Quantifying and Applying Green IR

Harry Scells

University of Tübingen

28.04.2025

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Information access systems impact our environment

Information access systems impact our environment

What causes more emissions? Google search vs. a ChatGPT response

Information access systems impact our environment

What causes more emissions? Google search vs. a ChatGPT response

Data center emissions probably 662% higher than big tech claims. Can it keep up the ruse?

Emissions from in-house data centers of Google, Microsoft, Meta and Apple may be 7.62 times higher than official tally

Overview of Green IR

Measuring Utilisation

Corpus Subsampling

Emma Strubell et al. (2020). “Energy and Policy Considerations for Modern Deep Learning Research.”. In: *AAAI*, pp. 13693–13696

Why?

Large (pre-trained) neural language models, now LLMs

Why?

Large (pre-trained) neural language models, now LLMs

- Expend high energy for training and inference
  compared to traditional models

Why?

Large (pre-trained) neural language models, now LLMs

- Expend high energy for training and inference
  compared to traditional models
- The energy demands expected to continue growing
  as size and complexity of models increase

Why?

Large (pre-trained) neural language models, now LLMs

- Expend high energy for training and inference
  compared to traditional models
- The energy demands expected to continue growing
  as size and complexity of models increase
- Data centers and other infrastructure
  used to run these models also consume energy (and water[1])

[1] Guido Zuccon et al. (2023). “Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.”. In: *ICTIR*, pp. 283–289.

Emma Strubell et al. (2020). “Energy and Policy Considerations for Modern Deep Learning Research.”. In: *AAAI*, pp. 13693–13696

But what are emissions?

- **Energy**: amount of work done
  ➔ Measured in **joules**

But what are emissions?

- **Energy**: amount of work done
  ➔ Measured in **joules**

- **Power**: energy per unit time
  ➔ Measured in **watts**; 1 watt = 1 joule/second
  ➔ kWh: energy consumed at a rate of 1 kilowatt in 1 hour

But what are emissions?

- **Energy**: amount of work done
  ➔ Measured in **joules**

- **Power**: energy per unit time
  ➔ Measured in **watts**; 1 watt = 1 joule/second
  ➔ kWh: energy consumed at a rate of 1 kilowatt in 1 hour

- **Emissions**: by-products created by producing power
  Measured in $kgCO_2e$; kilograms of carbon dioxide equivalent

Emma Strubell et al. (2020). “Energy and Policy Considerations for Modern Deep Learning Research.”. In: *AAAI*, pp. 13693–13696

Retrieval Efficiency

Speed a system can retrieve relevant information in response to a query

Retrieval Efficiency

Speed a system can retrieve relevant information in response to a query

Factors that impact retrieval efficiency:

Retrieval Efficiency

Speed a system can retrieve relevant information in response to a query

Factors that impact retrieval efficiency:

- **Size and complexity** of the search corpus

Retrieval Efficiency

Speed a system can retrieve relevant information in response to a query

Factors that impact retrieval efficiency:

- **Size and complexity** of the search corpus
- Effectiveness of the **retrieval models** or techniques used

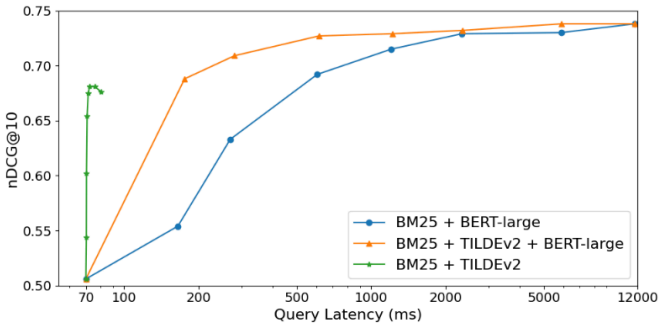
Retrieval Efficiency

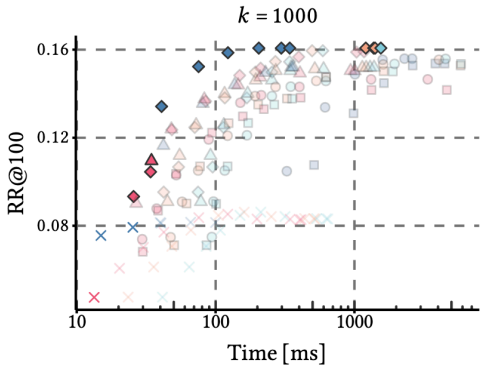
Speed a system can retrieve relevant information in response to a query

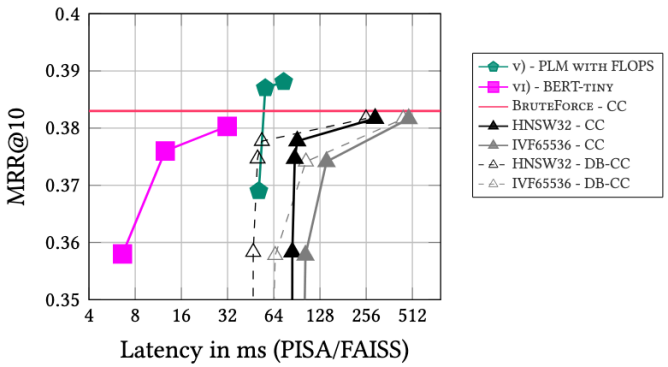
Factors that impact retrieval efficiency:

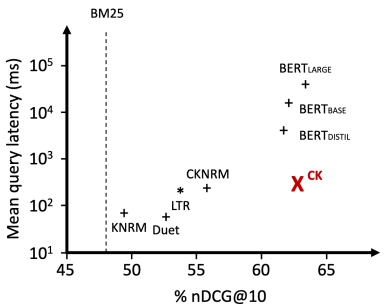
- **Size and complexity** of the search corpus
- Effectiveness of the **retrieval models** or techniques used
- Efficiency of the **hardware and infrastructure** used

