

# Learning Opinion Summarizers by Selecting Informative Reviews

**Arthur Bražiņskas**, Mirella Lapata, Ivan Titov  
The University of Edinburgh, Scotland

**EMNLP 2021**





# Opinion Summarization

# Customer Reviews

- Users often purchase items online (e.g., from Amazon)
- Seek **opinions** of other **users** expressed in **reviews**
- Use this information for **better purchasing decisions**

# Amazon Customer Reviews





All-new  
**echo show 8**  
13 MP camera with auto-framing

amazon music | spotify | apple music | prime video | netflix | photos | sky news | ring

Roll over image to zoom in

All-new Echo Show 8 | 2nd generation (2021 release), HD smart display with Alexa and 13 MP camera | Charcoal

Brand: Amazon

★★★★★ 1,929 ratings | 139 answered questions

Amazon's Choice for "echo show 8 2nd generation"


Climate Pledge Friendly

RRP: ~~£119.99~~  
Deal Price: **£89.99**  
You Save: **£30.00 (25%)**


Receive a **60-day free Audible trial** when you purchase this Echo. Eligibility requirements apply. [Click here to learn more](#)

**Note:** This item is eligible for **FREE Click and Collect** without a minimum order subject to availability. [Details](#)

Pick a version [See the differences](#)







8" HD smart display (2nd generation)  
All-new Echo Show 8 (2nd generation)  
★★★★★ 1,929  
From: £89.99



5" smart display (2nd generation)  
All-new Echo Show 5 (2nd generation)  
★★★★★  
From: £

Colour Name: **Charcoal**


Configuration: **Device only**

Share    


**£89.99**

**FREE delivery:** Tuesday, Sep 28 in the UK [Details](#)

Fastest delivery: **Tomorrow**  
Order within 7 hrs 2 mins [Details](#)


 [Select delivery location](#)

**In stock.**

Quantity: 1 

[Add to Basket](#)

[Buy Now](#)

 **Secure transaction**

Dispatches from Amazon EU Sarl  
Sold by Amazon EU Sarl  
Packaging Item arrives in packagin...

[Details](#)

Return policy: [This item is returnable](#)

[Save 25% with Trade-In](#)

☐ Yes, I want a free trial with **FREE Premium Delivery** on this order.  
[amazon prime](#)

1-Click ordering is not available for this item.

☐ This will be a gift



# Amazon Customer Reviews



jimmy

★★★★★ **It works**

Reviewed in the United Kingdom on 11 June 2021

Colour Name: Glacier White | Configuration: Device only | **Verified Purchase**

My first intention was to use this Echo Show device in the kitchen for recipes and also weather and radio, I am still experimenting with it, so far I am pleased with what it can do.

18 people found this helpful

Helpful

Report abuse



Mr. Brian R. Dougal

★★★★☆ **Fine as an Alexa device, but display is hopeless**

Reviewed in the United Kingdom on 11 June 2021

Colour Name: Charcoal | Configuration: Device only | **Verified Purchase**

Fine as an Alexa device, works just as well (or badly ?) as my 3 Dot's.

But the display offers far less than I hoped.

Worst is the incredibly limited amount of customisation allowed. It does what it wants, NOT what I want it to do.

Video calling to another similar unit may well be good - but of no use to me and I guess most other UK purchasers.

Works well with Ring doorbell.

39 people found this helpful

Helpful

Report abuse

# Challenge

- Some products have **thousands of reviews**
- Reading them is **time consuming**
- Automatic summarization can compress and fuse **opinions** to short texts
- Helps the user to make **faster** and **better** decisions

# Summarization

- There are two types of summarization systems:
  - extractive
  - abstractive

# Extractive Summarizers

- Mostly **unsupervised** or **weakly-supervised** (Ganesa et al 2010; Angelidis and Lapata, 2018; Isonuma et al. 2019)
- Select **summarizing input fragments**
- Concatenate to form a summary
- Can be **incoherent** and contain **unimportant details**

# Abstractive Summarizers

- Generate text with a **richer vocabulary** of words (Paulus et al. 2017; See et al. 2017; Liu et al., 2018)
- Can **compress** and **fuse** (Lebanoff et al., 2019)
- Can deal with **conflicting information**

# Challenge

- Supervised methods often require **large annotated datasets for training**
- Datasets in the domain are **very scarce**

# Available Datasets

	#Entities	#Summaries	Domain
MeanSum (Chu and Liu, 2019)	200	200	Yelp
Copycat (Bražinskas et al., 2020)	60	180	Amazon
FewSum (Bražinskas et al., 2020)	60	180	Amazon
SpaCe (Angelidis et al., 2020)	50	1,050	TripAdvisor

# Unsupervised Abstractive Methods

- **MeanSum** (Chu and Liu, 2019)
- **Copycat** (Bražiņskas et al. 2020)
- **OpinionDigest** (Suhara et al. 2020)
- **DenoiseSum** (Amplayo et al., 2020)
- **SelfSum** (Elsahar et al., 2020)
- **RecurSum** (Isonuma et al., 2020)
- **MultimodalSum** (Im et al., 2021)
- ...



