Green IR: Measuring and Applications

Maik Fröbe, Harry Scells

31.03.2025

Information Access Systems impact our Environment

Information Access Systems impact our Environment

Poll: What causes more emissions?

A Google search vs. a ChatGPT response

Information Access Systems impact our Environment

Poll: What causes more emissions?

A Google search vs. a ChatGPT response

The Guardian

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.

What about IR Research?

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

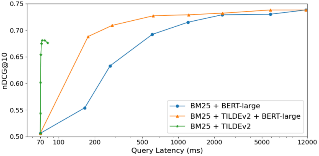
Retrieval Efficiency

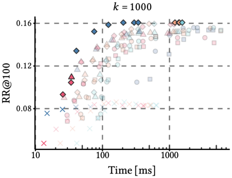
Speed a system can retrieve relevant information in response to a query

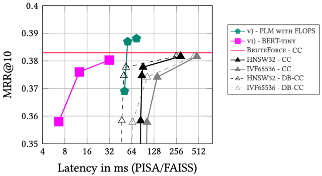
Factors that impact retrieval efficiency:

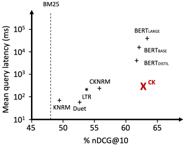
- **Size and complexity** of the search corpus
- Effectiveness of the **retrieval models** or techniques used
- Efficiency of the **hardware and infrastructure** used