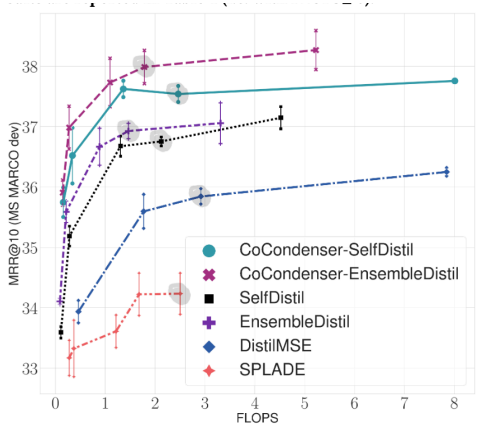
Retrieval Efficiency

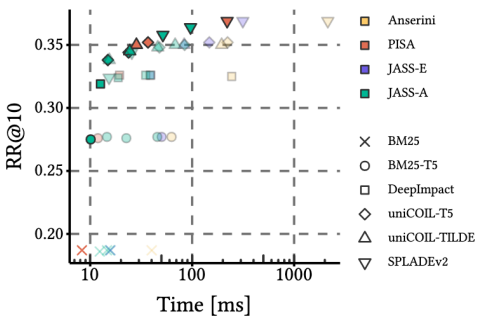


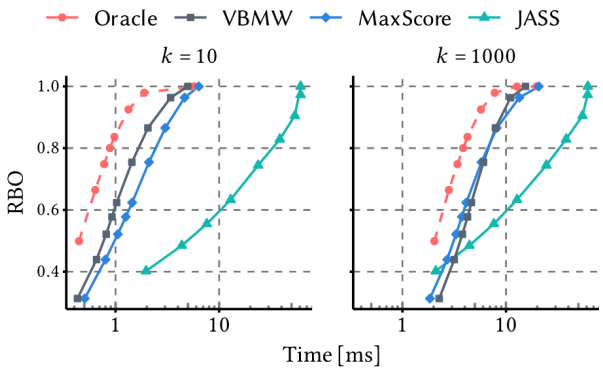




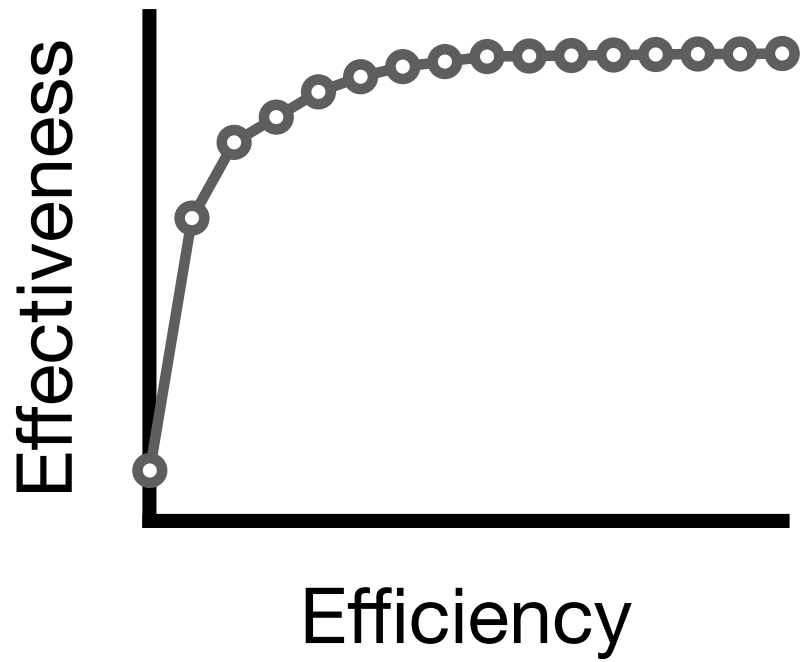




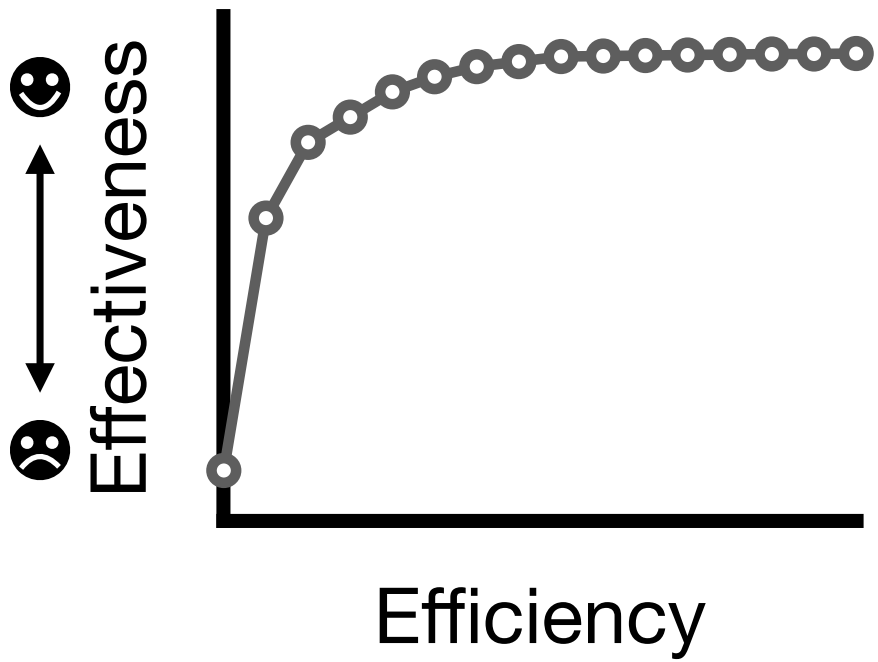




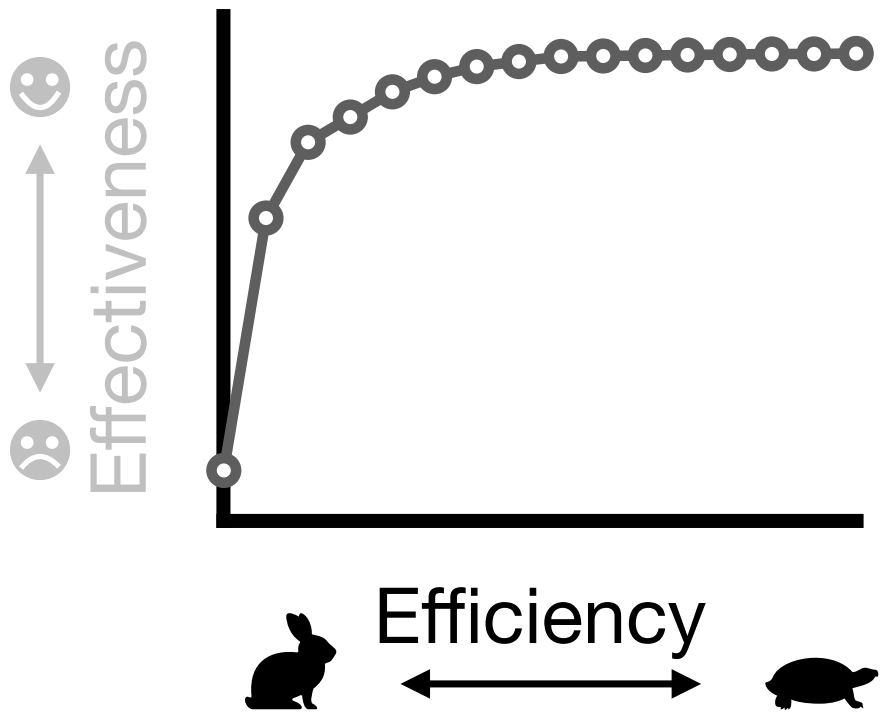
Retrieval Efficiency



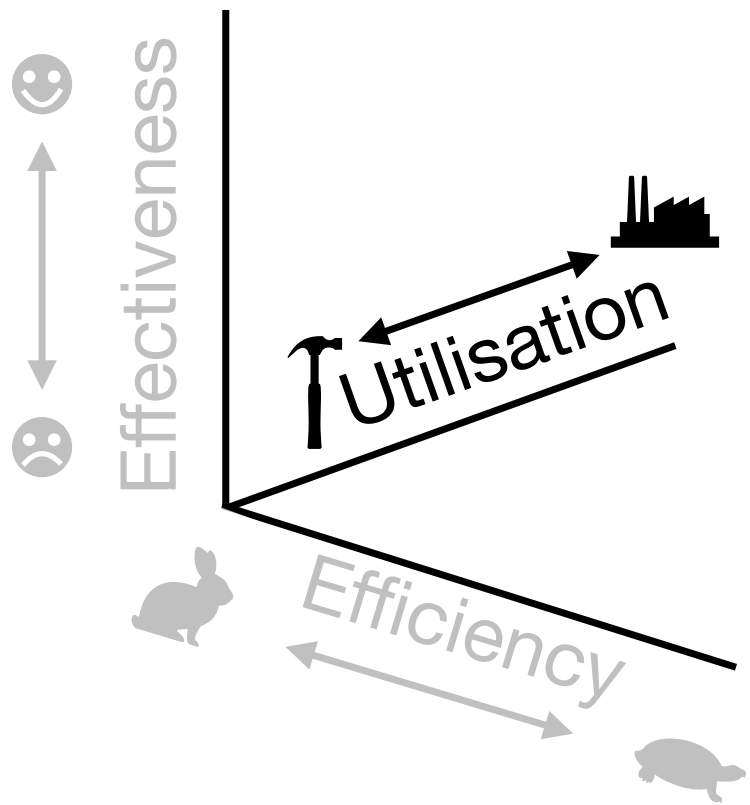
Retrieval Efficiency



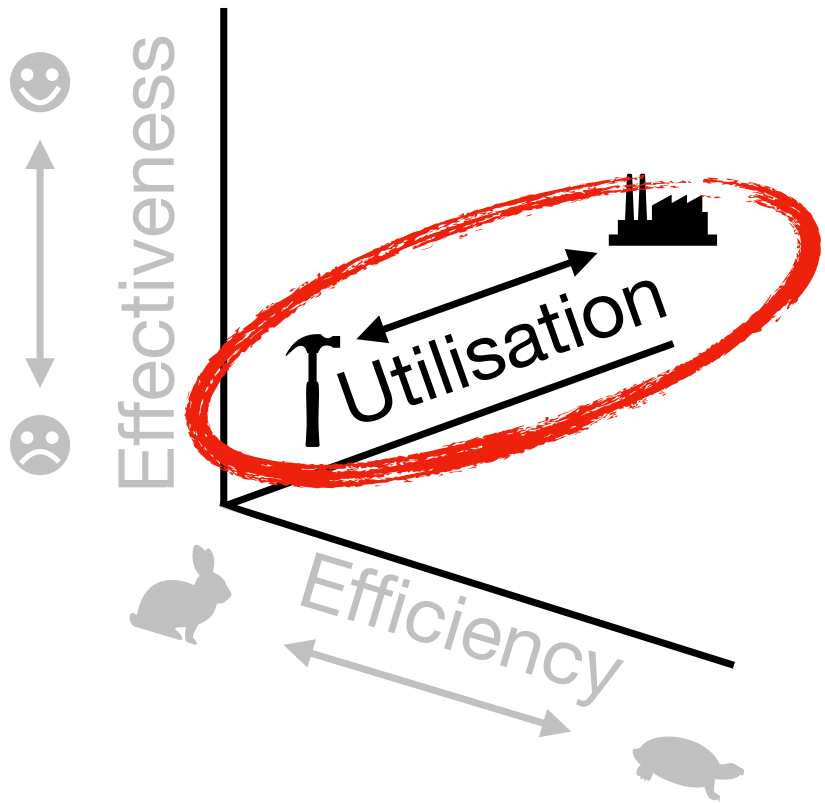
Retrieval Efficiency



Retrieval Efficiency

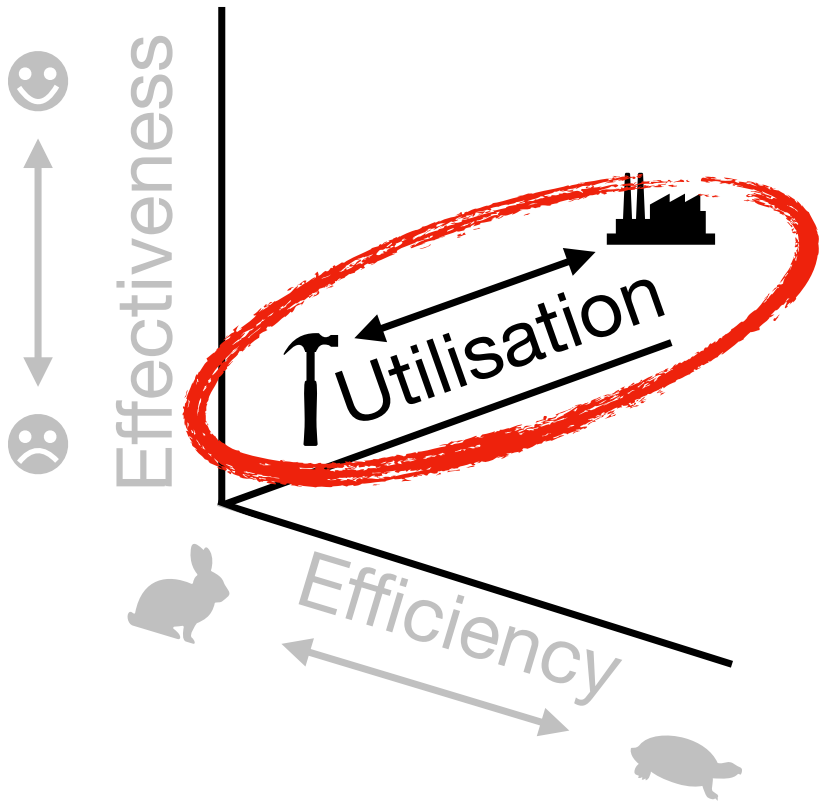


Retrieval Efficiency



Retrieval Efficiency

Okay, so what does this mean for IR?



Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*

Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

- **Orders of magnitude** more expensive to create and use

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment...

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment...

Missing dimension of IR evaluation:

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment...

Missing dimension of IR evaluation: effectiveness

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment...

Missing dimension of IR evaluation: effectiveness, efficiency

Utilisation and Green IR

Green IR is...

> *Research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.*
>
> Roy Schwartz et al. (2020). “Green AI.”. In: *Commun. ACM*, pp. 54–63

Neural methods require pre-trained LMs

- **Expensive** to create and use
- Have only become more expensive over time (e.g., GenIR methods)

Even more recently, LLMs used for IR

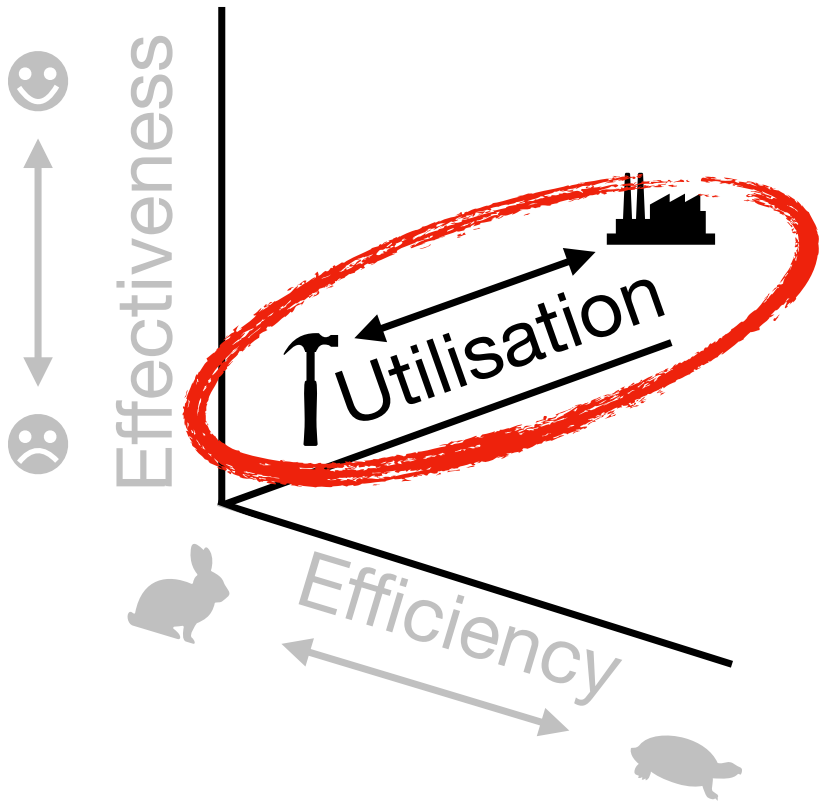
- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment...

Missing dimension of IR evaluation: effectiveness, efficiency, **utilisation**

Utilisation and Green IR

~~Okay, so what does this mean for IR?~~

Okay, so how can I measure this?



Overview of Green IR

Measuring Utilisation

Corpus Subsampling

Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:

$$p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000}$$

Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

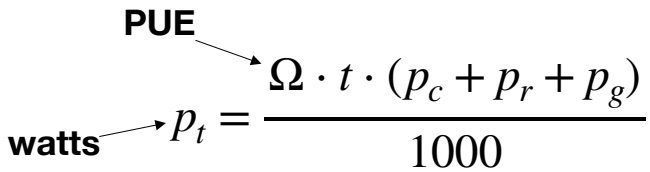
First, measure power consumption:

watts → $$p_t = \frac{\Omega \cdot t \cdot (p_c + p_r + p_g)}{1000}$$

Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

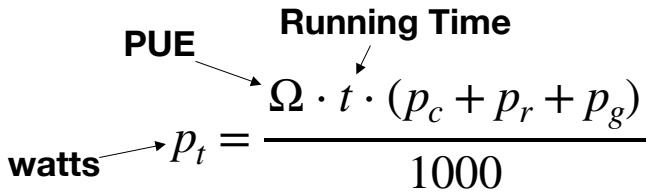
First, measure power consumption:



Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

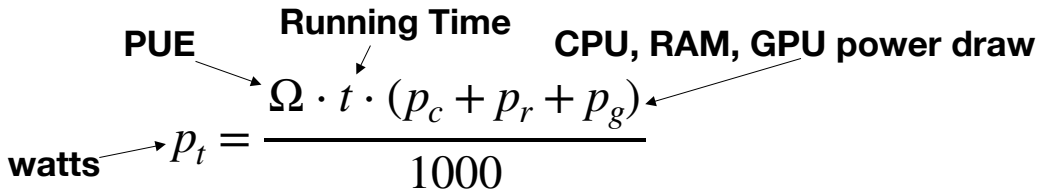
First, measure power consumption:



Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

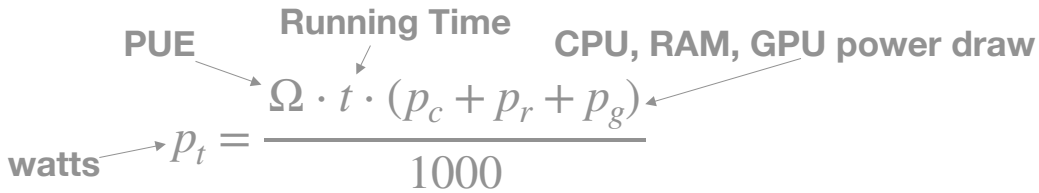
First, measure power consumption:



Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:

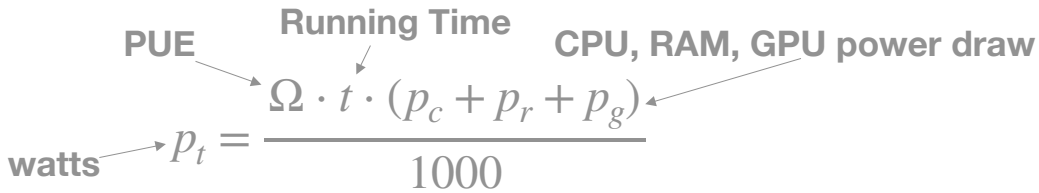


Next, measure emissions:

Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:



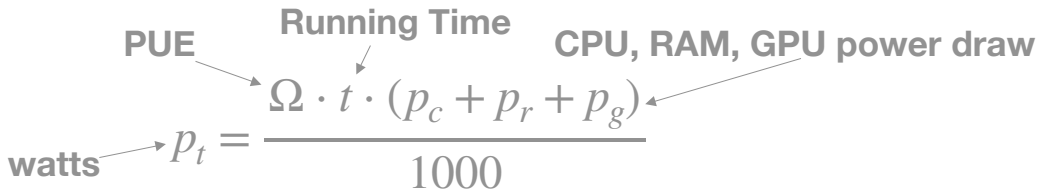
Next, measure emissions:

emissions → $\mathbf{kgCO_2e} = \theta \cdot p_t$

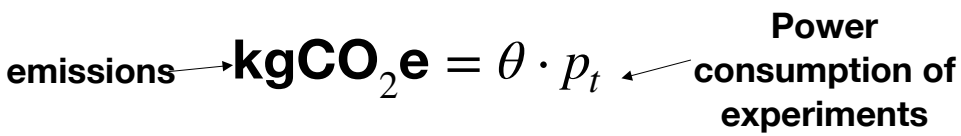
Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:



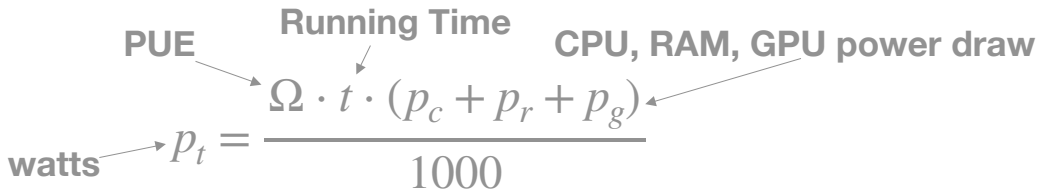
Next, measure emissions:



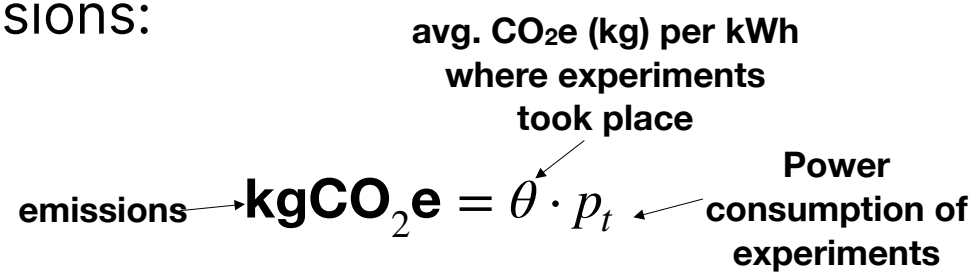
Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:



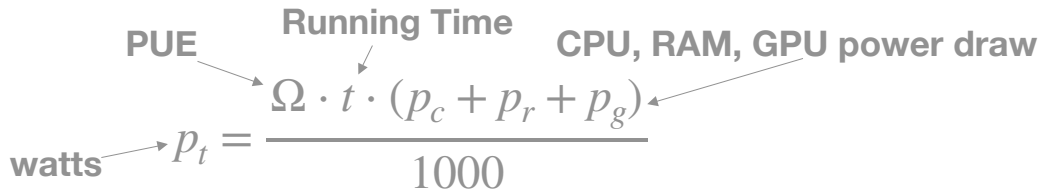
Next, measure emissions:



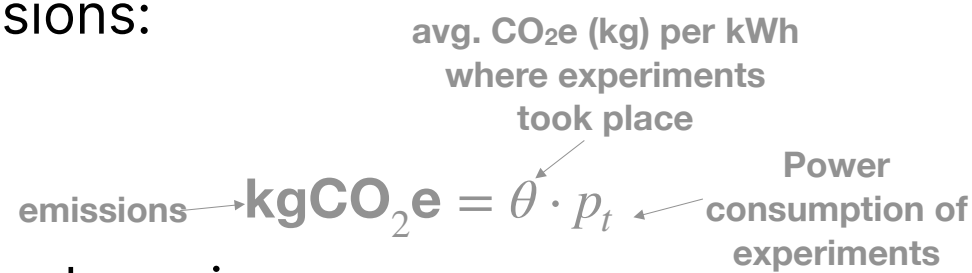
Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:



Next, measure emissions:



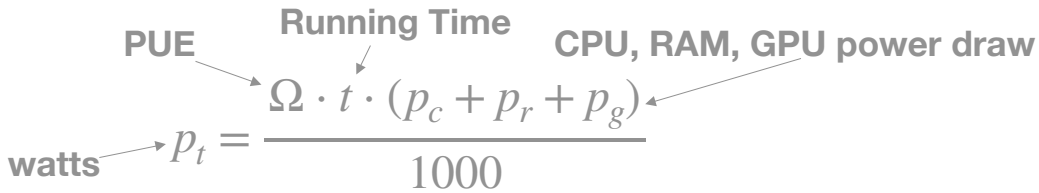
Emissions of my search engine:

$$\mathbf{kgCO_2e} = \theta \cdot \Delta_q \cdot p_q$$

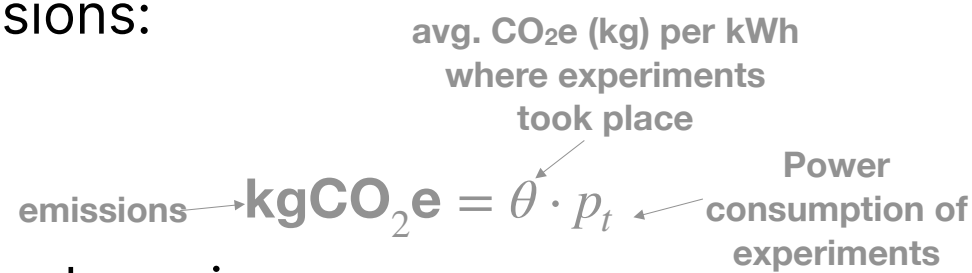
Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

First, measure power consumption:



Next, measure emissions:



Emissions of my search engine:

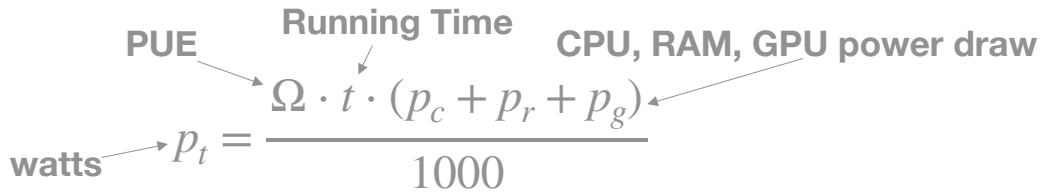
$$\mathbf{kgCO_2e} = \theta \cdot \Delta_q \cdot p_q$$

Power consumption of a single query

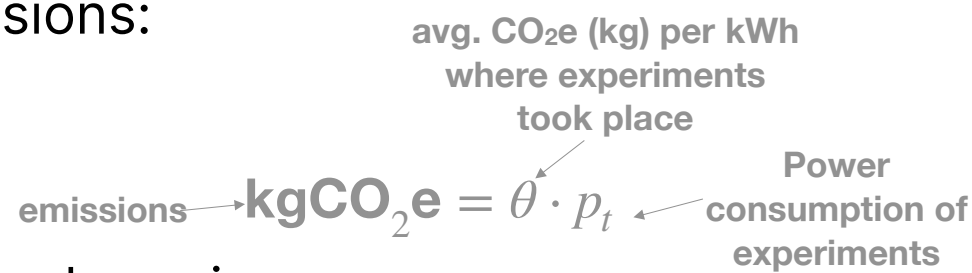
Measuring Energy/Emissions

Energy/emissions ➔ measures **direct** utilisation costs

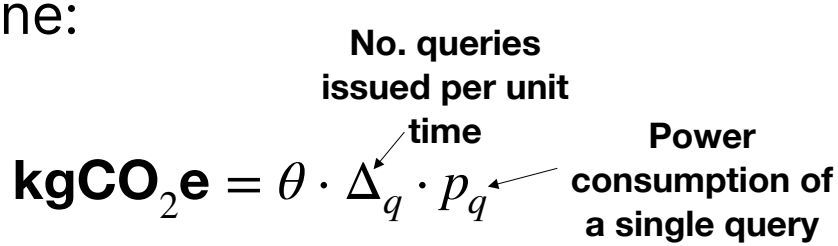
First, measure power consumption:



Next, measure emissions:

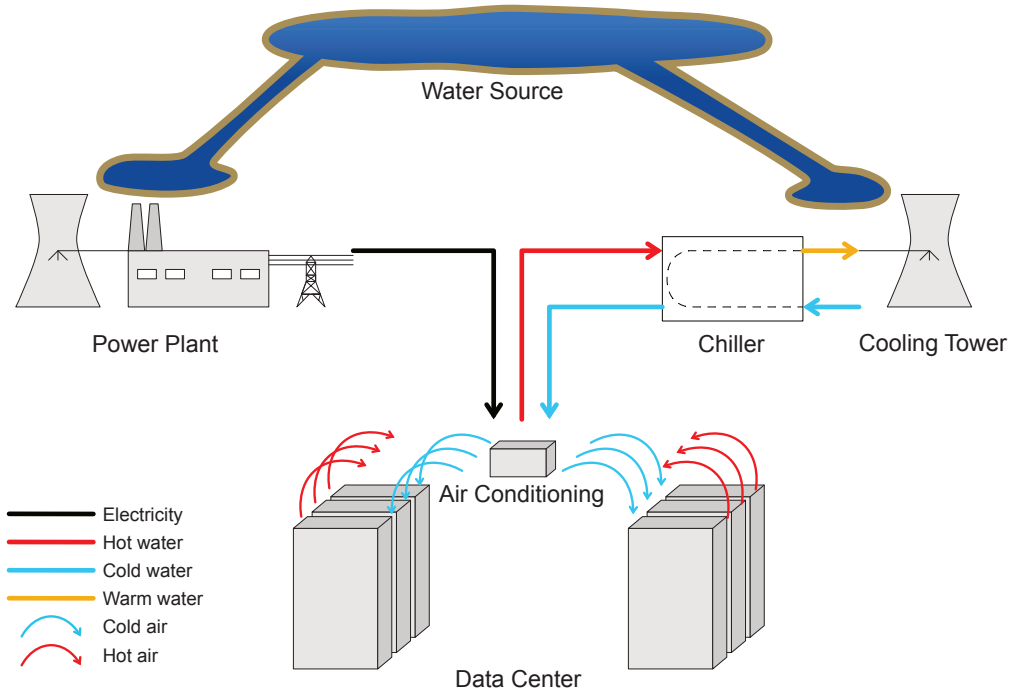


Emissions of my search engine:



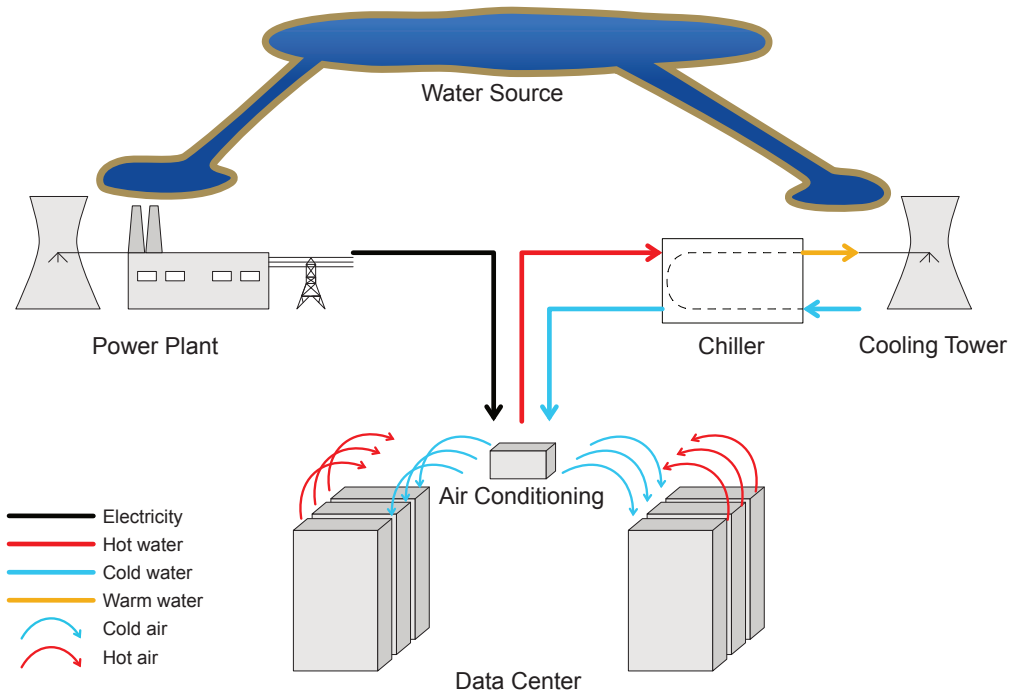
Measuring Water

Water → measures **indirect** utilisation costs



Measuring Water

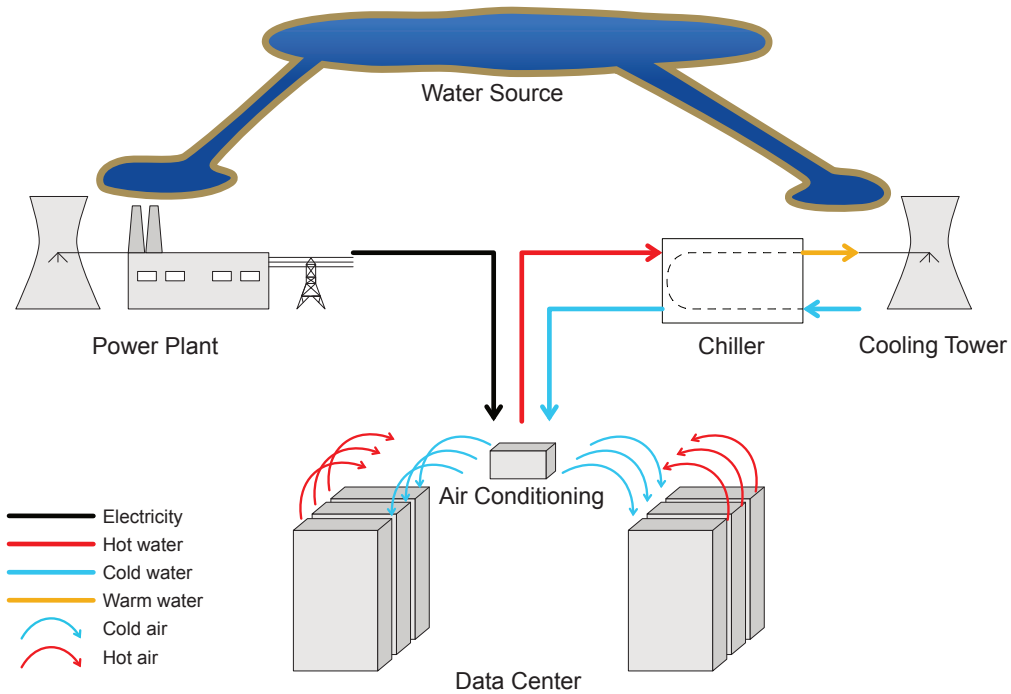
Water → measures **indirect** utilisation costs



In data centers, water is consumed through **evaporation** and **blow down**

Measuring Water

Water → measures **indirect** utilisation costs

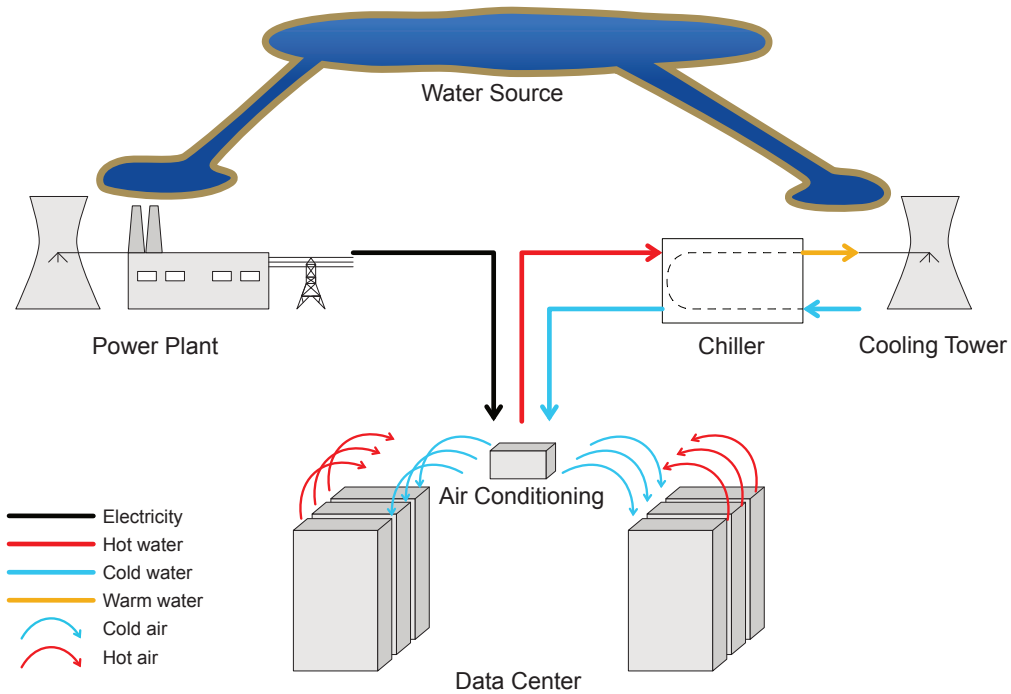


In data centers, water is consumed through **evaporation** and **blow down**

evaporation → inefficiency in chiller, **blow down** → flush water in system

Measuring Water

Water → measures **indirect** utilisation costs



In data centers, water is consumed through **evaporation** and **blow down**

evaporation → inefficiency in chiller, **blow down** → flush water in system

Water consumption of $\mathcal{M}$ → on-site cooling (W_{on}) and power plant (W_{off})

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Time

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Time

Energy used

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Time

Energy used

Water Usage Effectiveness[2]

[2] Guido Zuccon et al. (2023). “Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.”. In: *ICTIR*, pp. 283–289.

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Time

Energy used

Water Usage Effectiveness[2]

$$W_{off}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot PUE(t) \cdot WUE_{off}(t)$$

[2] Guido Zuccon et al. (2023). “Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.”. In: *ICTIR*, pp. 283–289.

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Time

Energy used

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Water Usage Effectiveness[2]

$$W_{off}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot PUE(t) \cdot WUE_{off}(t)$$

[2] Guido Zuccon et al. (2023). “Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.”. In: *ICTIR*, pp. 283–289.

Measuring Water

Water ➔ measures **indirect** utilisation costs

We want to measure $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Time

Energy used

$$W_{on}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot WUE_{on}(t)$$

Water Usage Effectiveness[2]

Power Usage Efficiency

$$W_{off}(\mathcal{M}) = \sum_{t=1}^{T} e(\mathcal{M}, t) \cdot PUE(t) \cdot WUE_{off}(t)$$

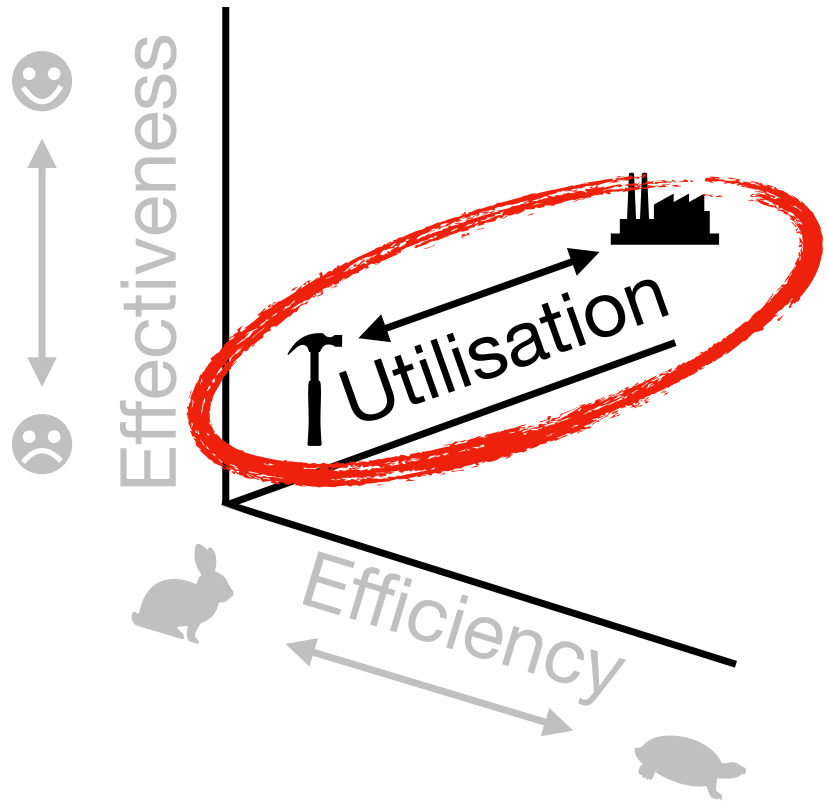
[2] Guido Zuccon et al. (2023). “Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.”. In: *ICTIR*, pp. 283–289.

Utilisation and Green IR

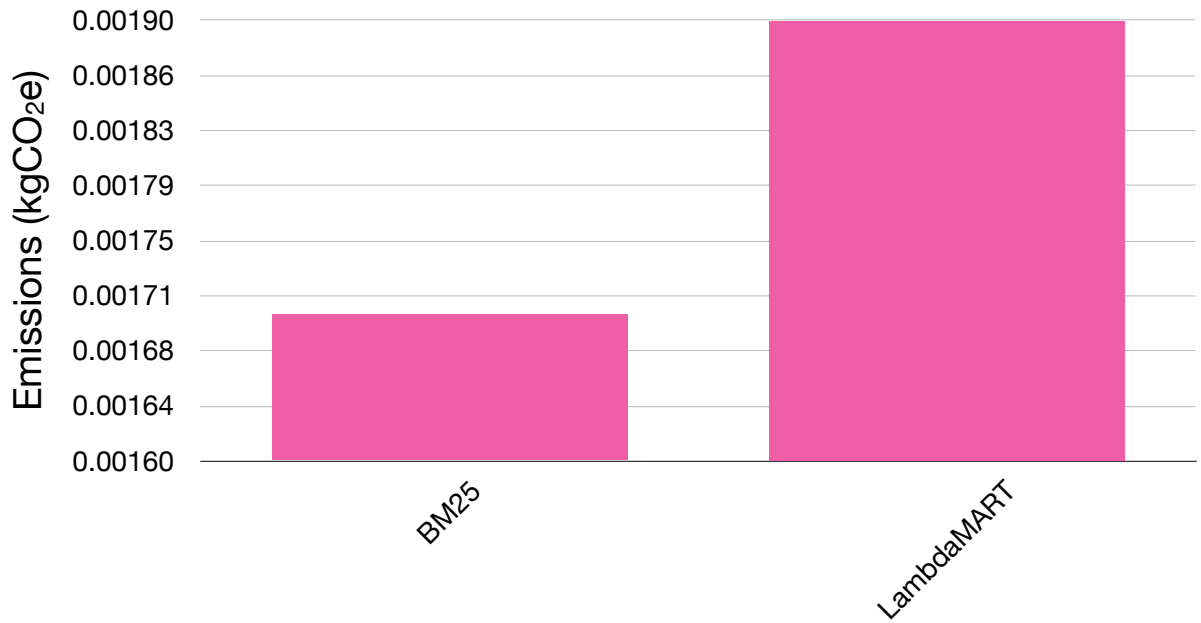
~~Okay, so what does this mean for IR?~~

~~Okay, so how can I measure this?~~

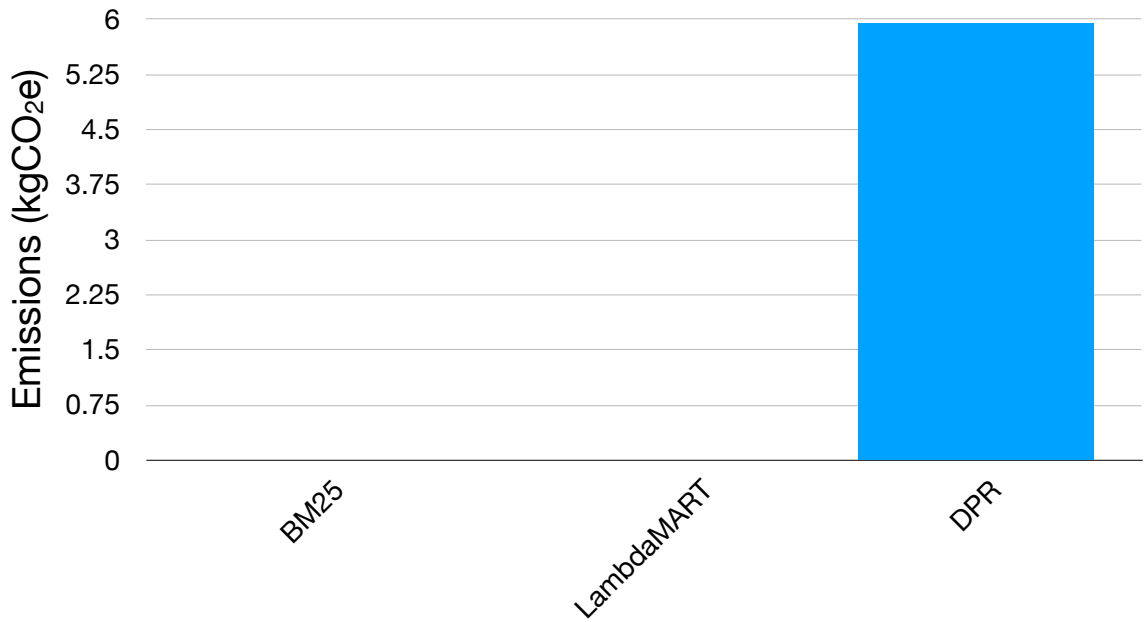
Okay, so show me what this means in IR research practice!