# Low-resource Methods

- **FewSum** (Bražiņskas et al. 2020)
- **PASS** (Oved and Levy, 2021)

# Contributions

- We provide the **largest dataset** for multi-document abstractive opinion summarization
- A novel model that **selects** and **summarizes** reviews from large collections **end-to-end**

# AmaSum

# AmaSum

- More than **33,000 summaries** for more than **31,000** Amazon products
- Each paired with more than **320 reviews**, on average
- Human-written by **professional product reviewers**
- Extracted from popular web portals

# AmaSum

	# Entities	Rev/Ent	# Summaries	Domain
<b>AmaSum (this work)</b>	<b>31,483</b>	<b>326</b>	<b>33,324</b>	<b>Amazon</b>
SpaCe (Angelidis et al., 2020)	50	100	1,050	Tripadvisor
Copycat (Bražiński et al., 2020)	60	8	180	Amazon
FewSum (Bražiński et al., 2020)	60	8	180	Amazon
MeanSum (Chu and Liu, 2019)	200	8	200	Yelp

# AmaSum

- Summaries consist of:
  - Verdicts
  - Pros and cons

# Example



Olympus E-500 EVOLT

# Verdict

The Olympus Evolt E-500 is a compact, easy-to-use digital SLR camera with a broad feature set for its class and very nice photo quality overall.



# Pros

- Compact design
- Strong autofocus performance
- Intuitive and easy-to-navigate menu system

# Cons

- Unreliable automatic white balance
- Slow start-up time when dust reduction is enabled

# Challenges

- Each summary is paired with more than **320 reviews**, on average
- Standard encoding-decoding can be challenging
- Not all **reviews content** covers the **summary content**
- Training on **random review subsets** leads to **hallucinations in test time** (show in this work)
- We address these challenges by introducing SelSum

# SelSum

# SelSum

- A probabilistic latent model that **selects** and **summarizes** reviews end-to-end
- Learns to select **subsets** of **summary relevant reviews** in training

# Review Selection



$r_1$

...

