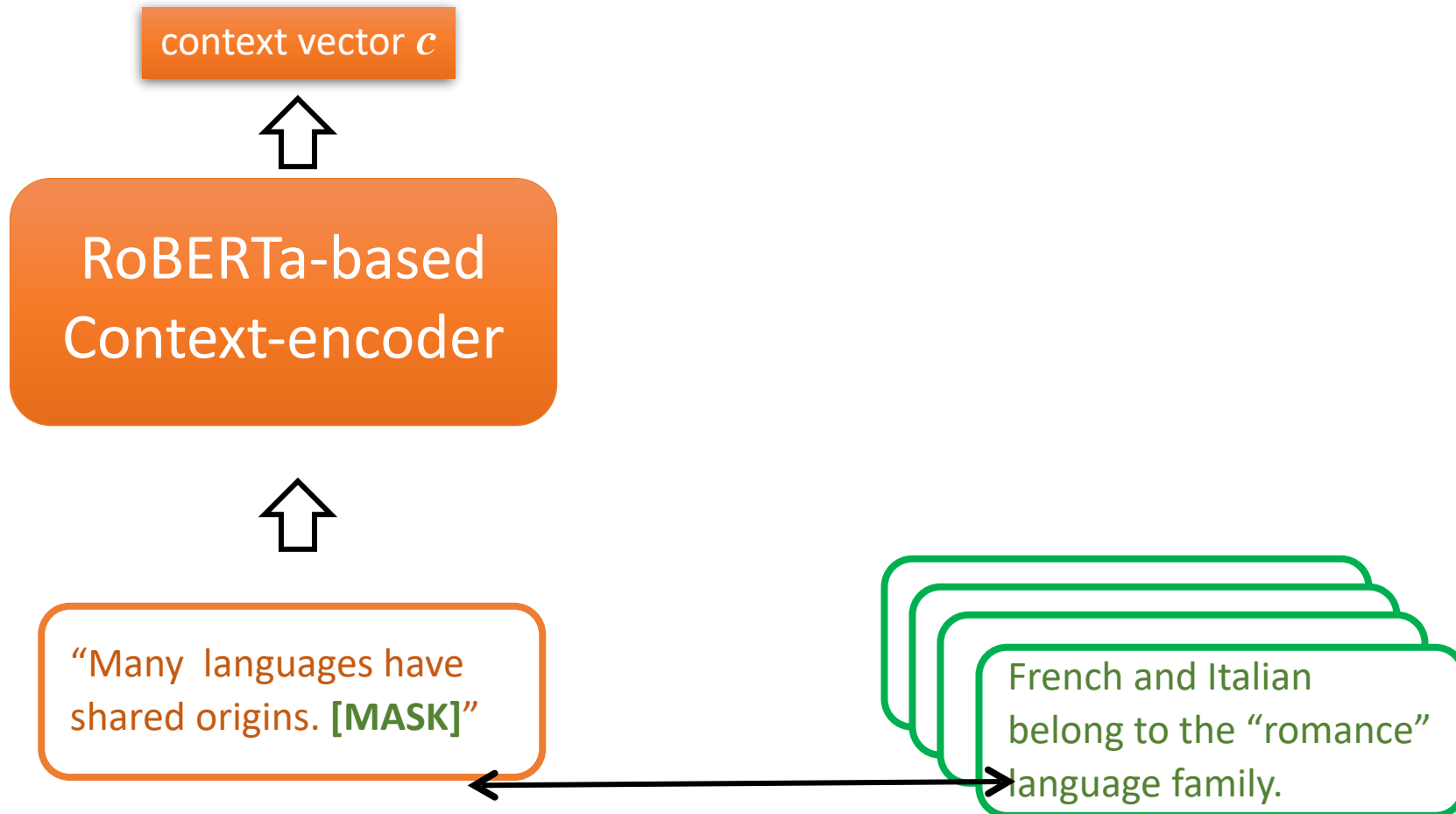
- Build ELI5 into a large collection of (query, value), where
 - Query = Anchor / context
 - Value = Example
 - This naturally fit into the retrieval framework