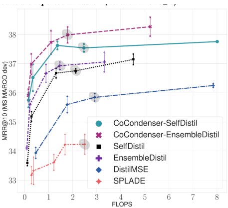
Retrieval Efficiency

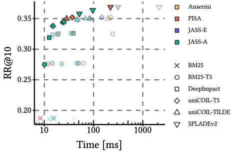


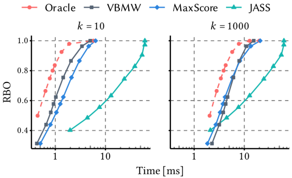




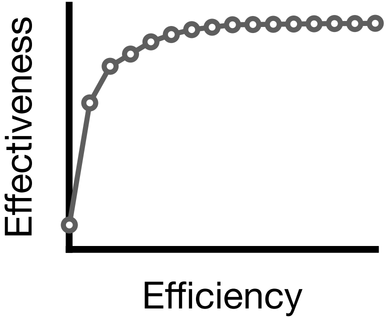




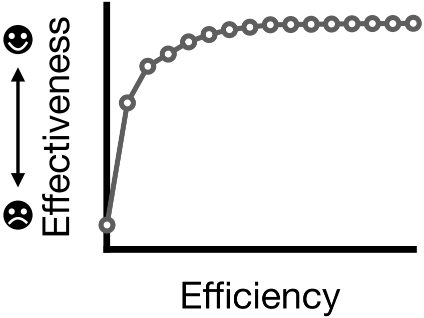




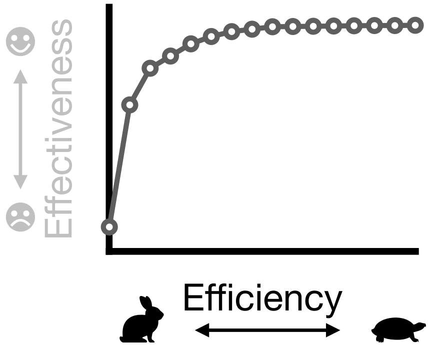
Retrieval Efficiency



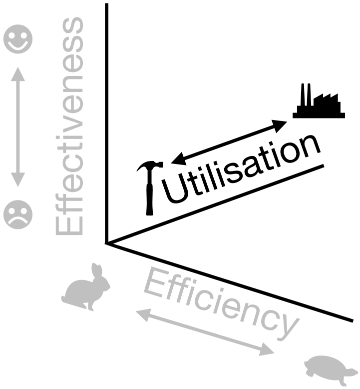
Retrieval Efficiency



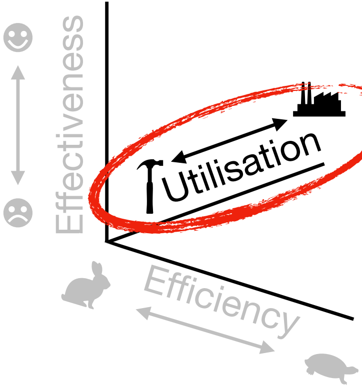
Retrieval Efficiency



Retrieval Efficiency

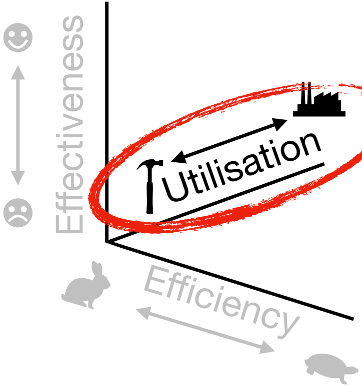


Retrieval Efficiency



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

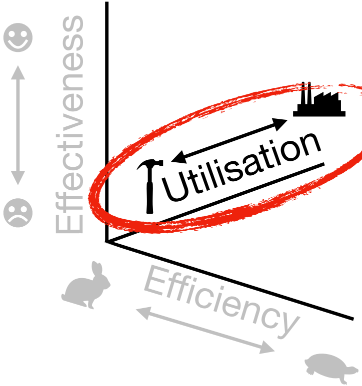
- **Orders of magnitude** more expensive to create and use
- Many applications: ranking, RAG, automatic assessment…

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:

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

Measuring Energy/Emissions

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

First, measure power consumption:

PUE

$$\underset{\textbf{watts}}{p_t} = \frac{\overset{\textbf{PUE}}{\Omega} \cdot t \cdot (p_c + p_r + p_g)}{1000}$$

Measuring Energy/Emissions

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

First, measure power consumption:

$$\underset{\text{watts}}{p_t} = \frac{\overset{\text{PUE}}{\Omega} \cdot \overset{\text{Running Time}}{t} \cdot (p_c + p_r + p_g)}{1000}$$

Measuring Energy/Emissions

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

First, measure power consumption:

PUE **Running Time** **CPU, RAM, GPU power draw**

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:

PUE

Running Time

CPU, RAM, GPU power draw

watts

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

Next, measure emissions:

Measuring Energy/Emissions

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

First, measure power consumption:

$$\underbrace{p_t}_{\text{watts}} = \frac{\underbrace{\Omega}_{\text{PUE}} \cdot \underbrace{t}_{\text{Running Time}} \cdot \underbrace{(p_c + p_r + p_g)}_{\text{CPU, RAM, GPU power draw}}}{1000}$$

Next, measure emissions:

$$\underbrace{\mathbf{kgCO_2e}}_{\text{emissions}} = \theta \cdot p_t$$

Measuring Energy/Emissions

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

First, measure power consumption:

$$\underset{\text{watts}}{p_t} = \frac{\underset{\text{PUE}}{\Omega} \cdot \underset{\text{Running Time}}{t} \cdot \underset{\text{CPU, RAM, GPU power draw}}{(p_c + p_r + p_g)}}{1000}$$

Next, measure emissions:

$$\underset{\text{emissions}}{\mathbf{kgCO_2e}} = \theta \cdot \underset{\text{Power consumption of experiments}}{p_t}$$

Measuring Energy/Emissions

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

First, measure power consumption:

$$\underbrace{p_t}_{\text{watts}} = \frac{\underbrace{\Omega}_{\text{PUE}} \cdot \underbrace{t}_{\text{Running Time}} \cdot \underbrace{(p_c + p_r + p_g)}_{\text{CPU, RAM, GPU power draw}}}{1000}$$

Next, measure emissions:

$$\underbrace{\mathbf{kgCO_2e}}_{\text{emissions}} = \underbrace{\theta}_{\text{avg. CO}_2\text{e (kg) per kWh where experiments took place}} \cdot \underbrace{p_t}_{\text{Power consumption of experiments}}$$

Measuring Energy/Emissions

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

First, measure power consumption:

PUE, Running Time, CPU, RAM, GPU power draw, watts

$$\underbrace{p_t}_{\text{watts}} = \frac{\underbrace{\Omega}_{\text{PUE}} \cdot \underbrace{t}_{\text{Running Time}} \cdot \underbrace{(p_c + p_r + p_g)}_{\text{CPU, RAM, GPU power draw}}}{1000}$$