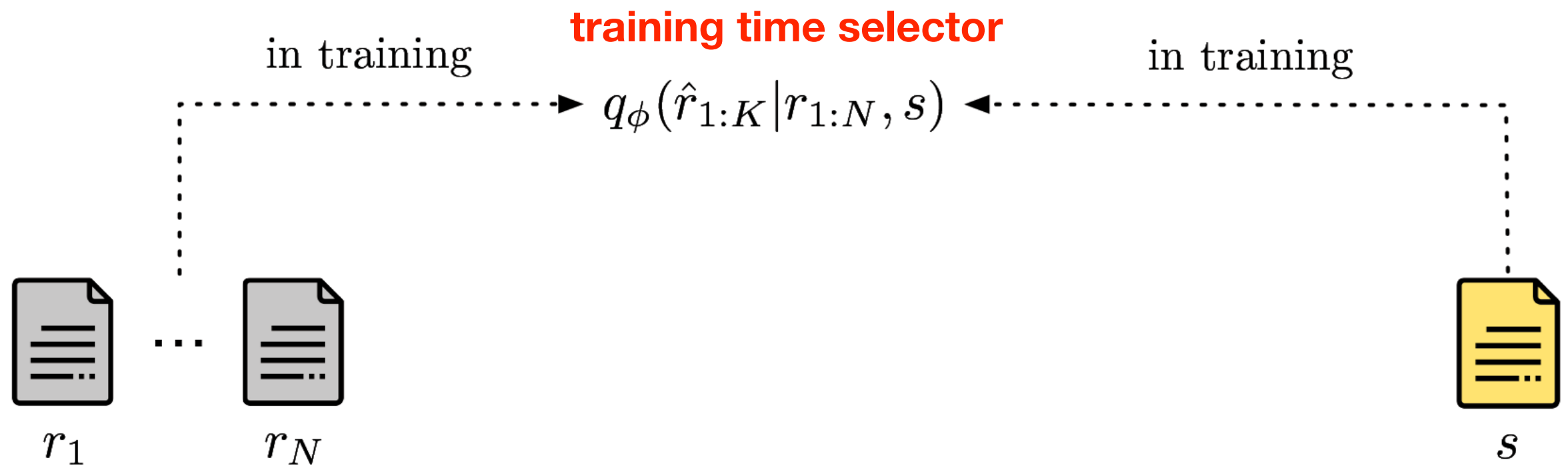


$r_N$

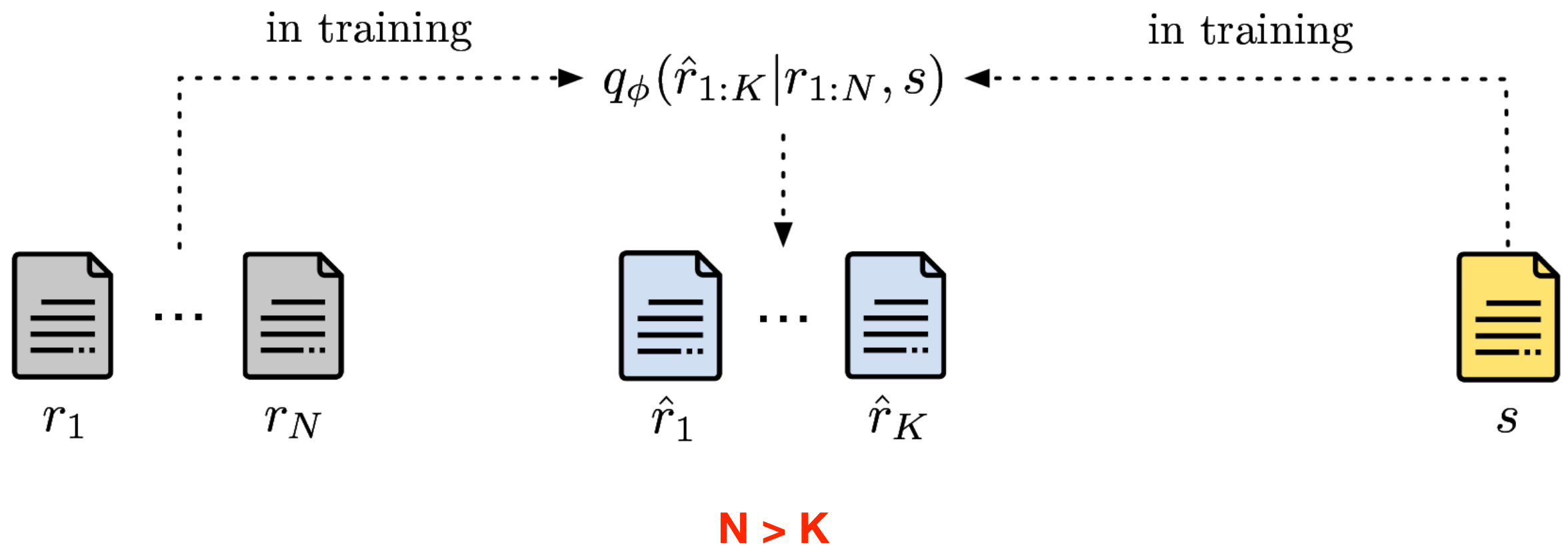


$s$

# Review Selection

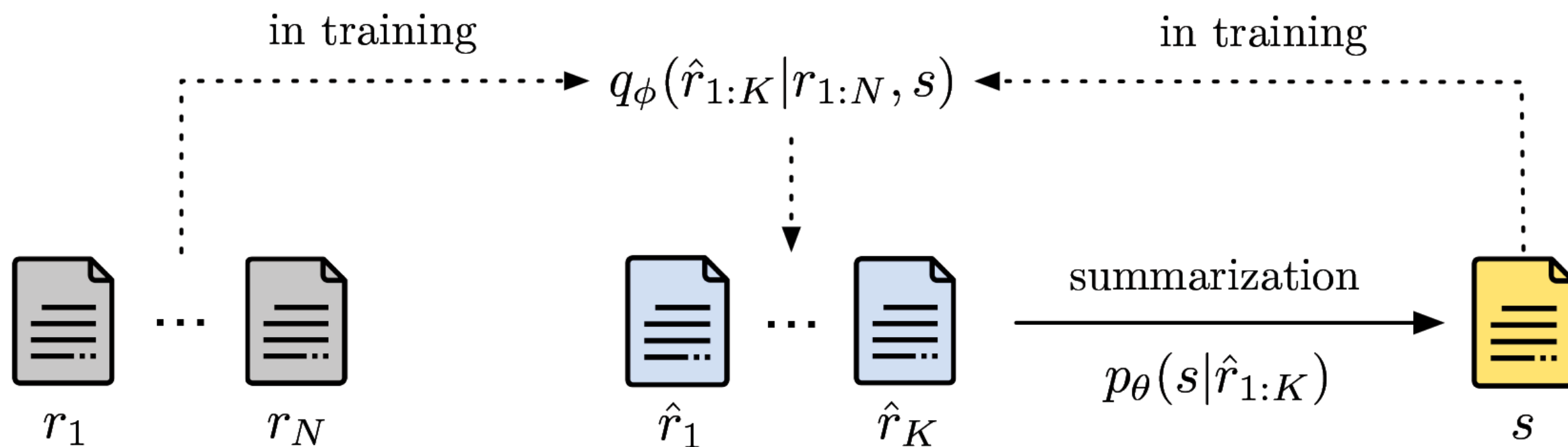


# Review Selection





# Review Selection



# Training Time Selector

- Review subsets are treated as vectors of **categorical variables** (K slots)
- **Sampling without replacement**

# Training Time Selector



$s$



$r_1$



$r_2$



$r_3$



$r_4$



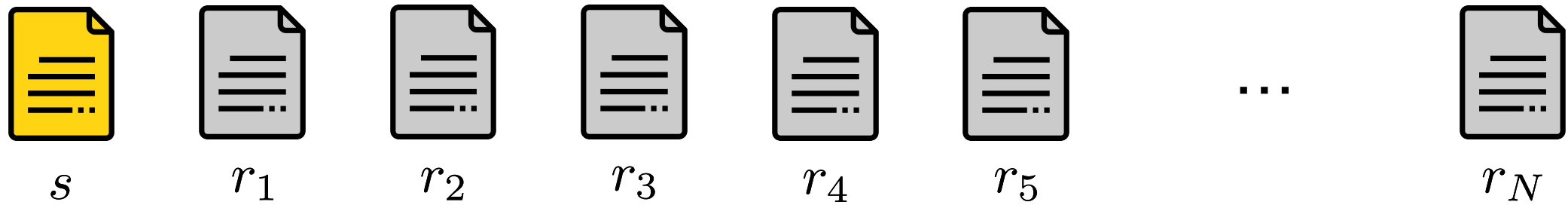
$r_5$

...