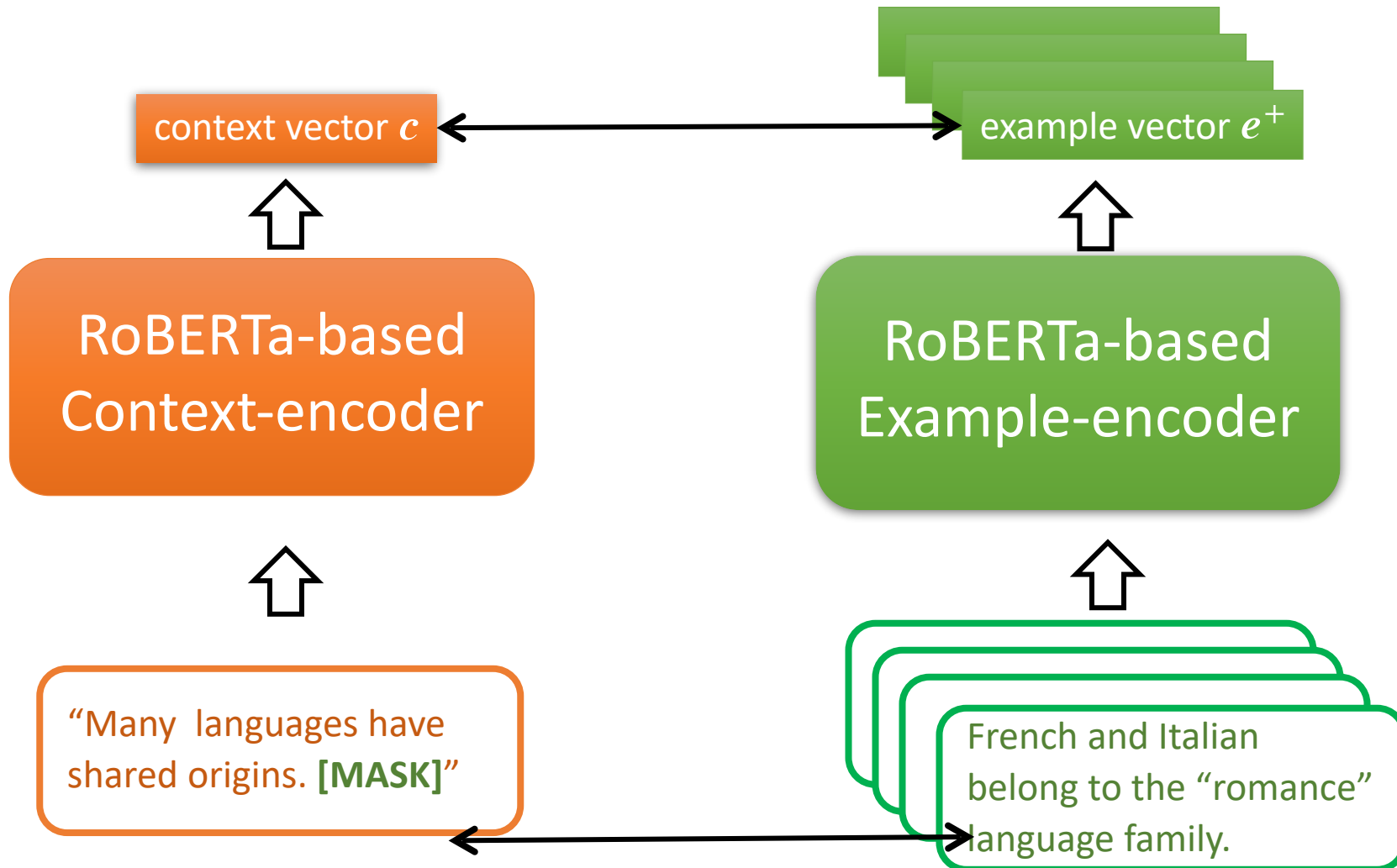
EGRET: an example retriever for long-form questions answering