Next, measure emissions:

$$\underbrace{\text{kgCO}_2\text{e}}_{\text{emissions}} = \underbrace{\theta}_{\text{avg. CO}_2\text{e (kg) per kWh where experiments took place}} \cdot \underbrace{p_t}_{\text{Power consumption of experiments}}$$

Emissions of my search engine:

$$\textbf{kgCO}_2\textbf{e} = \theta \cdot \Delta_q \cdot p_q$$

Measuring Energy/Emissions

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

First, measure power consumption:

$$\underset{\text{watts}}{p_t} = \frac{\overset{\text{PUE}}{\Omega} \cdot \overset{\text{Running Time}}{t} \cdot \overset{\text{CPU, RAM, GPU power draw}}{(p_c + p_r + p_g)}}{1000}$$

Next, measure emissions:

$$\underset{\text{emissions}}{\mathbf{kgCO_2e}} = \overset{\text{avg. CO}_2\text{e (kg) per kWh where experiments took place}}{\theta} \cdot \underset{\text{Power consumption of experiments}}{p_t}$$

Emissions of my search engine:

Technology	50th percentile (g CO_2-eq/ kWh_e)
Hydroelectric	4
Wind	12
Natural gas	469
Coal	1001

$$\mathbf{kgCO_2e} = \theta \cdot \Delta_q \cdot \underset{\textbf{Power consumption of a single query}}{p_q}$$

Measuring Energy/Emissions

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

First, measure power consumption:

PUE

Running Time

CPU, RAM, GPU power draw

watts

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

Next, measure emissions:

avg. CO_2e (kg) per kWh where experiments took place

emissions

$$\text{kgCO}_2\text{e} = \theta \cdot p_t$$

Power consumption of experiments

Emissions of my search engine:

Technology	50th percentile (g CO_2-eq/ kWh_e)
Hydroelectric	4
Wind	12
Natural gas	469
Coal	1001

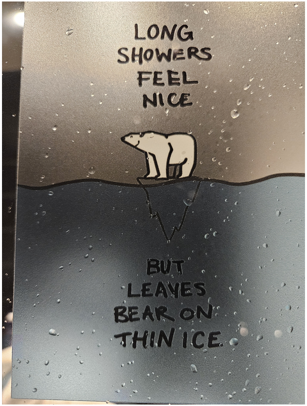
No. queries issued per unit time

$$\text{kgCO}_2\text{e} = \theta \cdot \Delta_q \cdot p_q$$

Power consumption of a single query

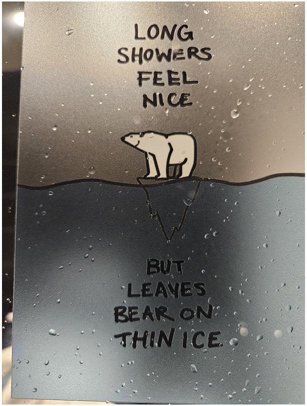
Measuring Energy/Emissions

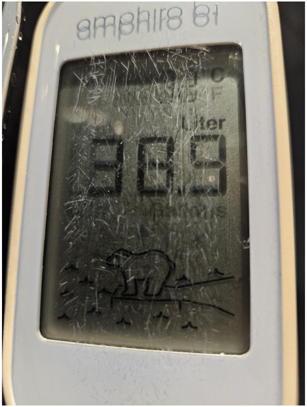
An Example: Shower for ca. 5 minutes



Measuring Energy/Emissions

An Example: Shower for ca. 5 minutes

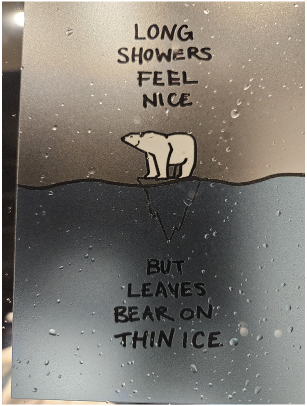




Water consumption 38.9 Liter

Measuring Energy/Emissions

An Example: Shower for ca. 5 minutes



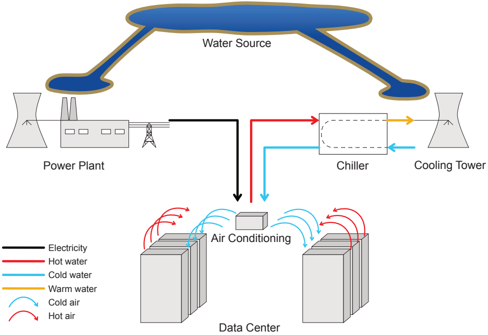
Water consumption 38.9 Liter

Assumed hydroelectric energy, this shower caused:

$4 \text{ gCO}_2\text{e} \cdot 1.4 \text{ kWH} = 5.6 \text{ gCO}_2\text{e} = 0.006 \text{ kgCO}_2\text{e}$

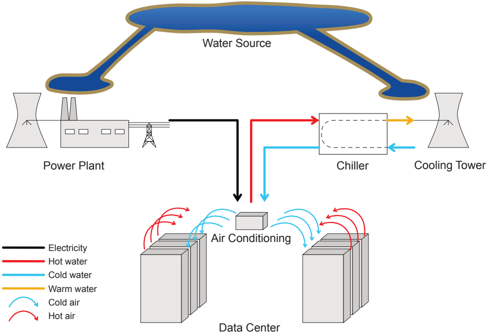
Measuring Water

Water ➔ measures **indirect** utilisation costs



Measuring Water

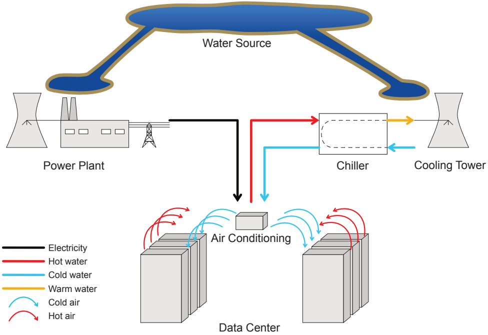
Water ➔ measures **indirect** utilisation costs



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

Measuring Water

Water → measures **indirect** utilisation costs

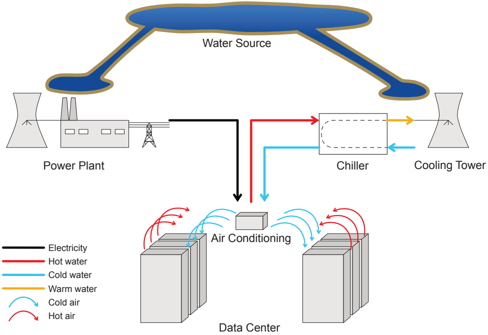


In data centers, water is consumed through **evaporation** and **blow down**

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 $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

Time

Energy used

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

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 $W_{\mathcal{M}} = W_{on}(\mathcal{M}) + W_{off}(\mathcal{M})$

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]

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

[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

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

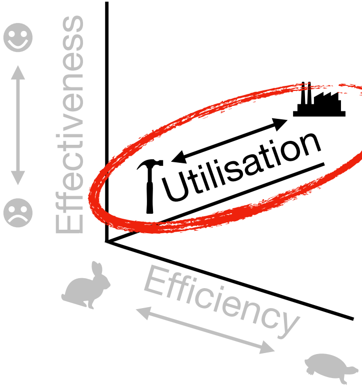
[2] Guido Zuccon et al. (2023). "Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models.". In: *ICTIR*, pp. 283–289.

Utilisation and Green IR

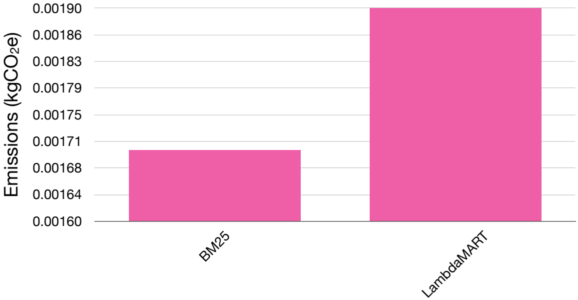
~~Okay, so what does this mean for IR?~~

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

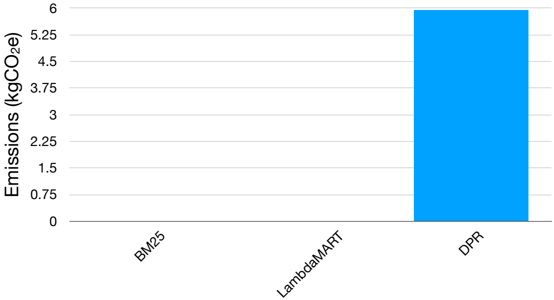
Okay, so show me what this means in IR research practice!