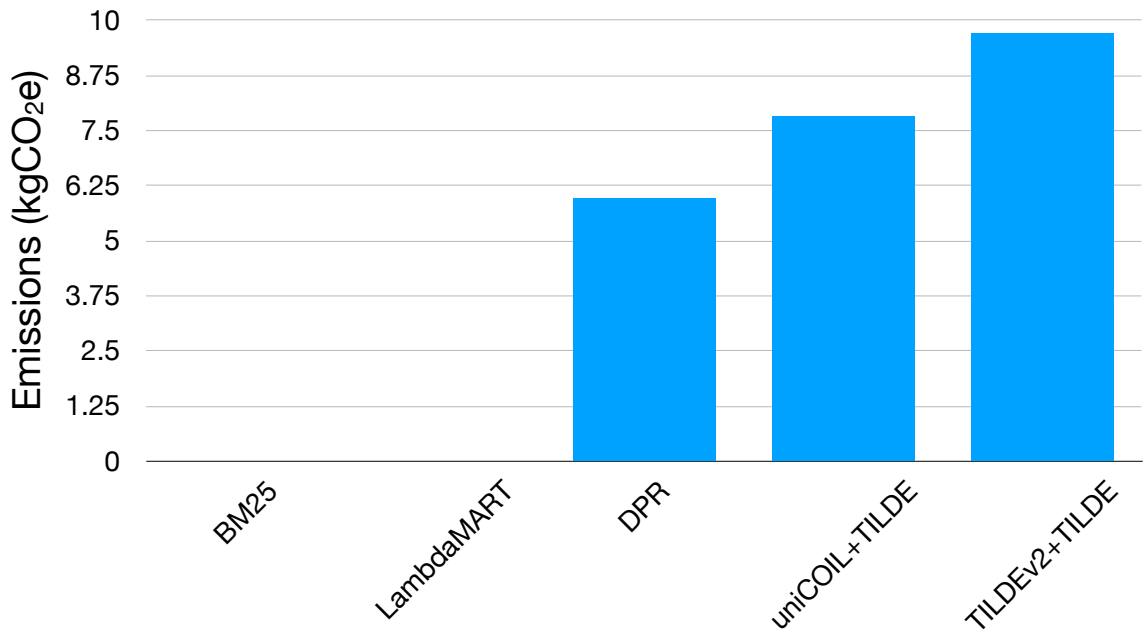
How many emissions produced to obtain a single result?



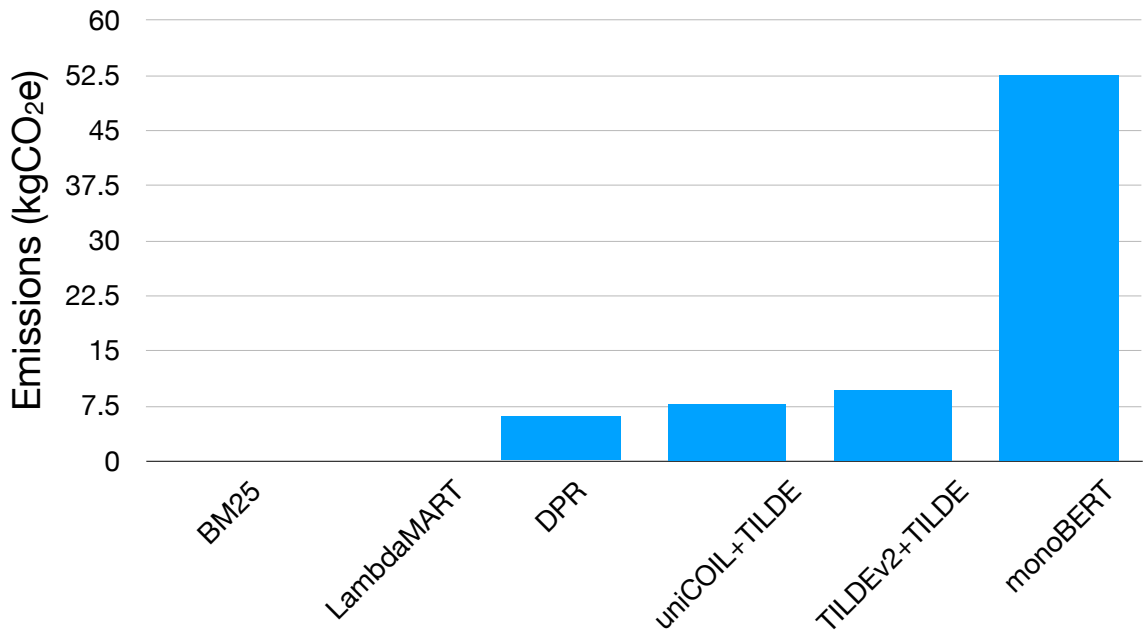
How many emissions produced to obtain a single result?



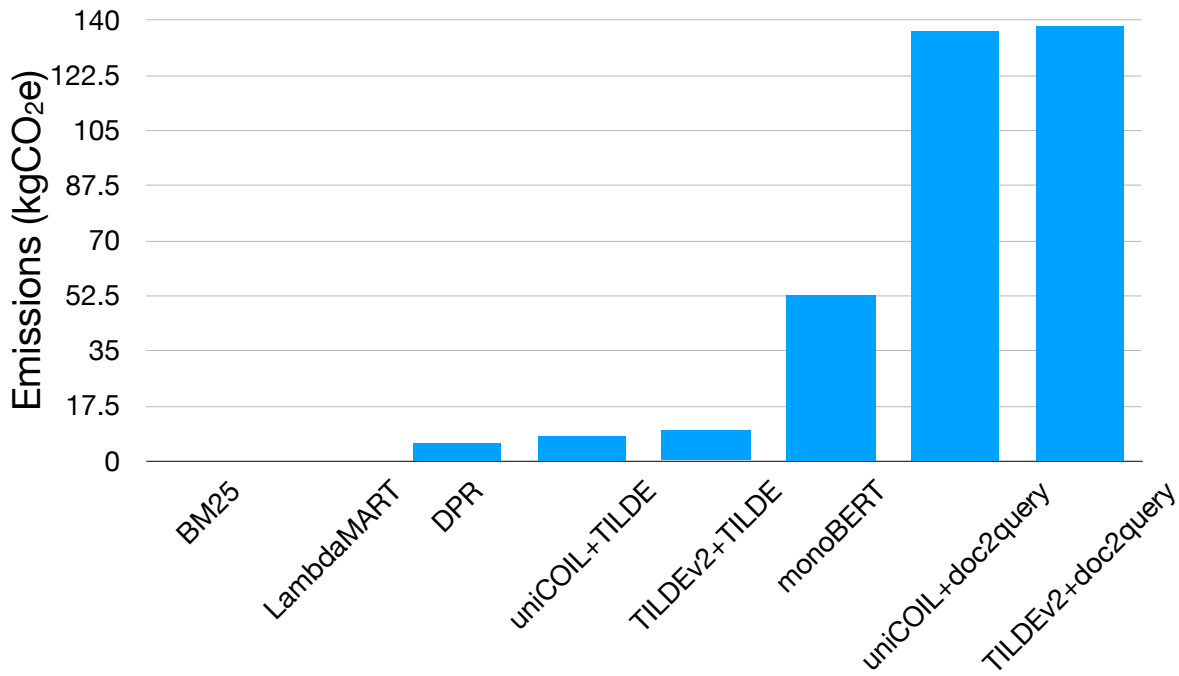
How many emissions produced to obtain a single result?



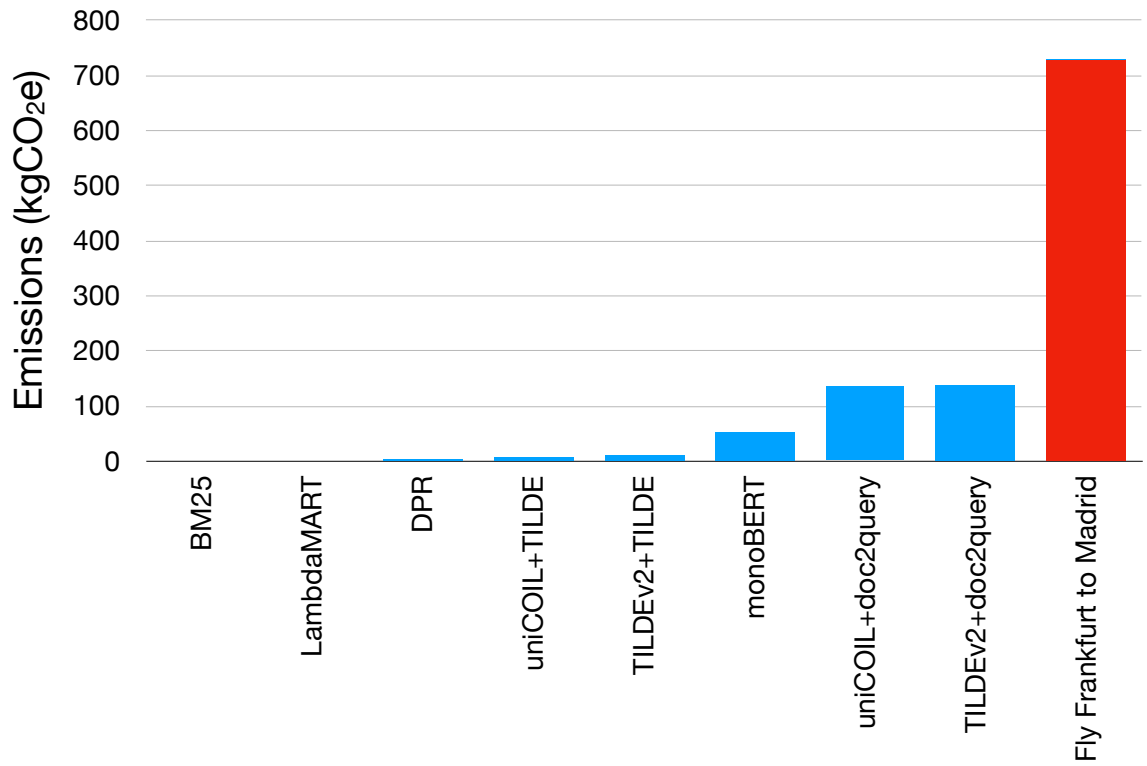
How many emissions produced to obtain a single result?



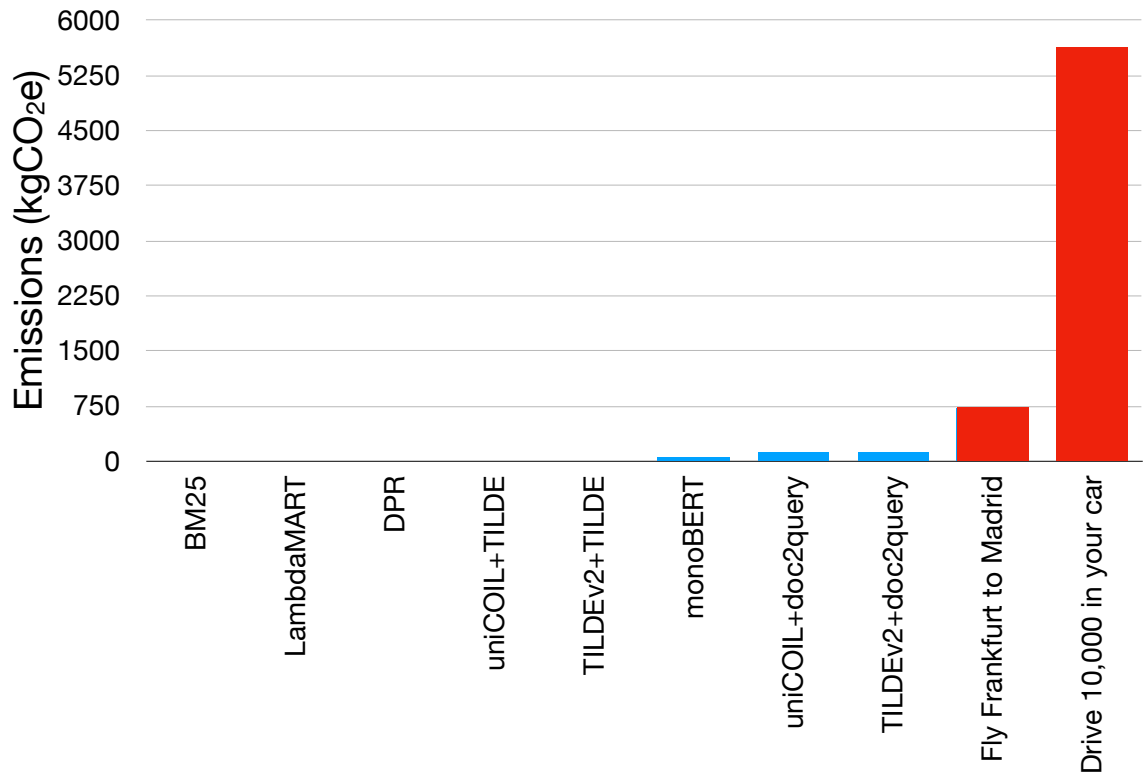
How many emissions produced to obtain a single result?



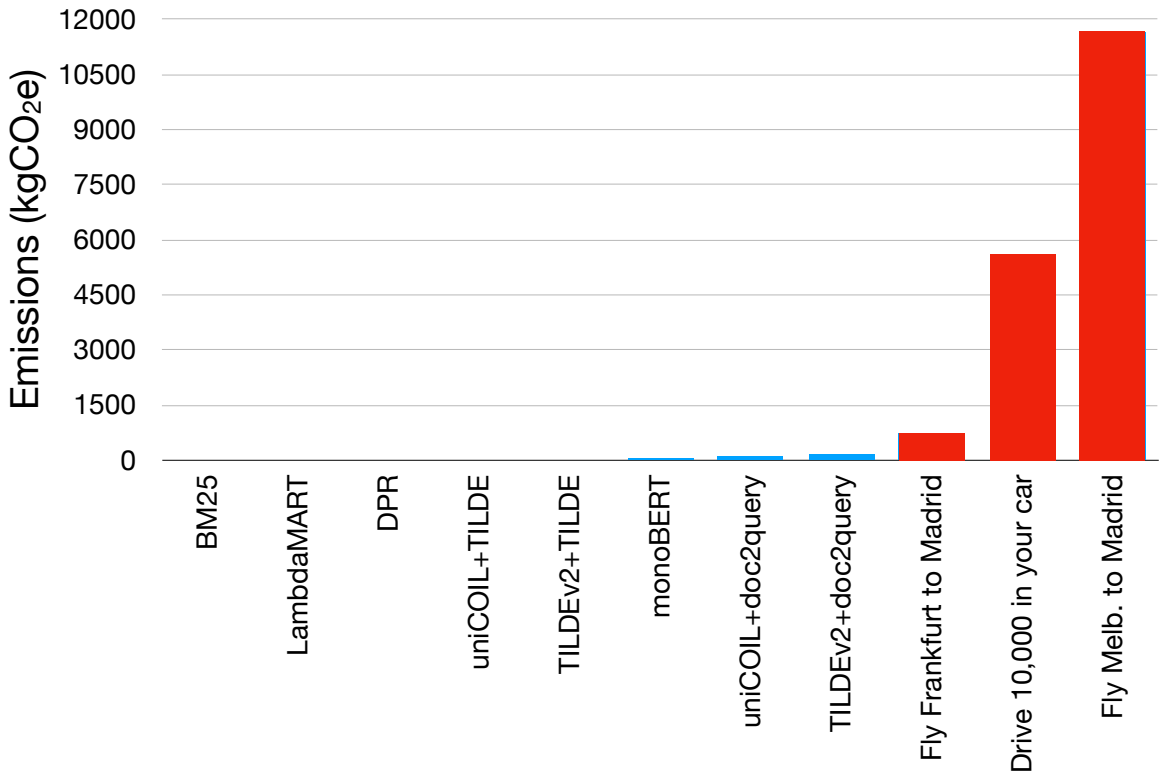
How many emissions produced to obtain a single result?