EGRET: an example retriever for long-form questions answering

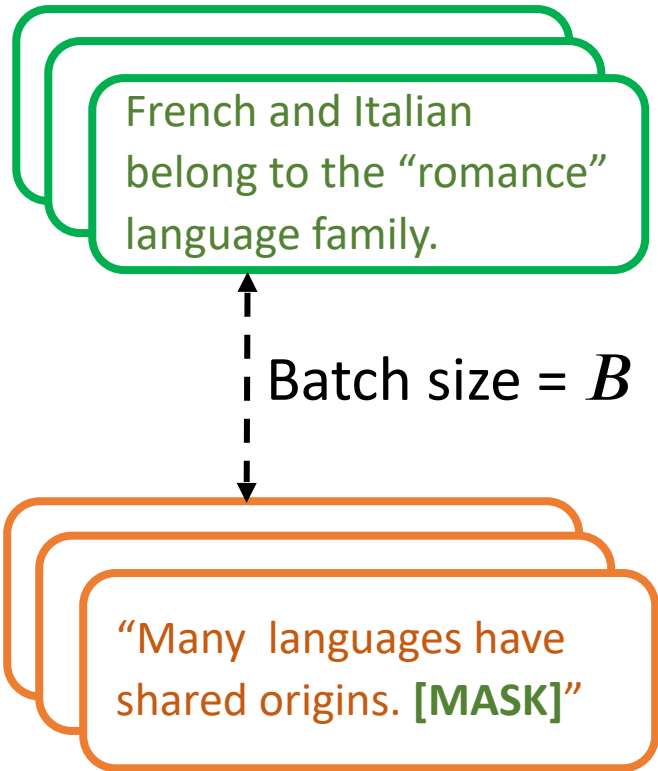


EGRET: an example retriever for long-form questions answering

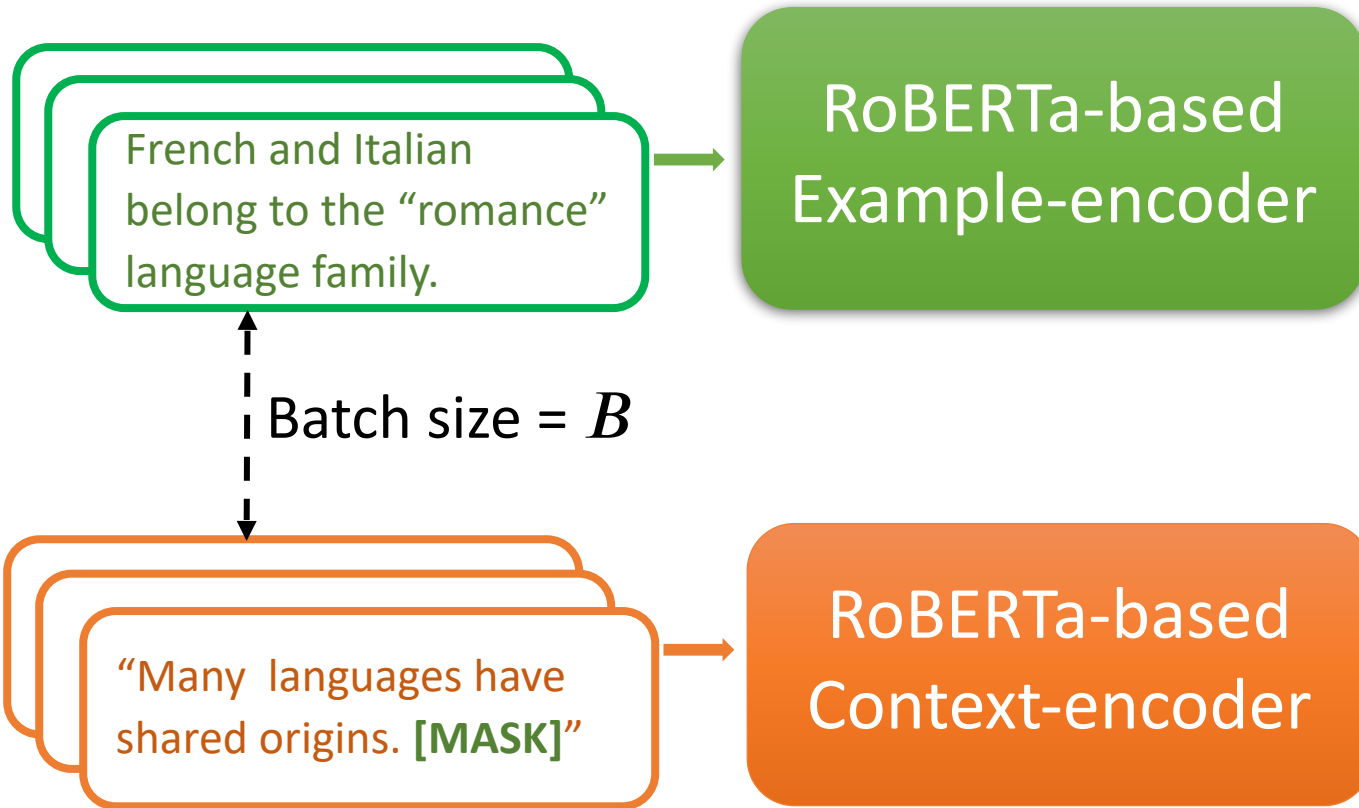