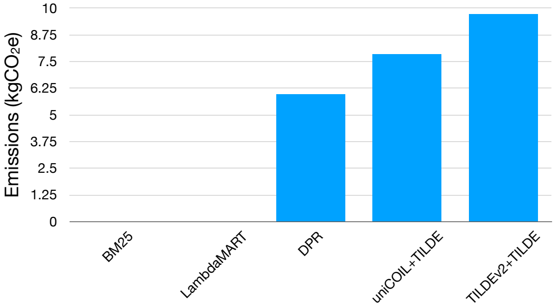
How many emissions produced to obtain a single result?



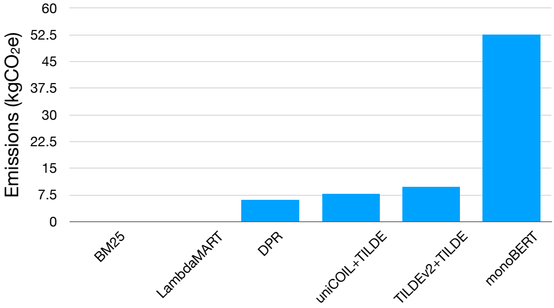
How many emissions produced to obtain a single result?



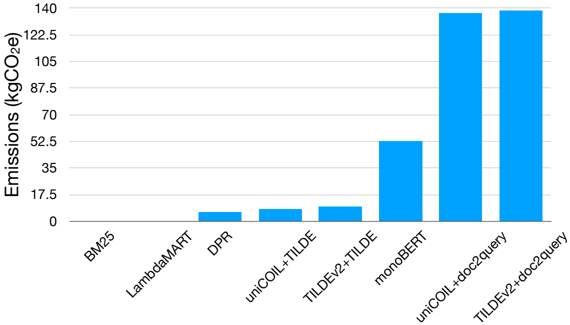
How many emissions produced to obtain a single result?



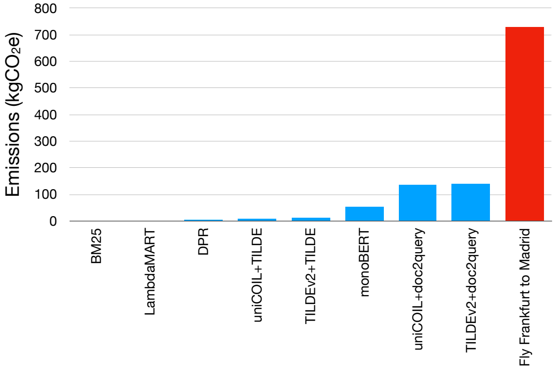
How many emissions produced to obtain a single result?



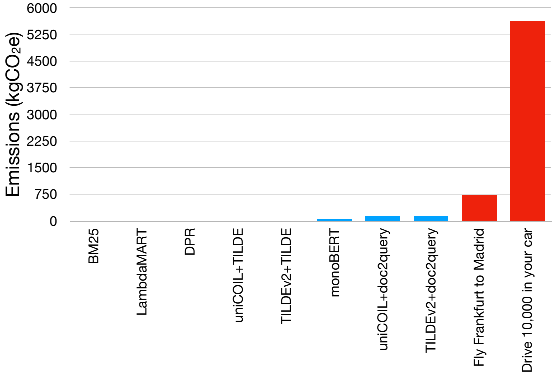
How many emissions produced to obtain a single result?