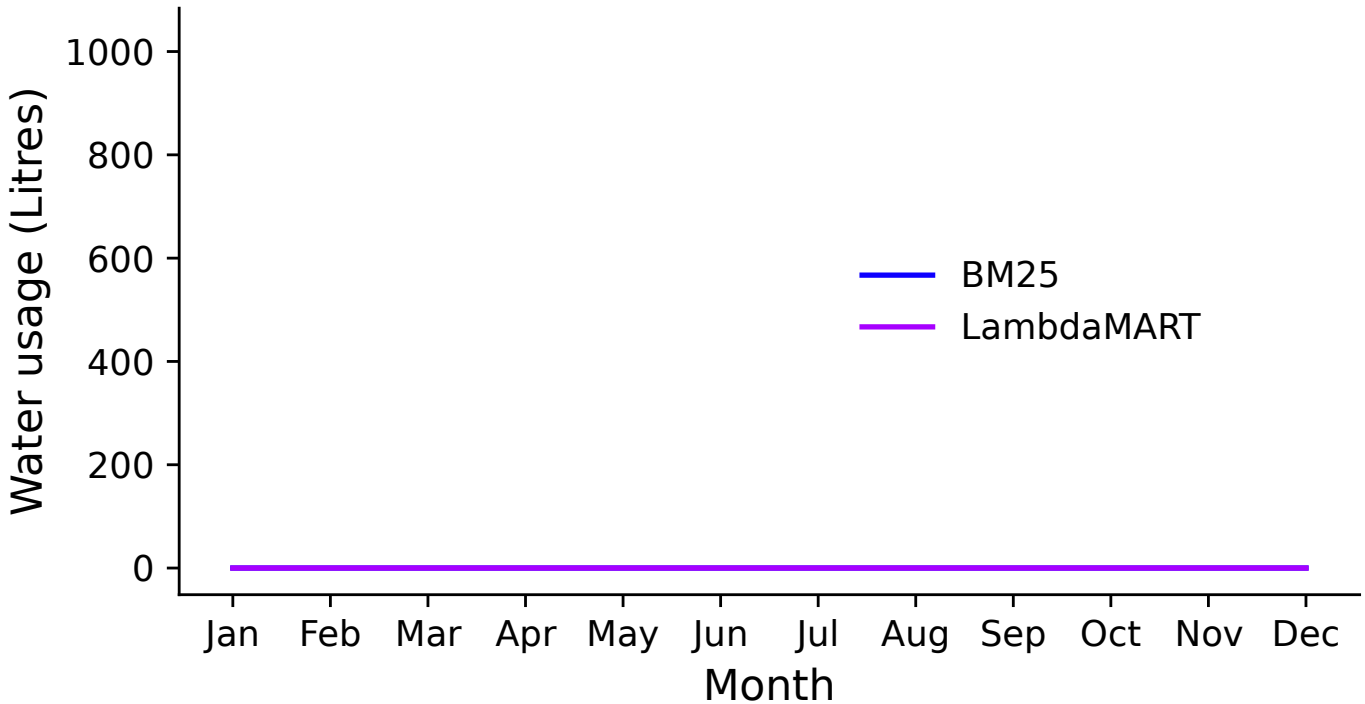
How many emissions produced to obtain a single result?



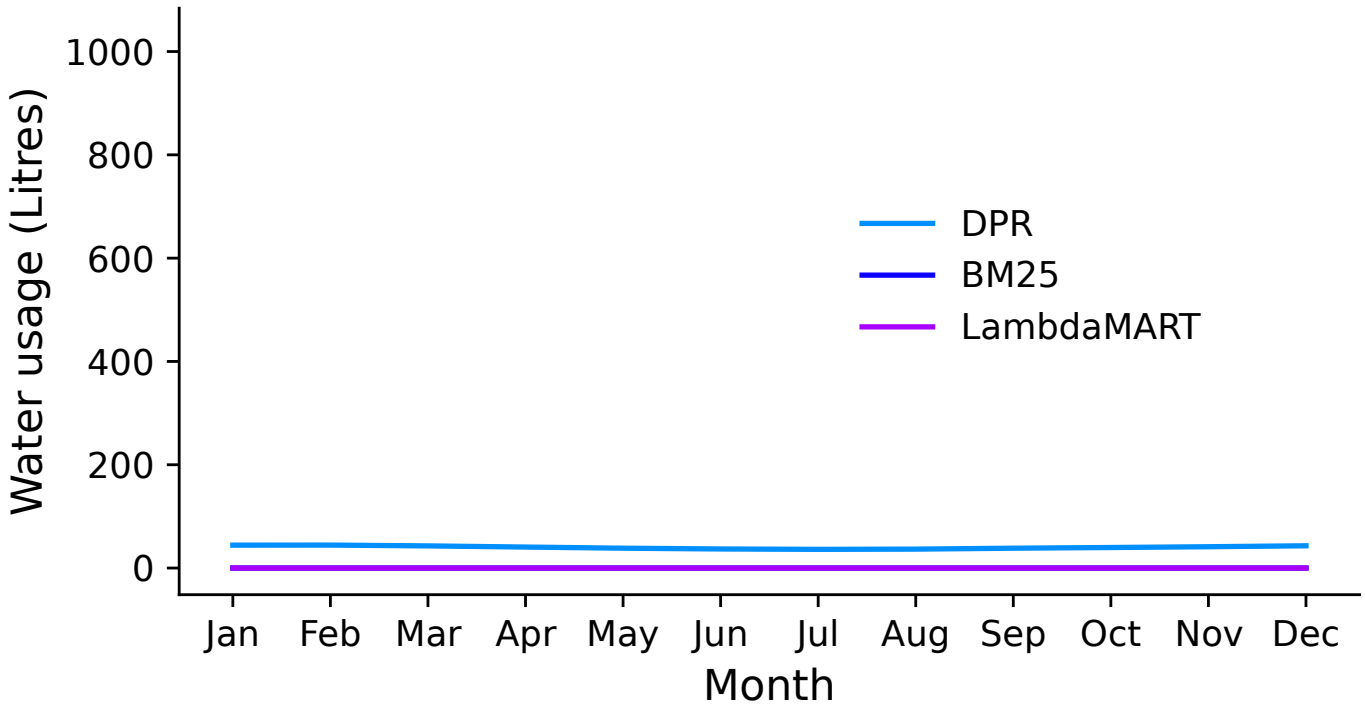
How many emissions produced to obtain a single result?



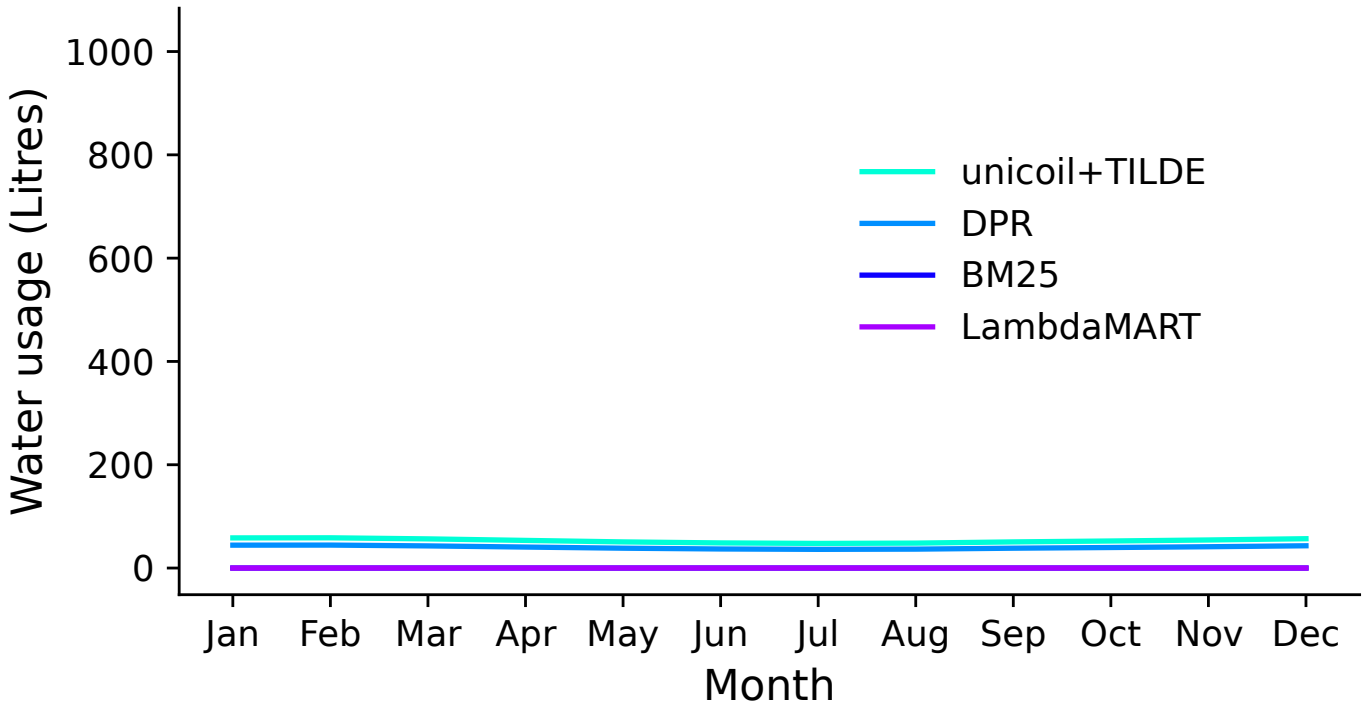
How much water used to produced to obtain a single result?



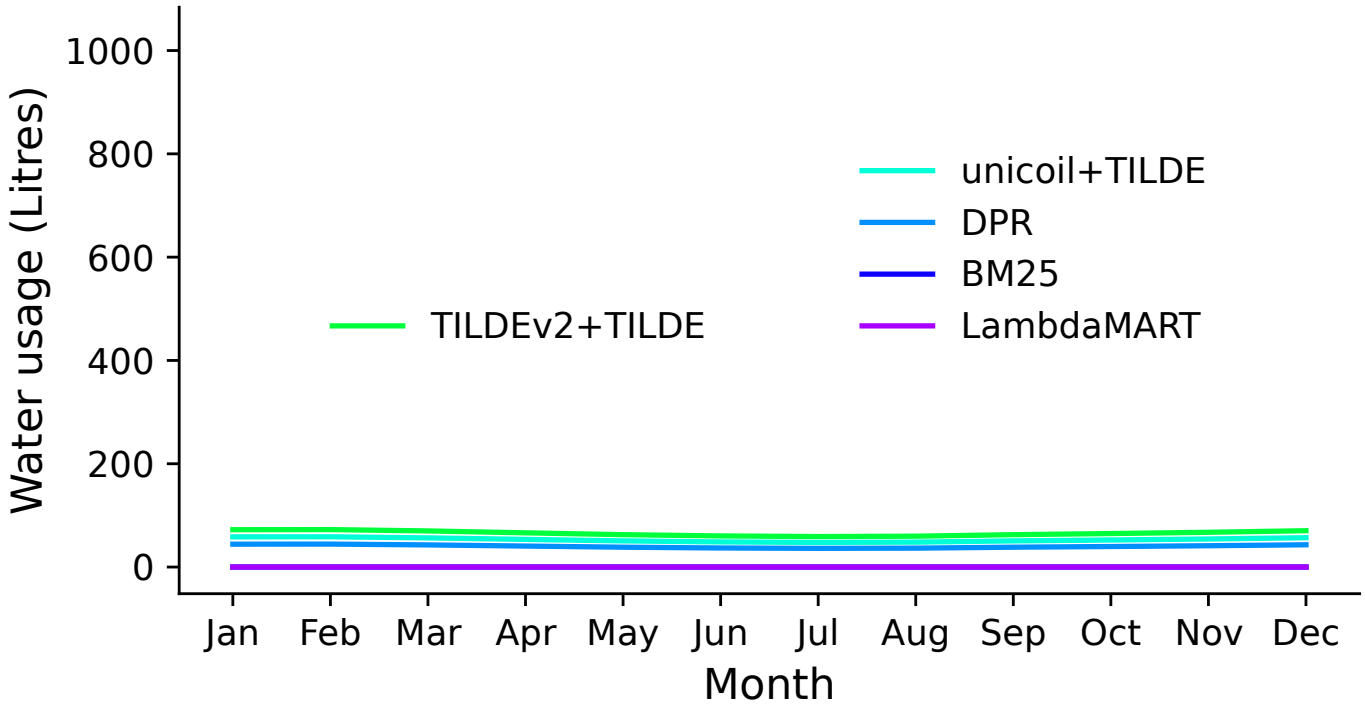
How much water used to produced to obtain a single result?



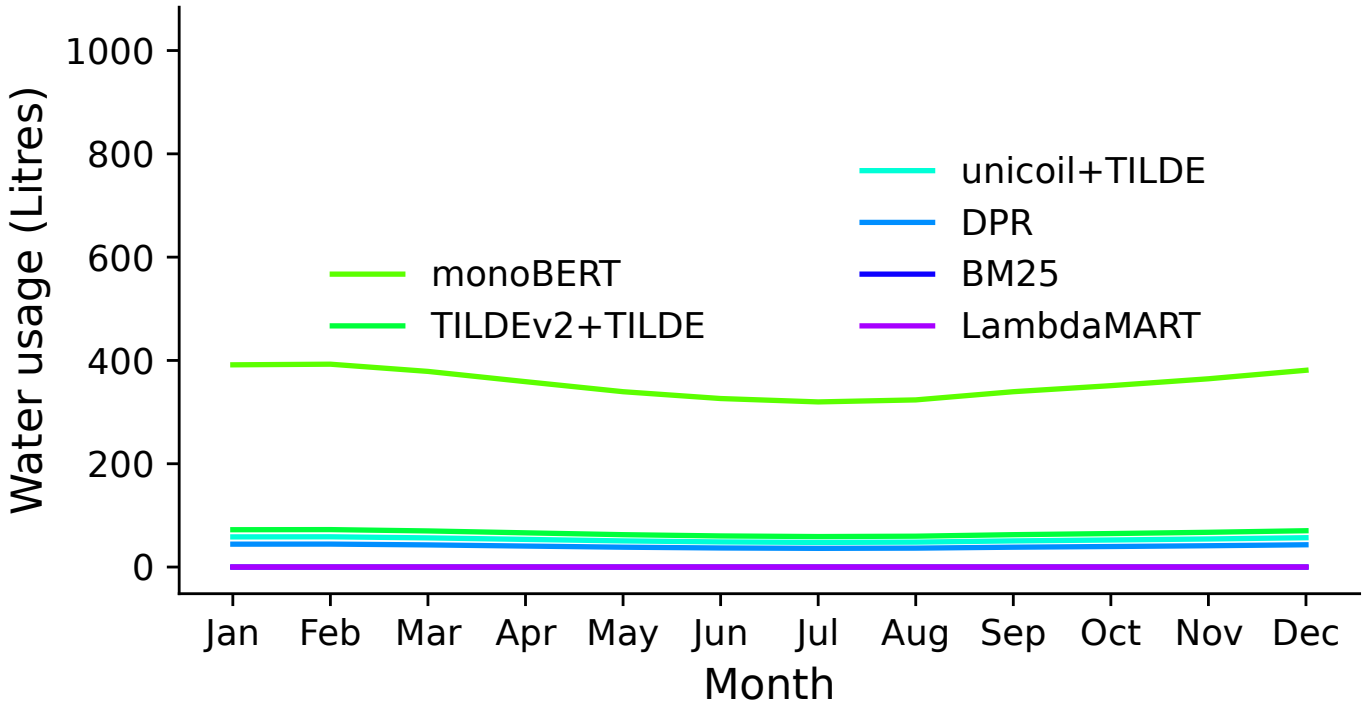
How much water used to produced to obtain a single result?



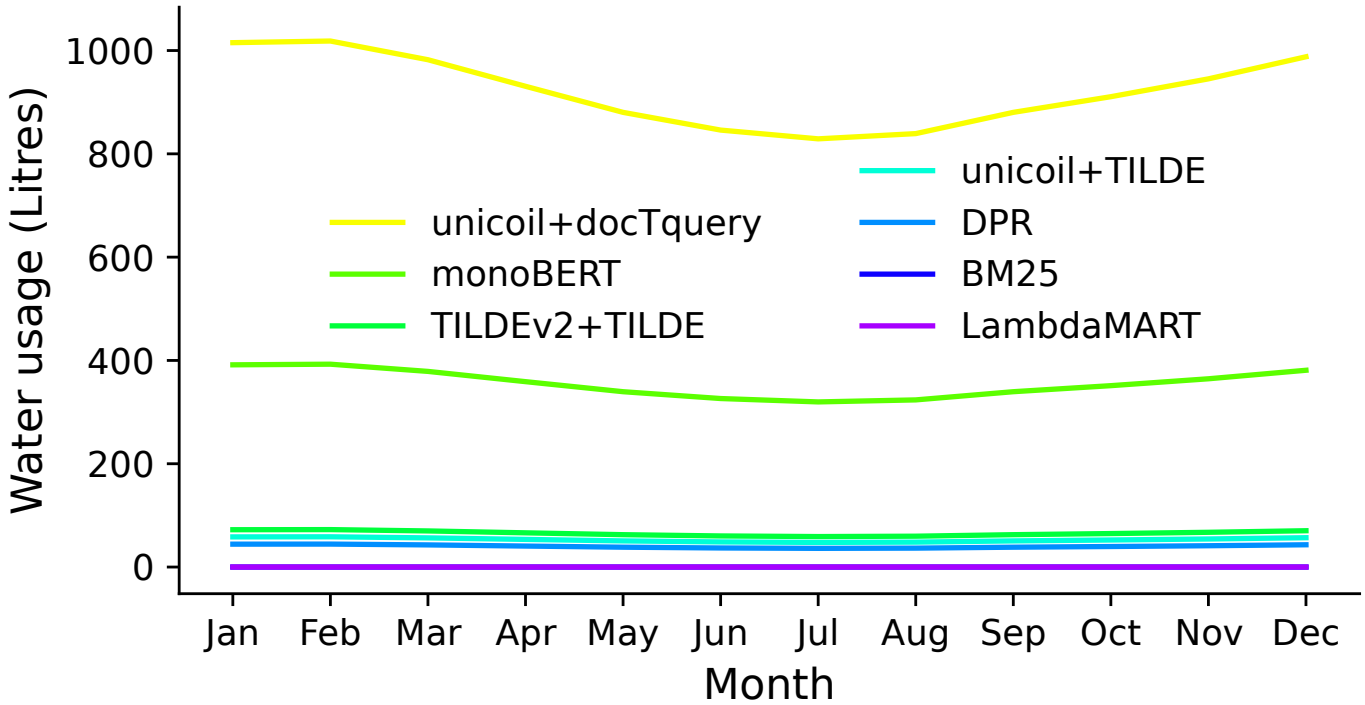
How much water used to produced to obtain a single result?