EGRET is trained with contrastive learning,
using in-batch negatives

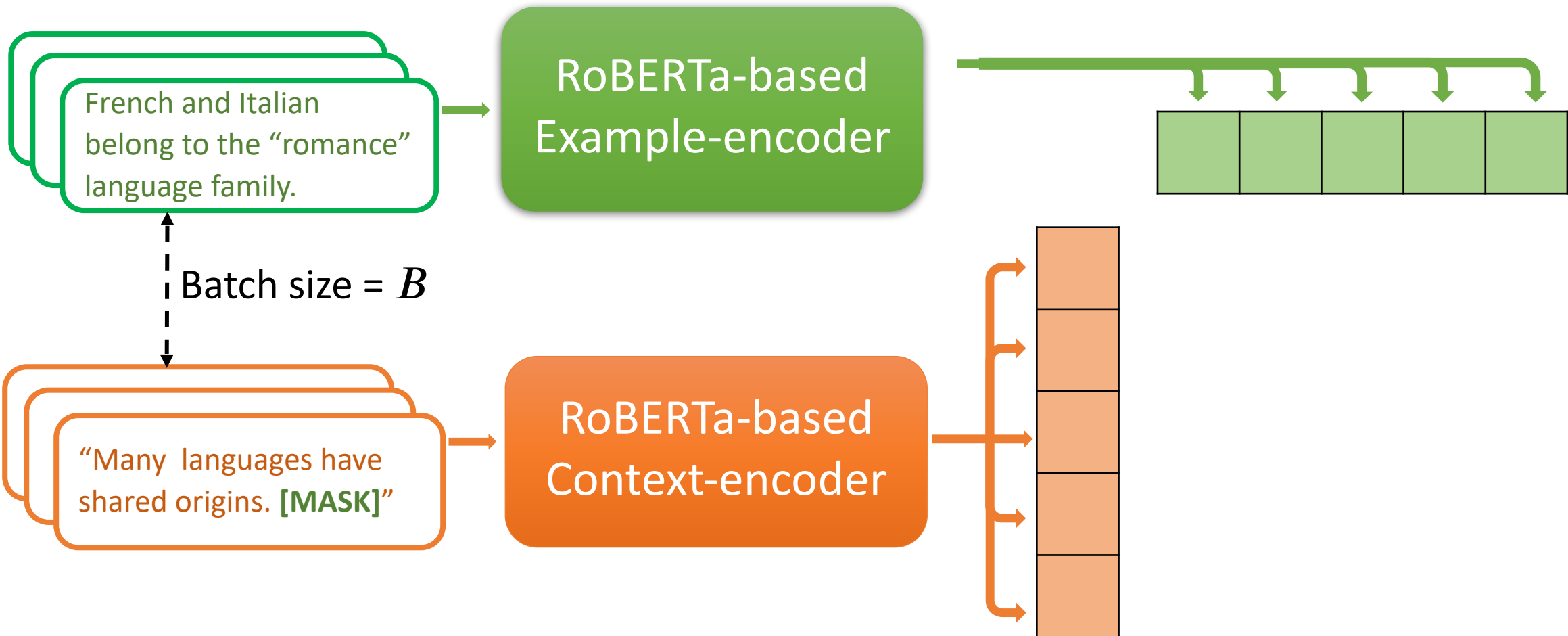
EGRET is trained with contrastive learning, using in-batch negatives