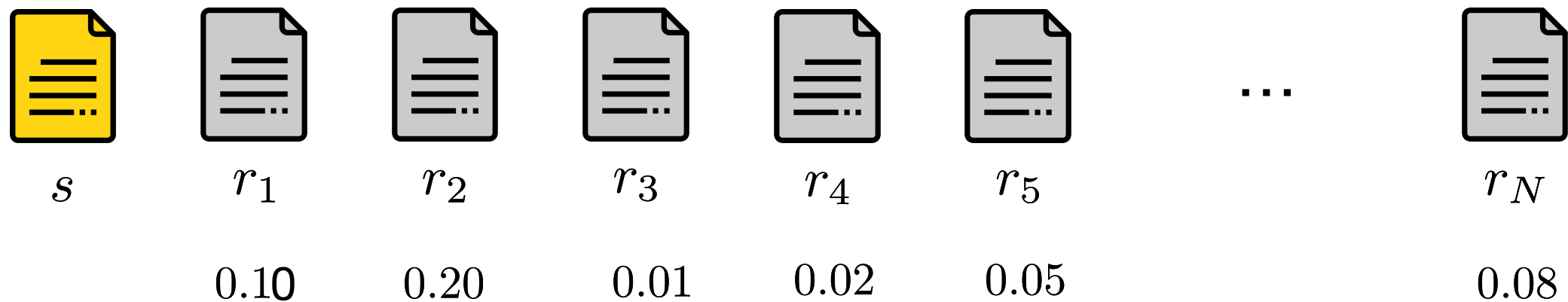
$r_N$

# Training Time Selector



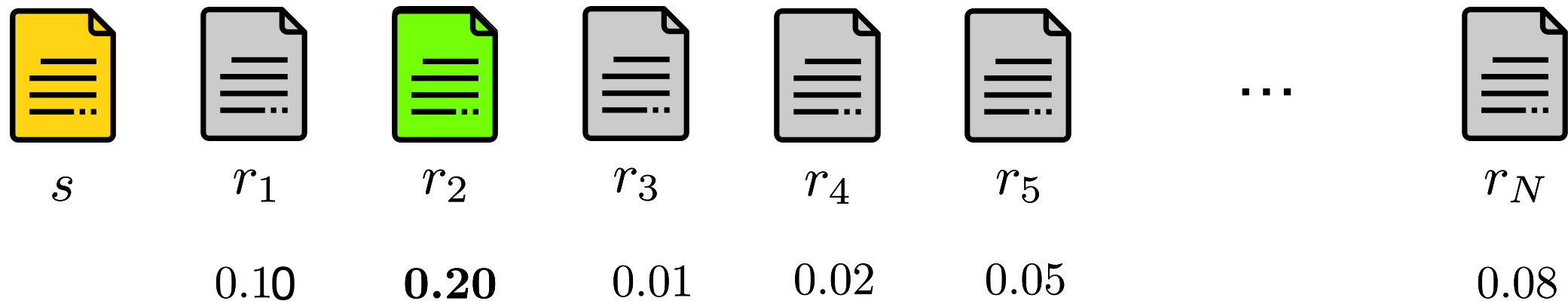
$$q_{\phi}(\hat{r}_1 | r_{1:N}, s)$$

# Training Time Selector



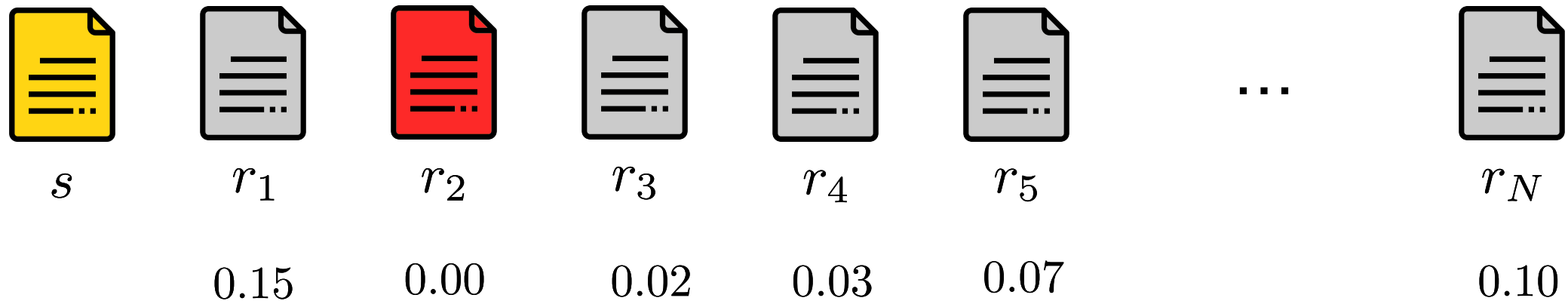
$$q_{\phi}(\hat{r}_1 | r_{1:N}, s)$$

# Training Time Selector



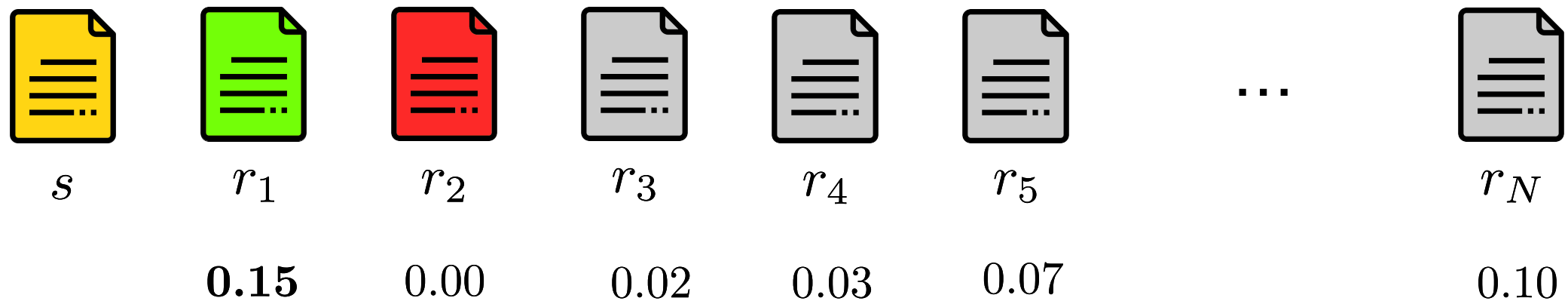
$$\hat{r}_1 \sim q_\phi(\hat{r}_1 | r_{1:N}, s)$$

# Training Time Selector



$$q_{\phi}(\hat{r}_2 | r_{1:N}, \hat{r}_1, s)$$

# Training Time Selector



$$\hat{r}_2 \sim q_\phi(\hat{r}_2 | r_{1:N}, \hat{r}_1, s)$$



# Model Training

- Sampling categorical variable assignments is **not differentiable**
- To train the selector and summarizer **end-to-end** we use:
  - Amortized variational inference (Kingma and Welling, 2013; Cremer et al., 2018)
  - REINFORCE (Williams, 1992)

# Review Selection

- **Computational and memory savings**
  - Only the subset is encoded using the deep encoder
- Better **interpretability** of the generated output
- **Fewer hallucinations** (as we show)

# Lexical Features

- Training time selector inputs **review representations**
- Represent each review in the collection with **pre-computed 23 features**
- Feed to a tiny non-linear neural network ( $< 0.1\%$  params of the model)
- Minimal computational burden in training

# Feature Examples

- ROUGE scores between a **review** and **summary**
- ROUGE scores between a **review** and the **other ones in the collection** (measures uniqueness)
- Aspect keyword-based scores
  - Used a vocabulary of aspect keywords
  - Counted their occurrence in reviews and summaries
  - Computed recall and precision scores
- ...

# Test Time Selector

- In test time would like to select and summarize **informative reviews**
- Can't use the **training time selector**
  - summary is **not available** in **test time**
  - fit a **test time selector** that relies only on reviews (Razavi et al., 2019)

# Test Time Selector

- Select reviews using the **training time selector**
- Fit the **test time selector** to predict the selected reviews