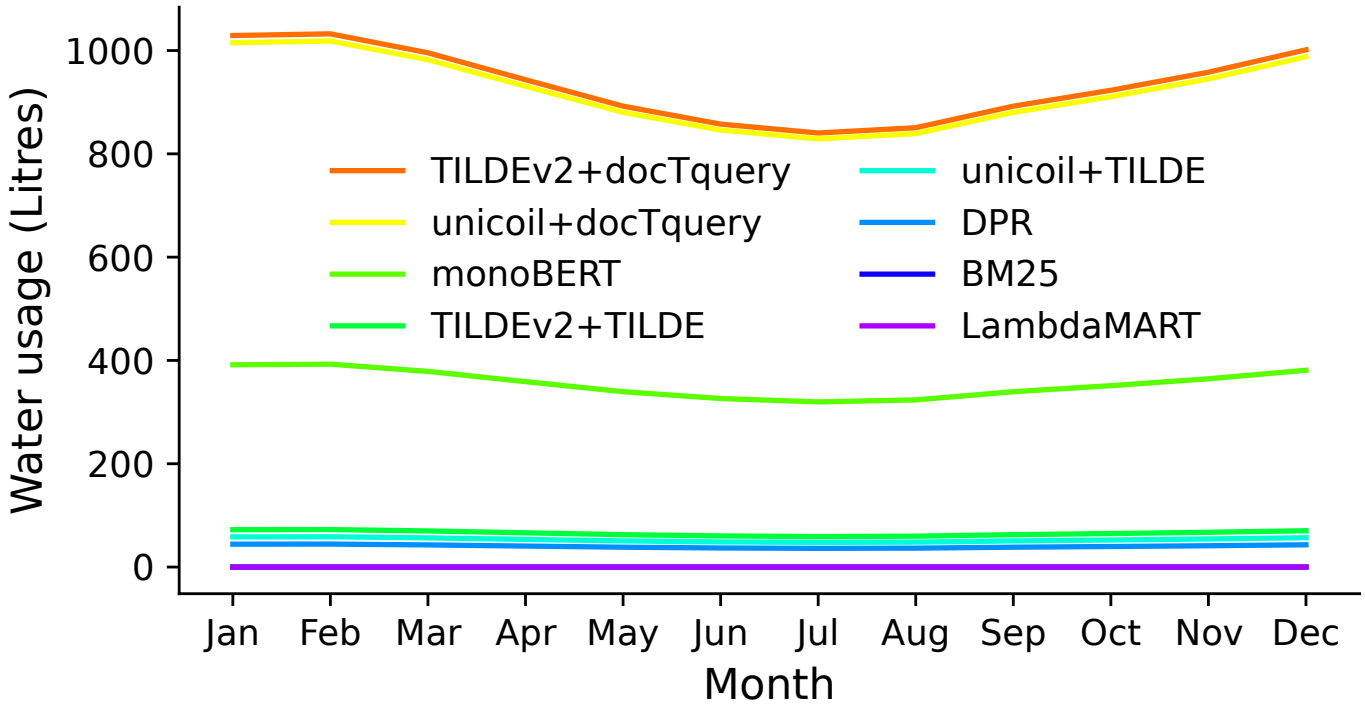
How much water used to produced to obtain a single result?



How much water used to produced to obtain a single result?



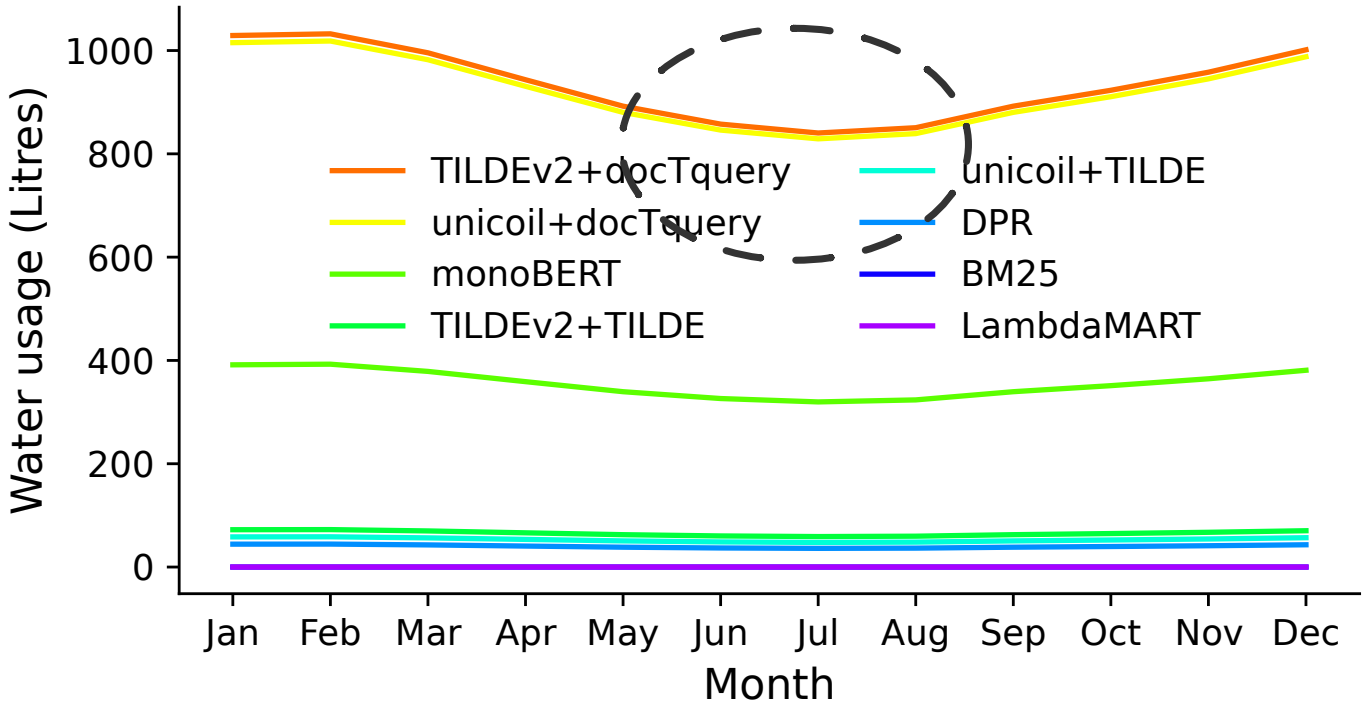
How much water used to produced to obtain a single result?



How much water used to produced to obtain a single result?

Time of year is important to how much water is used

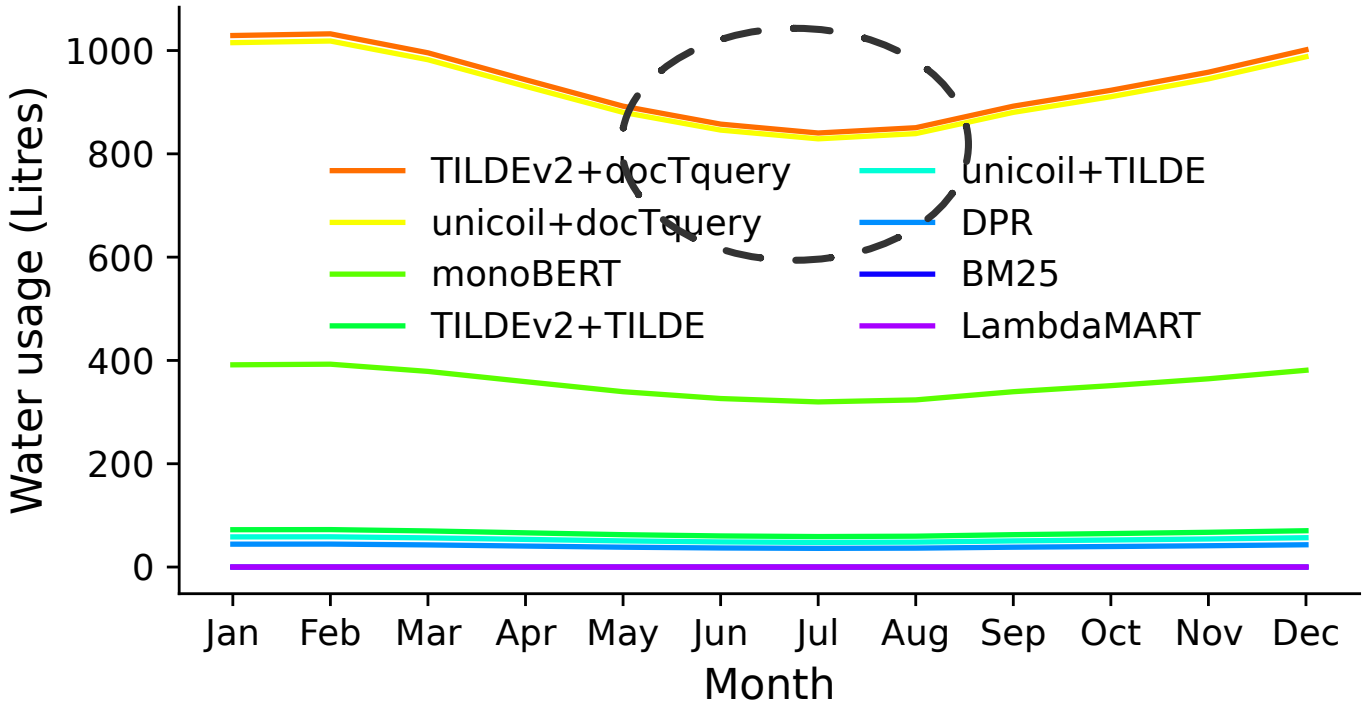
experiments performed in Australia



How much water used to produced to obtain a single result?

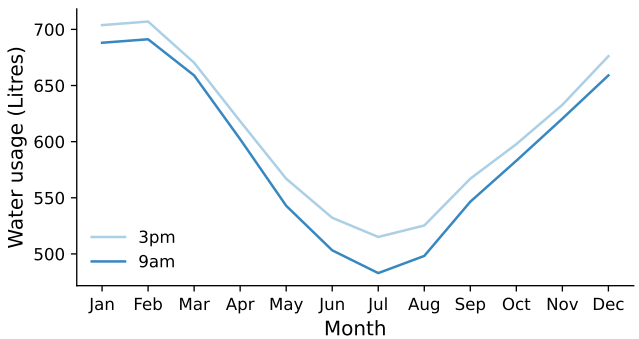
Time of year is important to how much water is used

experiments performed in Australia

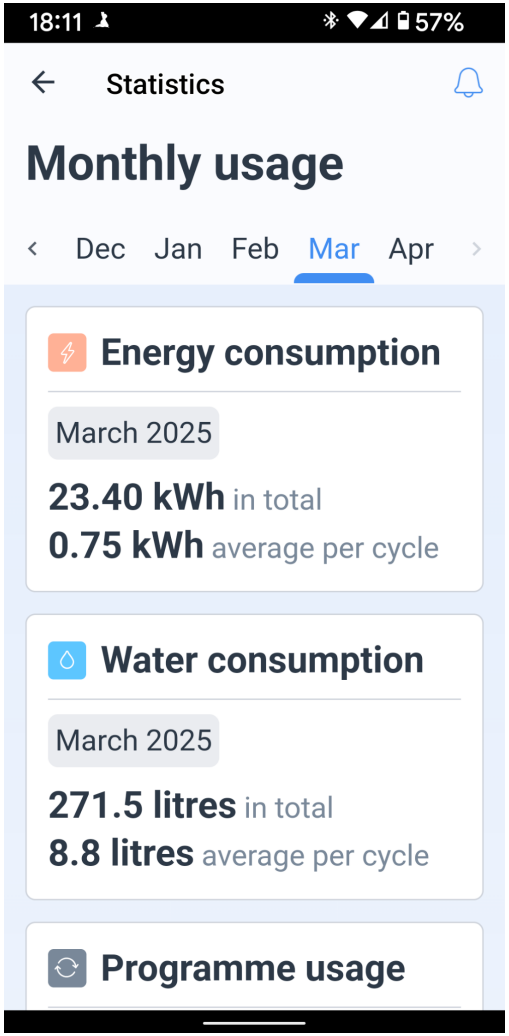


Time of day is equally important

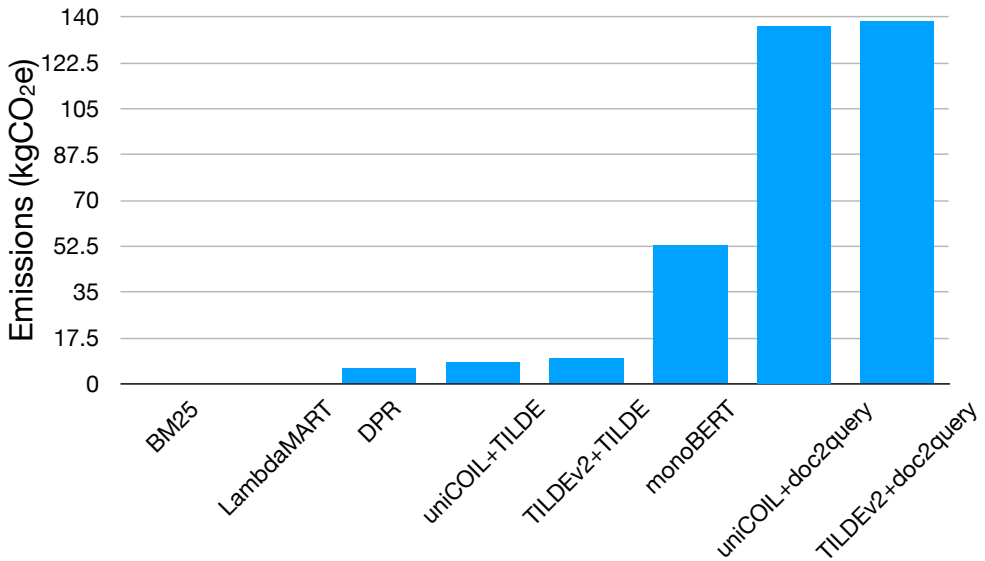
TILDEv2+docTquery

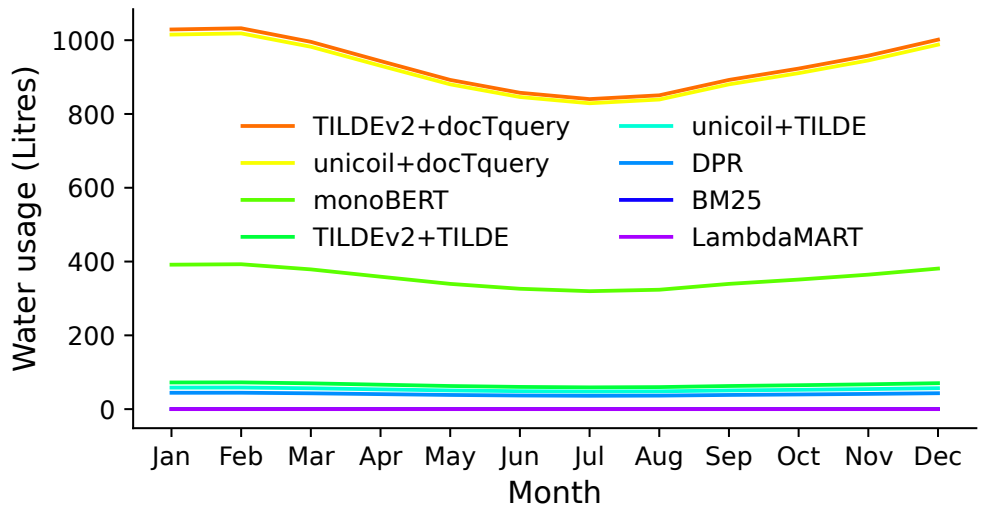


Is your model better than Harry's dishwasher?



total → 17.92kgCO_2e
avg. → 0.57kgCO_2e





Overview of Green IR

Measuring Utilisation

Corpus Subsampling

Retrieval Effectiveness

Evaluate how well our system can retrieve **relevant** documents

Retrieval Effectiveness

Evaluate how well our system can retrieve **relevant** documents

Problem: Our evaluation will always give us some number

➔ Is this number meaningful?