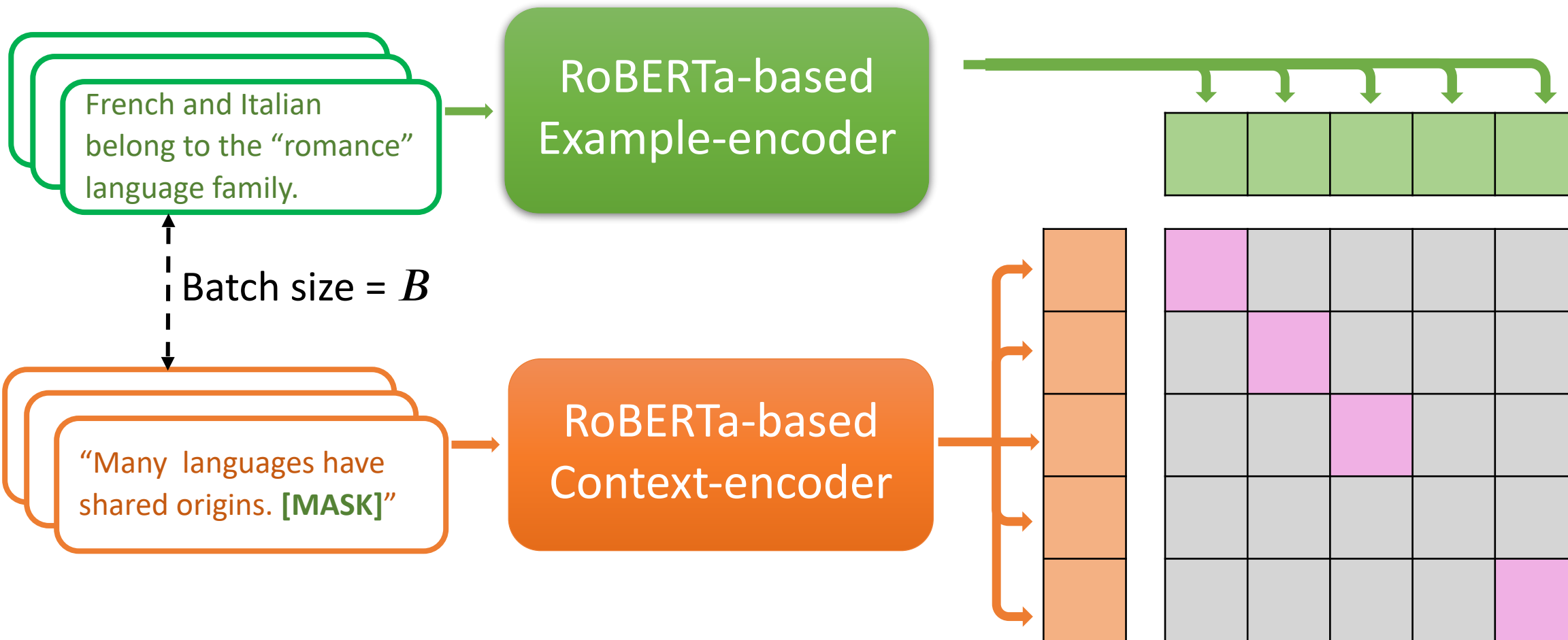
EGRET is trained with contrastive learning, using in-batch negatives