# Test Time Selector



$r_1$



$r_2$



$r_3$



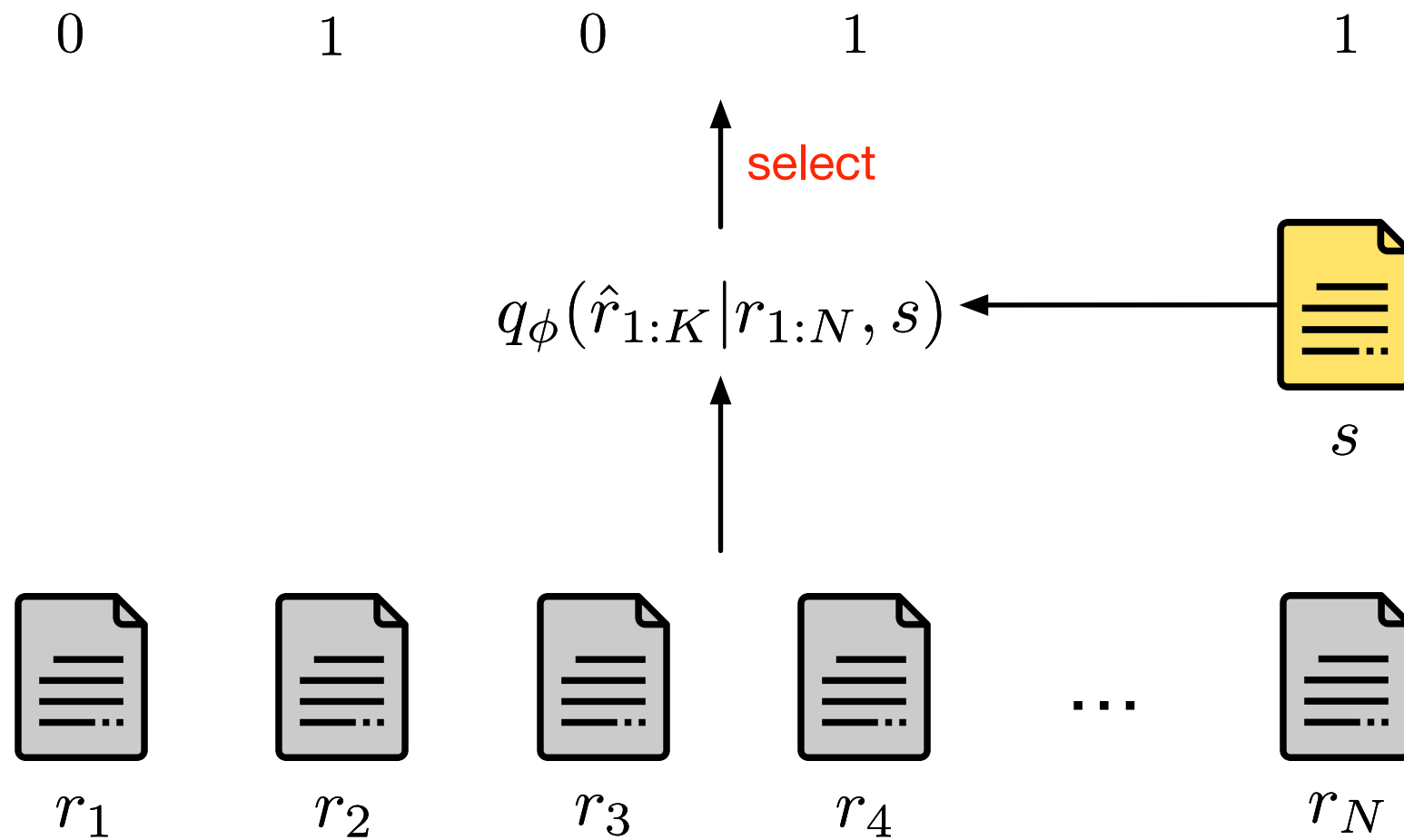
$r_4$

...