Retrieval Effectiveness

Evaluate how well our system can retrieve **relevant** documents

Problem: Our evaluation will always give us some number

➔ Is this number meaningful?

Solution: Ensure that our evaluation is **reliable**

➔ Observations transfer to similar scenarios with a high probability

System A > System B

Retrieval Effectiveness

Evaluate how well our system can retrieve **relevant** documents

Problem: Our evaluation will always give us some number

➔ Is this number meaningful?

Solution: Ensure that our evaluation is **reliable**

➔ Observations transfer to similar scenarios with a high probability

System A > System B

Two main aspects impact reliability[3]

- Subjectiveness of relevance judgements
- Incompleteness of relevance judgements

[3] Ellen M. Voorhees (2019). “The Evolution of Cranfield.”. In: *Information Retrieval Evaluation in a Changing World*, pp. 45–69.

Relevance judgements are highly subjective

Relevance judgements are highly subjective

hydrogen liquid at what temperature?

Relevance judgements are highly subjective

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Relevance judgements are highly subjective

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Relevance judgements are highly subjective

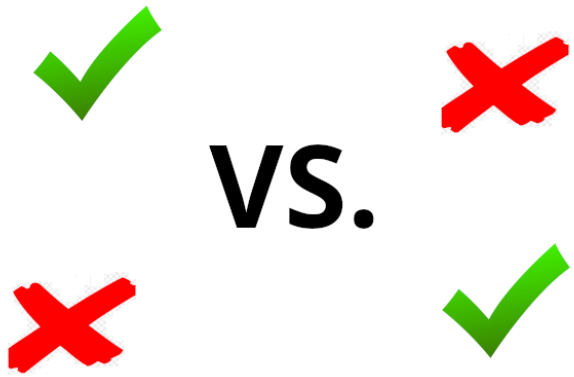
hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K



Relevance judgements are highly subjective

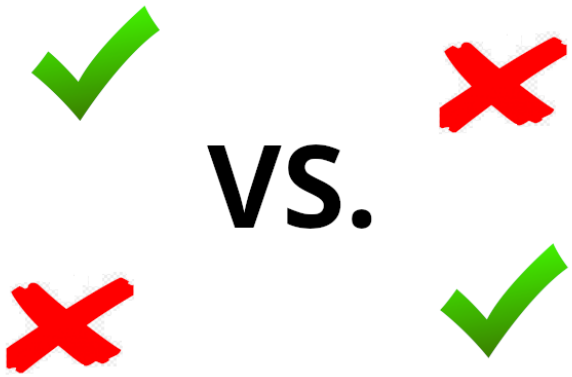
hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K



➜ Humans disagree substantially but rarely impacts system rankings

Relevance judgements are often incomplete

Relevance judgements are often incomplete

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Relevance judgements are often incomplete

hydrogen liquid at what temperature?

What is the **temperature** of **liquid hydrogen**?

Hydrogen

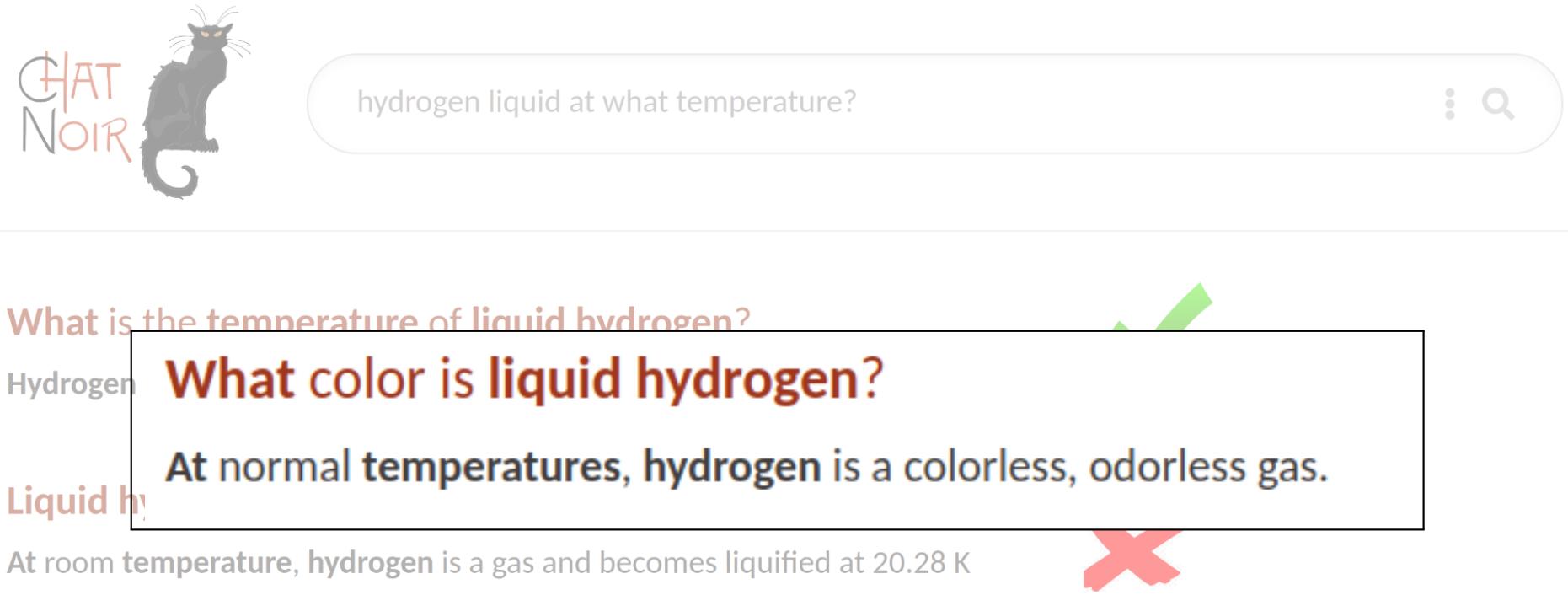
What color is **liquid hydrogen**?

At normal **temperatures**, **hydrogen** is a colorless, odorless gas.

Liquid h

At room **temperature**, **hydrogen** is a gas and becomes liquified at 20.28 K

Relevance judgements are often incomplete



Default assumption: Relevance judgements are **essentially complete**

- An unjudged document is assumed to be non-relevant
- New systems that retrieve new documents might be underestimated

Measuring the reliability of experiments [Breuer'20]

Ranking correlations can confirm the reliability of evaluations

Measuring the reliability of experiments [Breuer'20]

Ranking correlations can confirm the reliability of evaluations

Step 1: Create a system ranking

- Input: A set of retrieval systems and an evaluation measure
- Rank all systems by their effectiveness

System A > System B > System C > System D

Measuring the reliability of experiments [Breuer'20]

Ranking correlations can confirm the reliability of evaluations

Step 1: Create a system ranking

- Input: A set of retrieval systems and an evaluation measure
- Rank all systems by their effectiveness

System A > System B > System C > System D

Step 2: Repeat the experiment

- Observe new system rankings
- Calculate ranking correlation between old and new system ranking

Measuring the reliability of experiments [Breuer'20]

Ranking correlations can confirm the reliability of evaluations

Step 1: Create a system ranking

- Input: A set of retrieval systems and an evaluation measure
- Rank all systems by their effectiveness

System A > System B > System C > System D

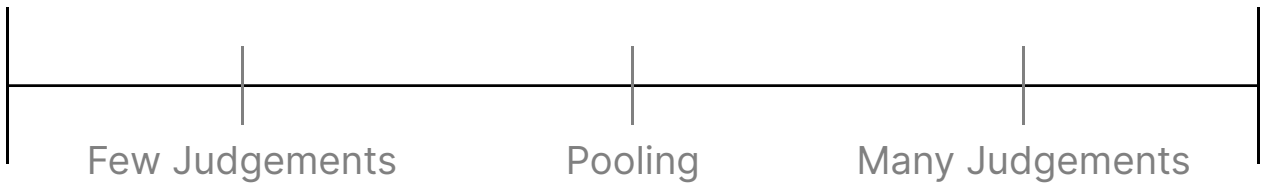
Step 2: Repeat the experiment

- Observe new system rankings
- Calculate ranking correlation between old and new system ranking

➔ Goal: **Green** and **Reliable** IR experiments

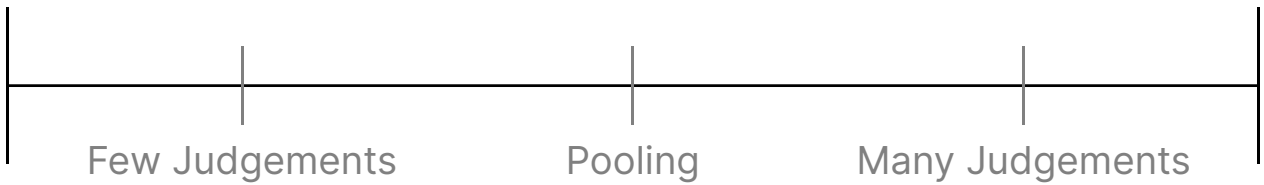
How build our test collection? Step 1: Queries

Many queries, few judgements or few queries, many judgements?



How build our test collection? Step 1: Queries

Many queries, few judgements or few queries, many judgements?

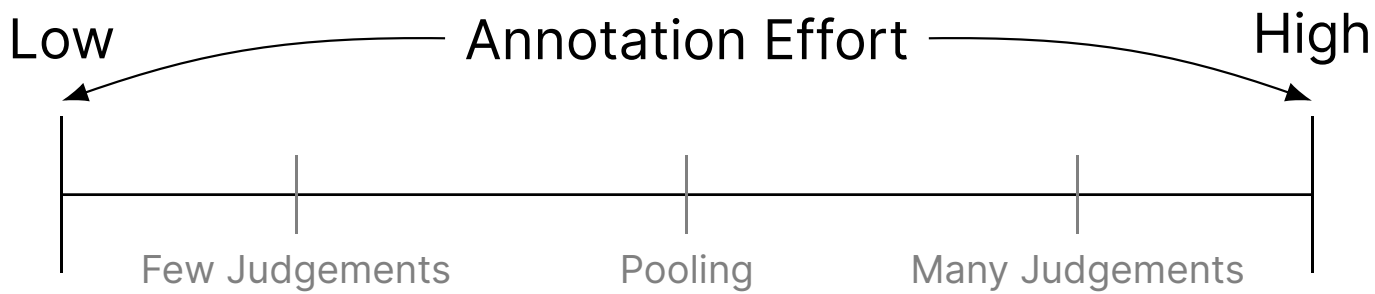


Few: E.g., one relevant document derived via click logs

Pooling: E.g., Judge the top-*k* results of each system (usually graded)

How build our test collection? Step 1: Queries

Many queries, few judgements or few queries, many judgements?

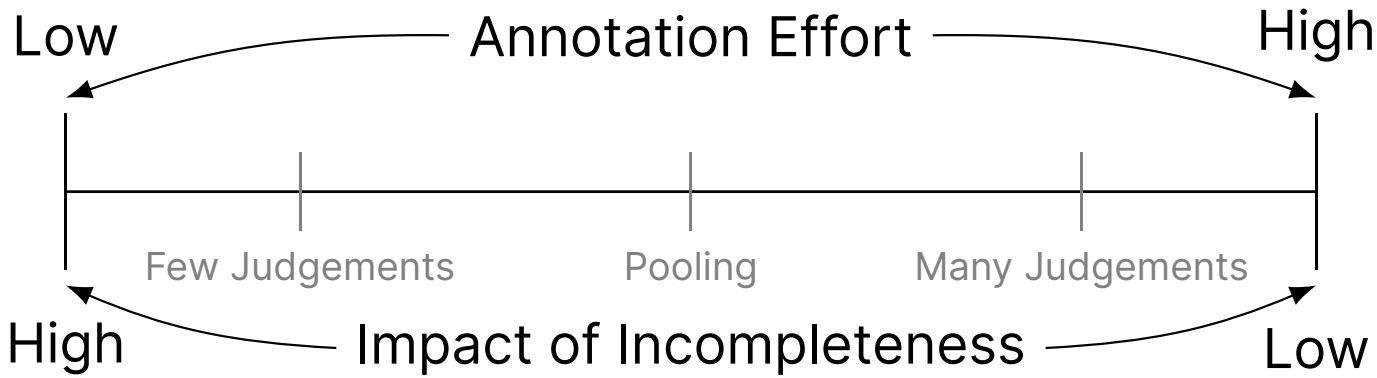


Few: E.g., one relevant document derived via click logs

Pooling: E.g., Judge the top-*k* results of each system (usually graded)

How build our test collection? Step 1: Queries

Many queries, few judgements or few queries, many judgements?

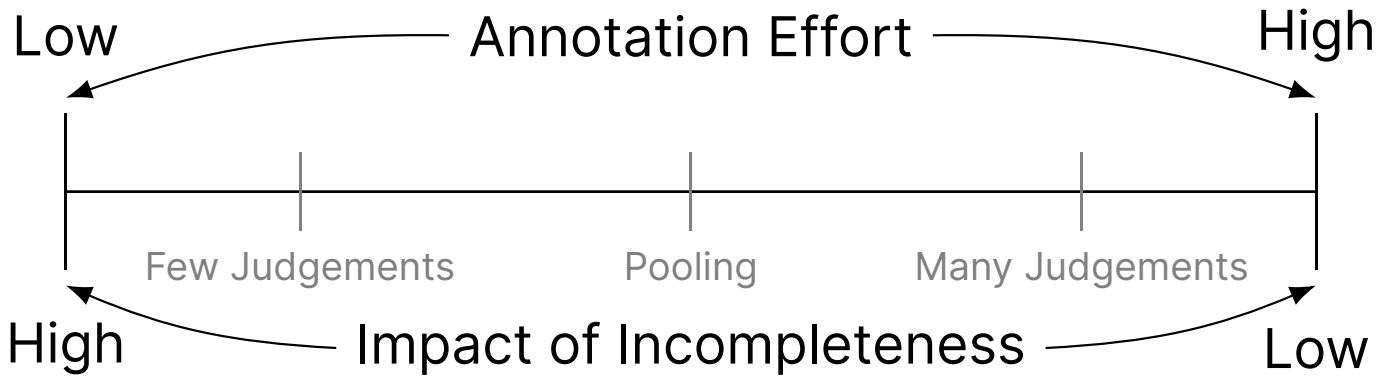


Few: E.g., one relevant document derived via click logs

Pooling: E.g., Judge the top-*k* results of each system (usually graded)

How build our test collection? Step 1: Queries

Many queries, few judgements or few queries, many judgements?



Few: E.g., one relevant document derived via click logs

Pooling: E.g., Judge the top-*k* results of each system (usually graded)

How many different rankings?

Many judgements advantageous from Green IR perspective

Labels				Top-10 Rankings
0	1	2	3	
∞	1	—	—	11
∞	10	10	10	$>$ 1 million

How build our test collection? Step 2: Documents

What documents should we include?

Evaluation Corpora with top-k pooling typically:

- Have **50 queries**
- Pool **30 to 100 systems**
- Between **10 million and 1 billion documents**

How build our test collection? Step 2: Documents

What documents should we include?

Evaluation Corpora with top-k pooling typically:

- Have **50 queries**
- Pool **30 to 100 systems**
- Between **10 million and 1 billion documents**

Considerations:

- A few million document suffice to satisfy most information needs
- We do not need to include all relevant documents
- We only need a subset that allows reliable evaluations

How build our test collection? Step 2: Documents

What documents should we include?

Evaluation Corpora with top-k pooling typically:

- Have **50 queries**
- Pool **30 to 100 systems**
- Between **10 million and 1 billion documents**

Considerations:

- A few million document suffice to satisfy most information needs
- We do not need to include all relevant documents
- We only need a subset that allows reliable evaluations

What documents to include to evaluate on ca. 50 pooled queries?