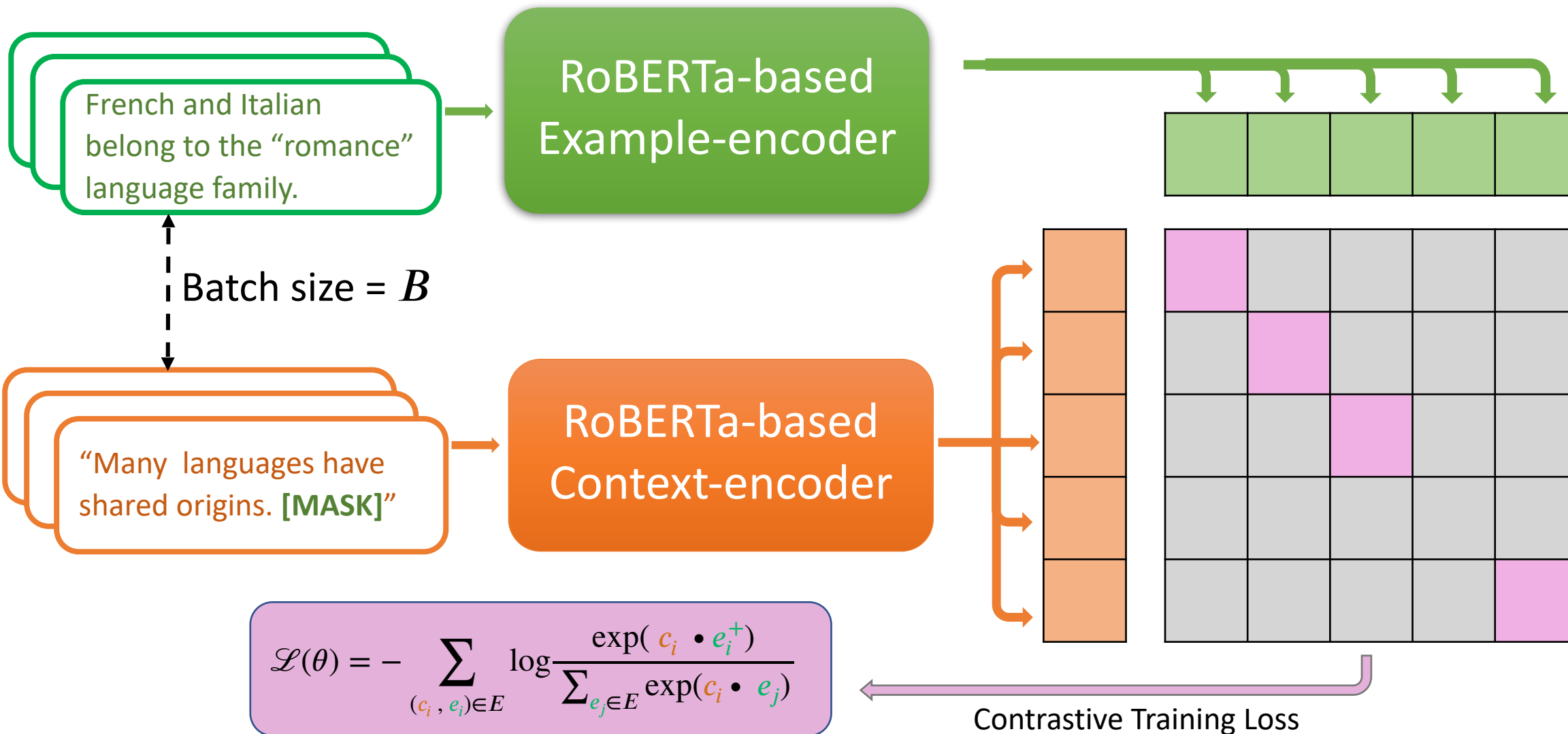
EGRET is trained with contrastive learning, using in-batch negatives



EGRET is trained with contrastive learning, using in-batch negatives



EGRET is trained with contrastive learning, using in-batch negatives



Evaluating example retrievals with recalls@k

- EGRET outperforms pretrained / non-parametric baselines,
- Exemplification cannot be solved by simple query-context similarity matching

		Recall@K					
	context	K=1	K=3	K=5	K=10	K=50	K=100
Random	L	0.0	0.0	0.01	0.02	0.08	0.15
BM25	L	4.6	9.5	12.1	16.2	25.6	30.4
DPR	L	2.7	5.2	7.1	9.7	20.3	27.5
CoBERT	L	6.0	11.8	14.3	18.2	31.2	36.3
SBERT	L	5.7	11.6	15.0	20.4	34.3	42.2
EGRET	L	13.0	22.8	29.3	36.5	55.2	64.0