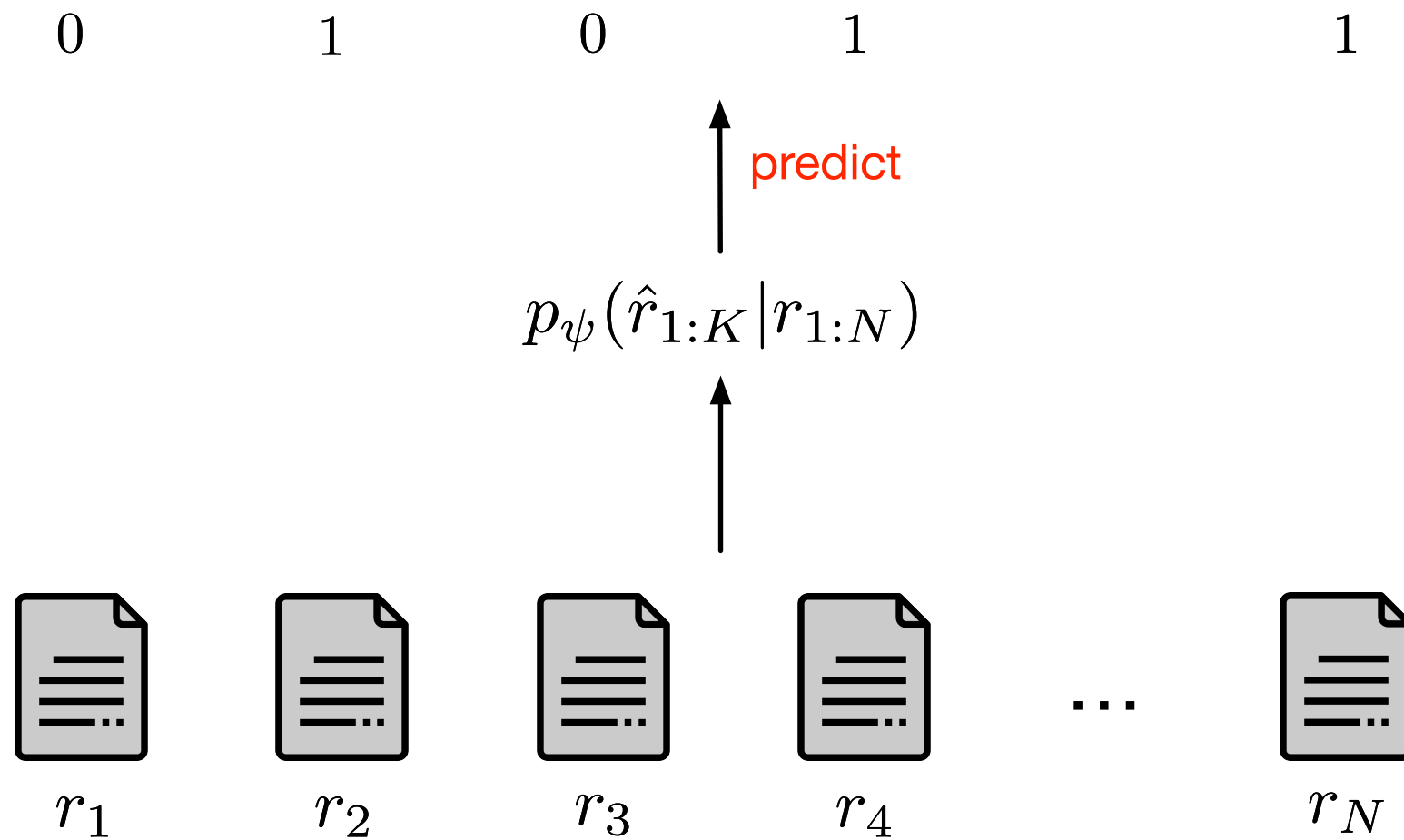
$r_N$

# Test Time Selector

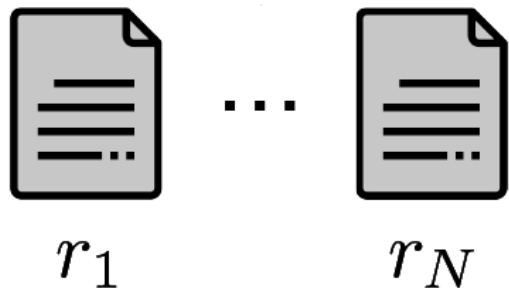




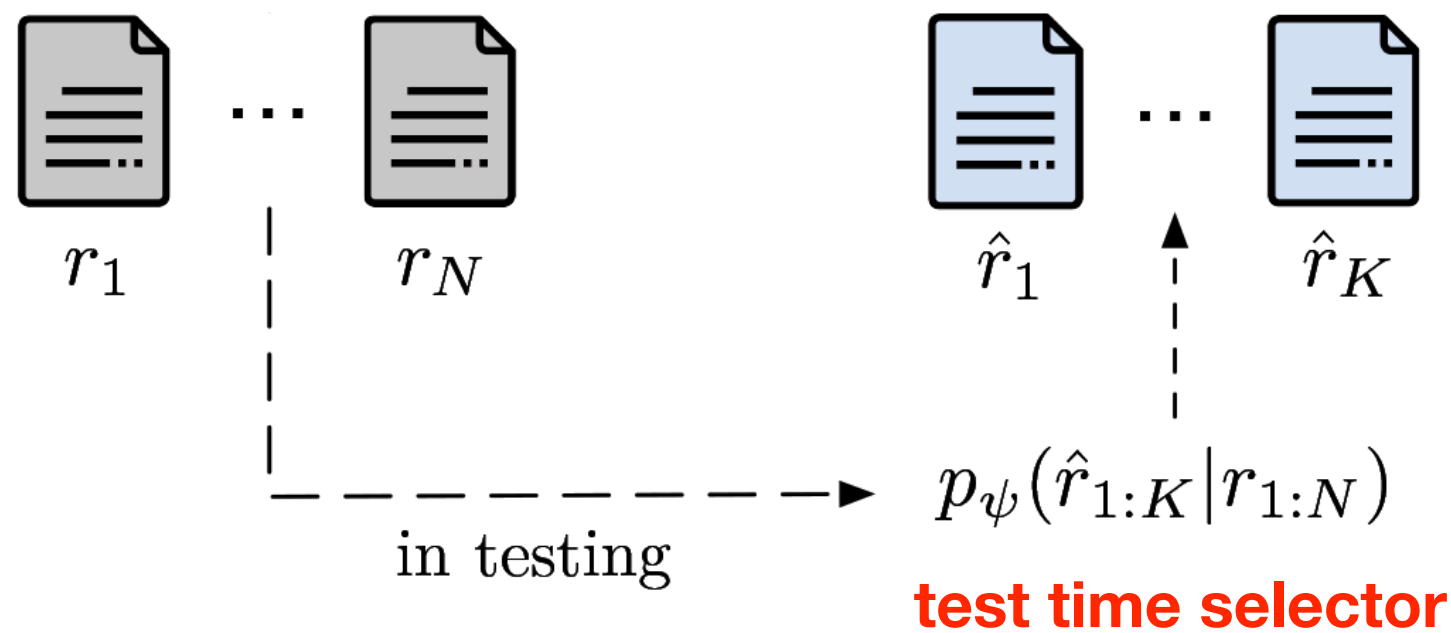
# Test Time Selector



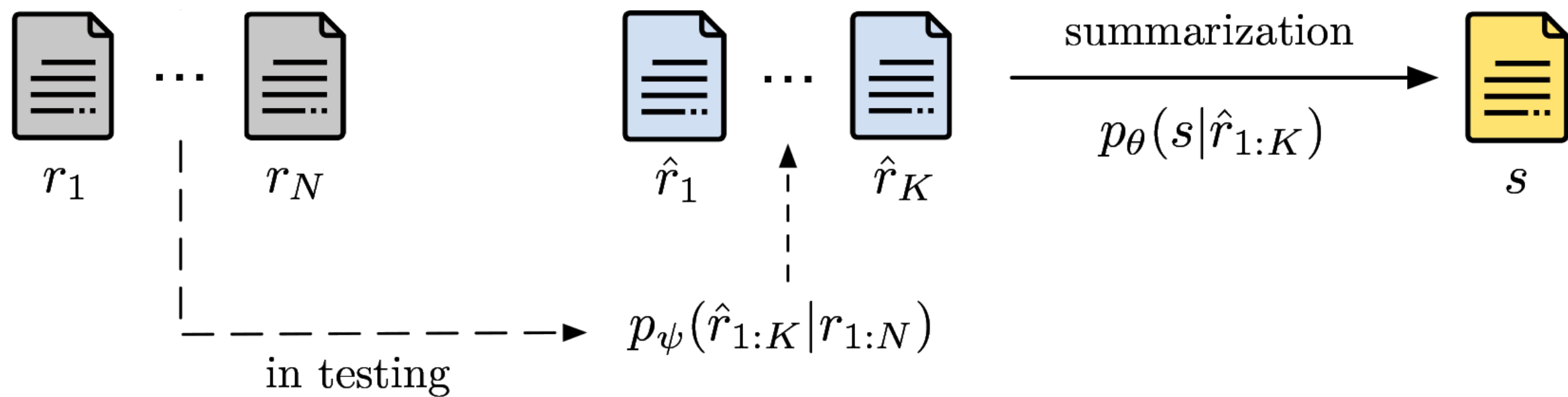
# Test Time Selector



# Test Time Selector



# Test Time Selector



# Setup and Results

# Splits

- **Training:** 26,660 summaries
- **Validation:** 3,302 summaries
- **Testing:** 3,362 summaries