How build our Evaluation Dataset? Step 2: Documents

Judegment Pool:

- Select all documents with a judegment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

How build our Evaluation Dataset? Step 2: Documents

Judegment Pool:

- Select all documents with a judegment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

How build our Evaluation Dataset? Step 2: Documents

Judegment Pool:

- Select all documents with a judegment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

Judegment Pool + Random

- All documents with a judgement plus random documents
- Disadvantage: Random documents are too easy negatives

How build our Evaluation Dataset? Step 2: Documents

Judegment Pool:

- Select all documents with a judegment. E.g., the top-10 pool
- Disadvantage: Effectiveness overestimated in post-hoc experiments [Sakai'08,Fröbe'23]

Re-Ranking:

- Select all documents retrieved by a model. E.g., the top-1k of BM25
- Disadvantage: Bias towards the first stage model

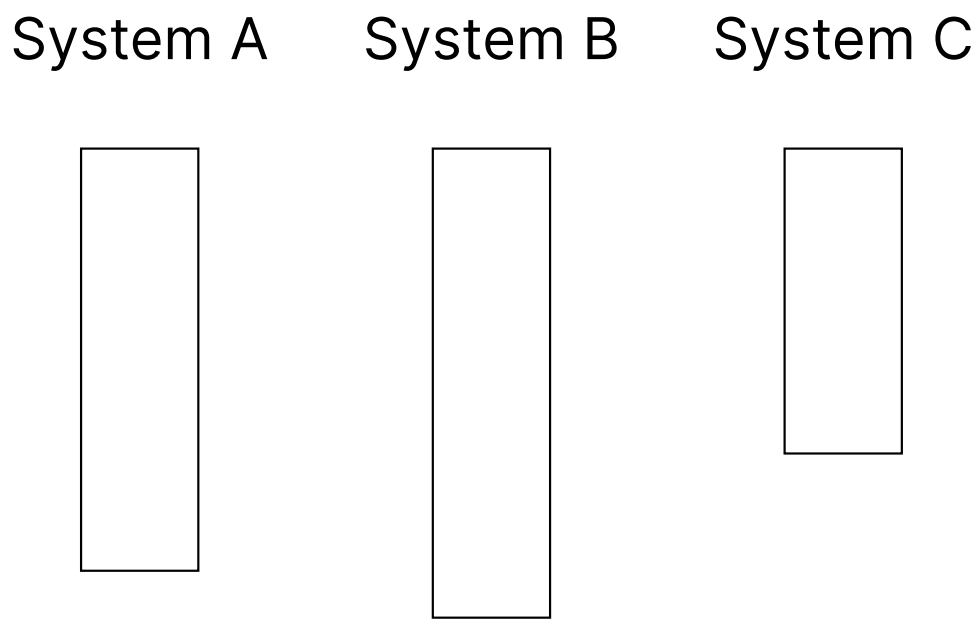
Judegment Pool + Random

- All documents with a judgement plus random documents
- Disadvantage: Random documents are too easy negatives

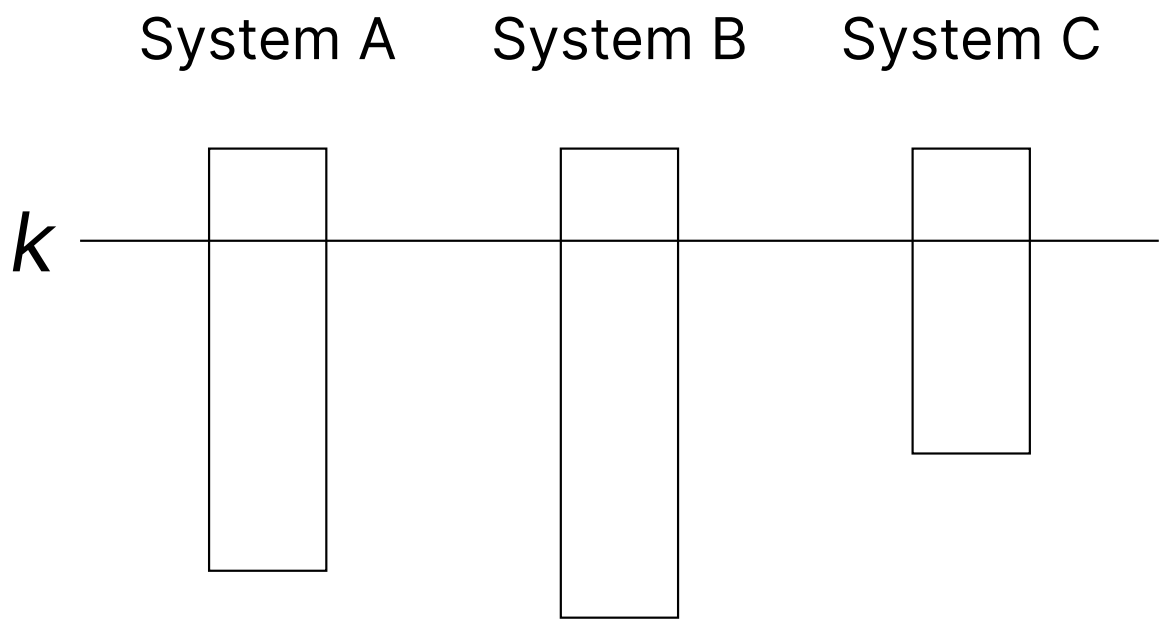
Re-Pooling

- Re-Pool to $k' >> k$. E.g., top-100 or 1k for a top-10 judgement pool
- Addresses all disadvantages of the three above approaches

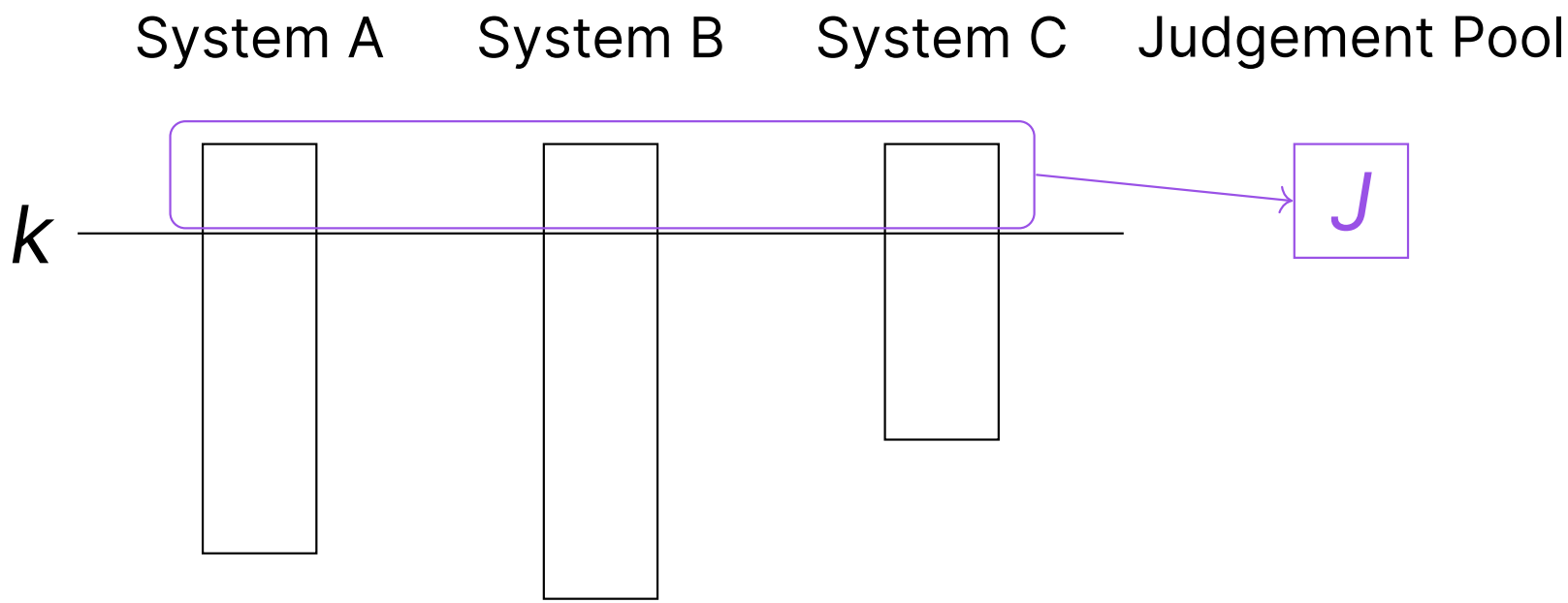
How build our Evaluation Dataset? Step 2: Documents



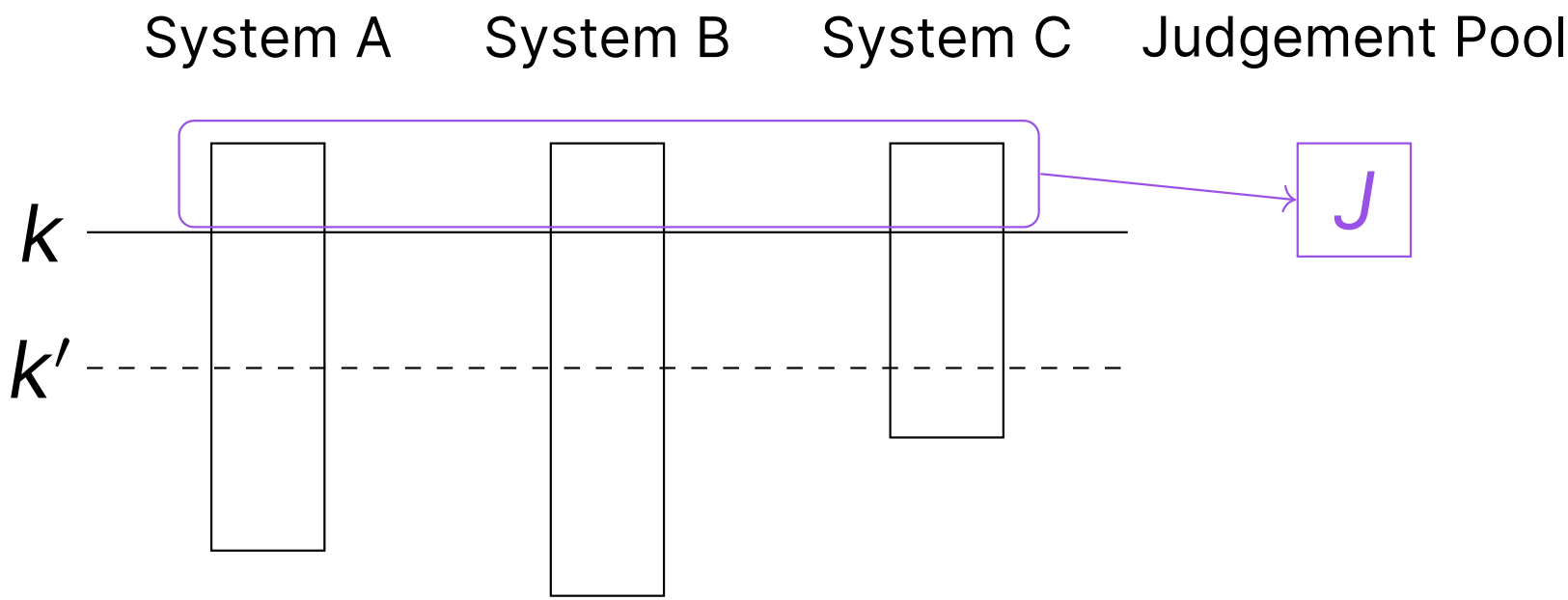
How build our Evaluation Dataset? Step 2: Documents



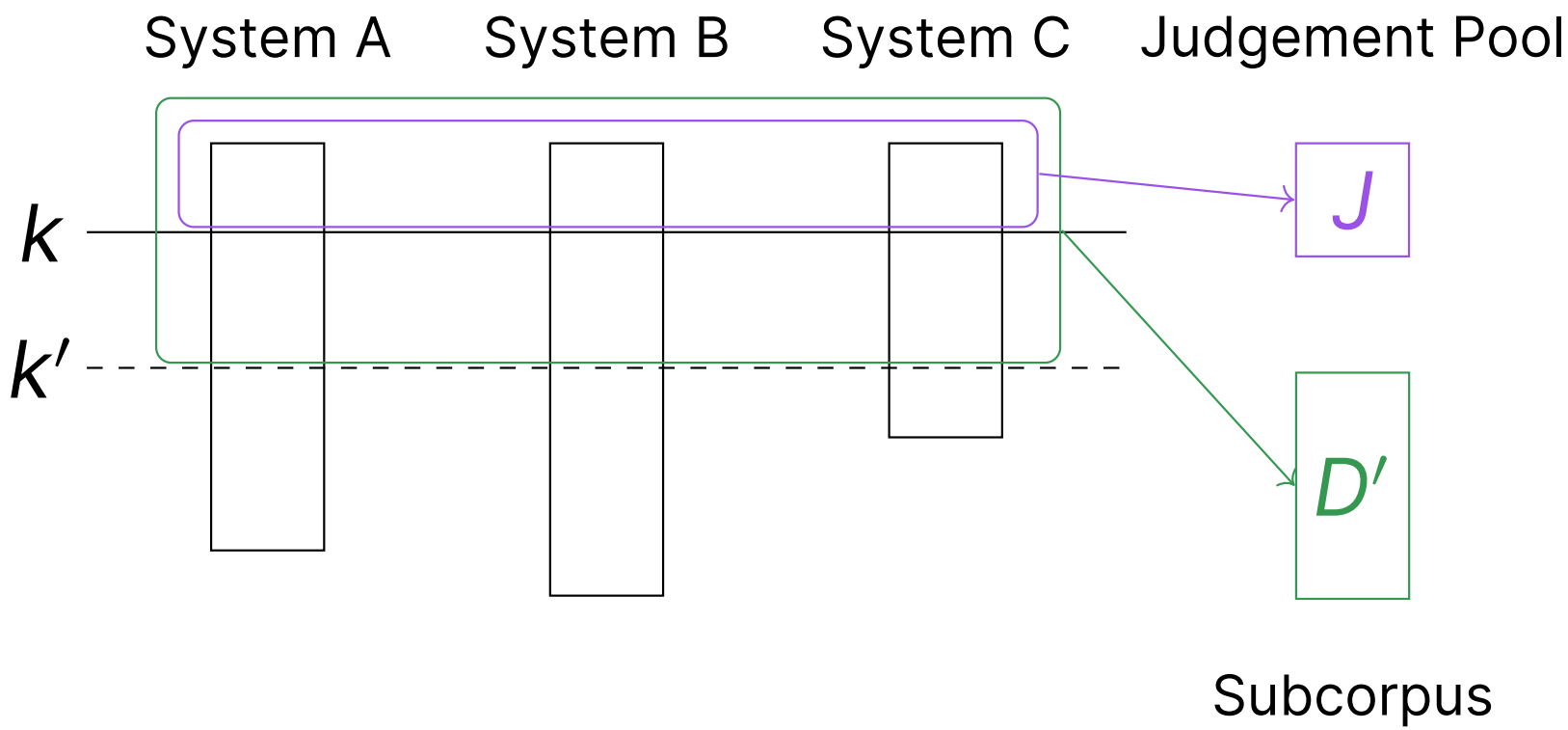
How build our Evaluation Dataset? Step 2: Documents



How build our Evaluation Dataset? Step 2: Documents



How build our Evaluation Dataset? Step 2: Documents



Results of Corpus Subsampling

How big are the resulting subcorpora?

Corpus	Complete			Subsampled		
	Docs.	$\notin_J$	Size	Docs.	$\notin_J$	Size
ClueWeb09	1.0 b	99 %	4.0 TB	0.3 m	73 %	0.9 GB
ClueWeb12	0.7 b	99 %	4.5 TB	0.1 m	72 %	0.5 GB
Disks 4/5	0.5 m	41 %	0.6 GB	0.4 m	31 %	0.5 GB
MS MARCO	8.8 m	99 %	2.9 GB	0.3 m	97 %	42.1 MB

$\notin_J$ ➔ amount of unjudged documents

Results of Corpus Subsampling

Similarity to ground truth:

Subsampling	τ			
	ClueWeb09	**ClueWeb12**	**Robust04**	**MS MARCO**
Judgement Pool	0.944	0.941	0.983	0.978
Re-Ranking BM25	0.936	0.938	0.836	0.994
Judgement Pool + Random	0.799	0.765	0.789	0.794
Re-Pooling $k^{'} = 100$	**0.980**	**0.987**	**0.995**	**0.999**

Results of Corpus Subsampling

Similarity to ground truth:

Subsampling	τ			
	ClueWeb09	**ClueWeb12**	**Robust04**	**MS MARCO**
Judgement Pool	0.944	0.941	0.983	0.978
Re-Ranking BM25	0.936	0.938	0.836	0.994
Judgement Pool + Random	0.799	0.765	0.789	0.794
Re-Pooling $k^{'} = 100$	**0.980**	**0.987**	**0.995**	**0.999**

Re-Pooling does not overestimate effectiveness:

Subsampling	$\Delta_{\text{nDCG@10}}$			
	ClueWeb09	**ClueWeb12**	**Robust04**	**MS MARCO**
Judgement Pool	0.030	0.031	0.005	0.011
Re-Ranking BM25	-0.013	-0.053	0.049	-0.005
Judgement Pool + Random	0.375	0.325	0.062	0.259
Re-Pooling $k^{'} = 100$	-0.030	-0.060	-0.004	-0.007

Conclusion and Future Work

Summary

- Utilisation of our experiments is not negligible
- Corpus subsampling ➔ reliable evaluation, small fraction of utilisation

Future Work

- Can corpus subsampling be incorporated into evaluation campaigns?
- How to holistically evaluate efficiency and effectiveness?
- Upcoming workshop on that: ReNeuIR 2025 at SIGIR

Maik Fröbe et al. (2025). “Corpus Subsampling: Estimating the Effectiveness of Neural Retrieval Models on Large Corpora.”. In: *ECIR*, pp. 453–471