Out-of-domain pretraining improves retrievals

- Books3 from the Pile consists of 200k books, out of which we extract 3.5m context-example pairs
- Pretraining with large-scaled out-of-domain Books3 data boosts EGRET's performance significantly

Recall@K

	Context	K=1	K=3	K=5	K=10	K=50	K=100
EGRET (ELI5)	L	13.0	22.8	29.3	36.5	55.2	64.0
EGRET (Books3 only)	L	19.3	30.4	36.8	44.1	63.1	69.0
EGRET (Books3 + Eli5)	L	21.1	33.5	39.2	46.8	66.7	73.0

Compare retrievals vs generations

Crowd workers' **ranking** on examples produced by models

	Ranking _{STD} (↓)	Krippendorff's
Generative Approach (c-REALM-RT)	2.26 _{0.271}	0.168
Retrieval Approach (EGRET)	1.88 _{0.252}	0.154
Ground Truth (Human-written)	1.71 _{0.284}	0.200

- Human readers show a clear preference for exemplifying units retrieved by EGRET compared to those generated by c-REALM-RT
- Exemplification may be better handled by retrieval models than generative models as of now

C-REALM-RT struggles at generating relevant and informative examples

Question: What prime numbers are and why they're important

Answer: A prime number is a number that can be divided only by 1 and by itself. ... For example,

Ground-truth example

4 is not prime because $4 = 2 \times 2$. $6 = 3 \times 2$, $9 = 3 \times 3$, on the other hand 17 is prime because $17 = 17 \times 1$.

EGRET-retrieved example:

For example, 2, 3, 5, 7, and 11 are the first primes.

c-REALM-RT-generated example:

every number has a unique, simple form.



!

!

C-REALM-RT struggles at generating relevant and informative examples

Often the C-REALM-RT model generates uninformative but common / generic expressions in in the ELI5 dataset:

- *“this is going to be something I could not ELI5 well.”*
- *“this is the better question for /r/linguistics.”*

High ROUGE ≠ Meaningful example

Question: Why is it we go into a state of "feeling low" for no apparent reason sometimes and how does it pass away?

Answer: Small bouts of depression are also thought to have an evolutionary origin. I wish I could make this clearer, but if you think about it, it's pretty a pretty simple scenario. For example,

Ground-truth example

you have a **situation**/problem in your life, you get down on yourself about it, you **think** about it, you change/objectively look at it, and viola it's not a problem/unsolved situation anymore.

c-REALM-RT-generated example, 17.1 Rouge-L

if you're referring to biological depression like mine, I don't **think** that we've any real idea, and I **think** it's one of those **situations** in which we don't really know

!

EGRET produces examples that human readers prefer, despite lower ROUGE

Question: Why is it we go into a state of "feeling low" for no apparent reason sometimes and how does it pass away?

Answer: Small bouts of depression are also thought to have an evolutionary origin. I wish I could make this clearer, but if you think about it, it's pretty a pretty simple scenario. For example,

Ground-truth example

you have a **situation**/problem in your life, you get down on yourself about it, you **think** about it, you change/objectively look at it, and viola it's not a problem/unsolved situation anymore.

c-REALM-RT-generated example, 17.1 Rouge-L

if you're referring to biological depression like mine, I don't **think** that we've any real idea, and I **think** it's one of those **situations** in which we don't really know

EGRET-retrieved example, 11.3 Rouge-L

if not getting mates or not getting enough food or other problems are facing the homo sapien, they could enter a state of depression, change their ways or solve their problem and promote their evolutionary fitness.

!



Conclusion

In summary,

- We conduct a detailed study of exemplification in LFQA and across datasets of different domains (ELI5, NaturalQA, Books from the Pile)
- We treat modeling (and evaluation) of exemplification in LFQA as a retrieval problem, instead of a generation problem, and proposed EGRET
- The retrieval approach allows us to produce more meaningful examples and to use more informative evaluation metrics than ROUGE

Future work:

- Modeling and evaluating other fine-grained linguistic / discourse phenomena in LFQA, e.g., hypotheticals, analogies, personal anecdotes
- More intelligent hybrid retrieval + generation approach for LFQA