# Summarizer

- Pre-trained BART (Lewis et al, 2020) **encoder-decoder**
- Verdicts, pros and cons were **concatenated** together as one string

# Training Time Selector

- Feed-forward network inputting static features
- Selecting **10** out of **100** reviews



# Test Time Selector

- Pre-trained BART encoder on the end-task to **represent reviews**
- Feed-forwards to tag reviews

# Baseline Models

- **Random**: random sentences from reviews
- **Oracle**: greedy selection of sentences with maximum ROUGE-1 and -2 scores to the summary
- **LexRank** (Erkan and Radev, 2004): unsupervised extractive
- **MeanSum** (Chu and Liu, 2019): unsupervised abstractive
- **Copycat** (Bražiņskas et al, 2020): unsupervised abstractive
- **ExtSum** (ours): supervised extractive summarizer

# Review Selectors

- Experimented with review selectors (**non-learned**)
- **RandSel:**
  - Random selection of reviews
- **R1 top-K:**
  - **K highest scored** reviews based on ROUGE-1 with respect to the **summary**
  - Before test time, fit the test time selector

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44



# Automatic Evaluation

	<b>Verdict</b>			<b>Pros</b>			<b>Cons</b>		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	22.01	3.94	19.84	14.93	2.57	12.96

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	22.01	3.94	19.84	14.93	2.57	12.96
SELSUM	24.33	5.29	18.84	21.29	4.00	19.39	14.96	2.60	13.07

# Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	<b>22.01</b>	3.94	<b>19.84</b>	14.93	2.57	12.96
SELSUM	<b>24.33</b>	<b>5.29</b>	<b>18.84</b>	21.29	<b>4.00</b>	19.39	<b>14.96</b>	<b>2.60</b>	<b>13.07</b>

# Content Support

- **ROUGE is not always reliable** for assessing how **input faithful** summaries are (Tay et al., 2019; Bražinskas et al., 2020)
- Generation of **input faithful** summaries is **crucial** for practical applications
- Remains an **open problem** (Maynez et al., 2020; Fabbri et al., 2020; Want et al., 2020)
- Performed **human evaluation** via Amazon Mechanical Turk (AMT)

# Content Support

- Evaluated different selectors
- Summarizer remained exactly the same

# Content Support

- Asked AMT workers to assess **faithfulness** of each summary sentence to input reviews by marking them as:
  - Fully supported
  - Partially supported
  - Not supported

# Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48



# Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89

# Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	<b>14.60</b>	70.48
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89
SELSUM	<b>66.08</b>	<b>25.15</b>	<b>8.77</b>	<b>70.21</b>	<b>17.99</b>	<b>11.80</b>	<b>38.41</b>	29.21	<b>32.38</b>

# Content Support

- Investigated the role of **better review subsets** in test time
- We selected reviews using the **SelSum's test time selector**
- Input them to the summarizer trained on **random review subsets (RandSel)**
- Indicated by \*

# Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
RANDSEL*	50.79	31.75	17.46	50.62	22.96	26.42	16.84	<b>13.75</b>	69.42

# Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
RANDSEL*	50.79	31.75	17.46	50.62	22.96	26.42	16.84	<b>13.75</b>	69.42
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89
SELSUM	<b>66.08</b>	<b>25.15</b>	<b>8.77</b>	<b>70.21</b>	<b>17.99</b>	<b>11.80</b>	<b>38.41</b>	29.21	<b>32.38</b>

# Take Away

- Random review subsets **might not cover well the content of summaries**
- A summarizer trained on these reviews **learns to hallucinate**
- Evident when better review subsets are provided in **test time**

# Wrap up

# Conclusions

- We contribute the **largest dataset** for **multi-document opinion summarization** (more than 33,000 summaries)
- Propose an end-to-end model **selecting** and **summarizing** reviews
- Show that learned review selection leads to generation of **input faithful summaries**



# Dataset and Codebase

Publicly available:

**<https://github.com/abrazinskas/SelSum>**

# Appendix

# Example Summary

---

**Verdict** If you like the idea of a **glass feeder**, this is the one to get. It has **a lot to offer for the price**.

---

- Pros**
- Has a **large opening** that makes it **easy to get in and out** of the feeder
  - Has a **nice design** that's **easy to clean**
- 

- Cons**
- The **lid is a little flimsy**, and it's **not as durable as some of the other models**
- 

**Reviews** ... looks just as nice as the **glass feeders** || ... Very happy with the **value, quality and function** ... || ... **the cheapest most flexible "jar"** I've ever seen ... || ... **Nice large opening** so it's easy to pour the sugar water || ... This feeder has a nice **large opening** ... || ... this is the **perfect design** and size ... || **The hummingbirds liked it and had no trouble feeding or perching....** || ... The main compartment is **easy to clean**... || ... **The top is a little flimsy** ... || ... **it fell out of the hanger it broke for good** ... there are so many other nice ones out there that have glass "jar's" or at least sturdier plastic ... || ... **The tray is easy to clean** ...

---