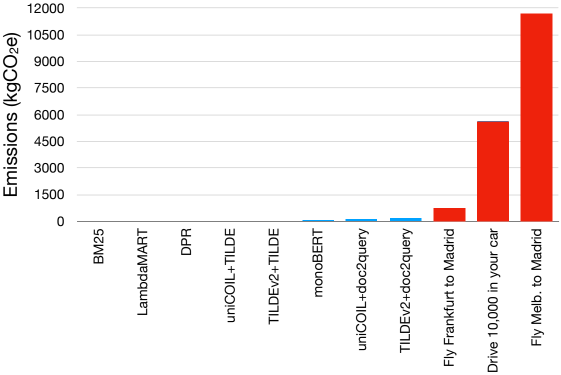
How many emissions produced to obtain a single result?



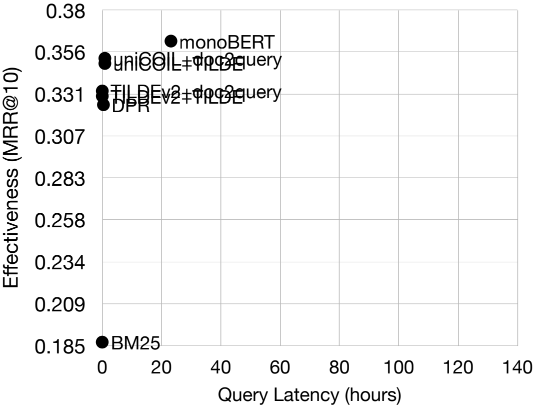
How many emissions produced to obtain a single result?



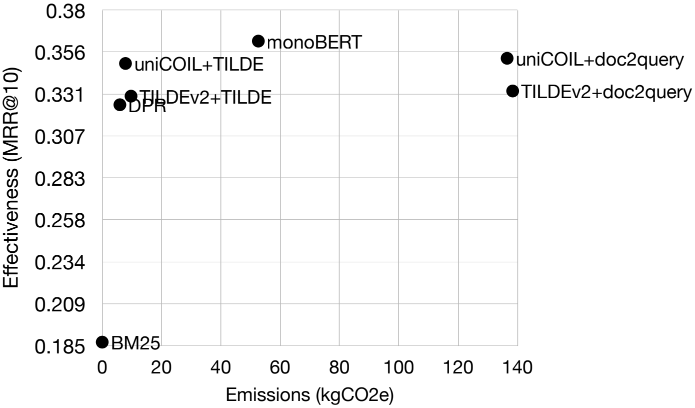
How many emissions produced to obtain a single result?



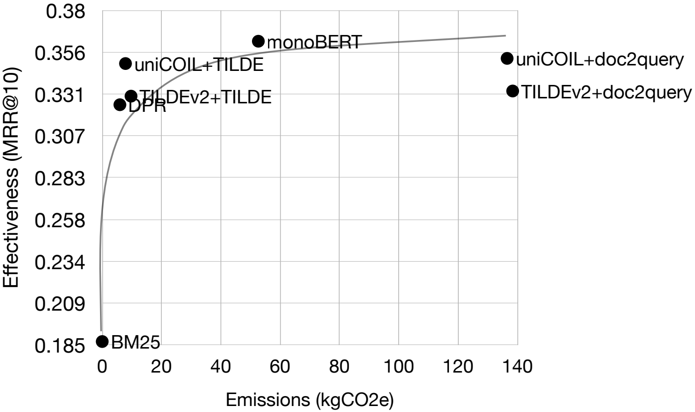
How many emissions produced to obtain a single result?



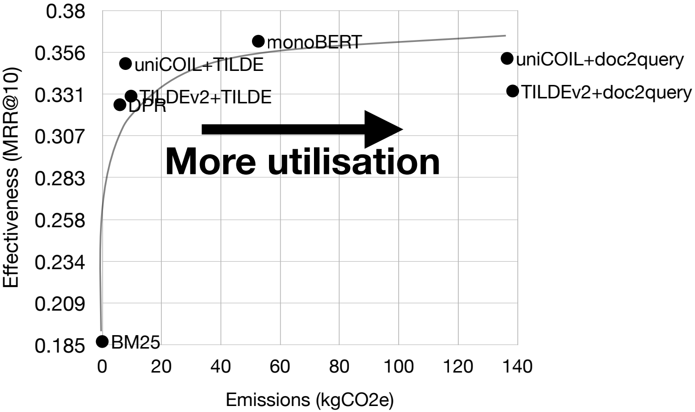
How many emissions produced to obtain a single result?



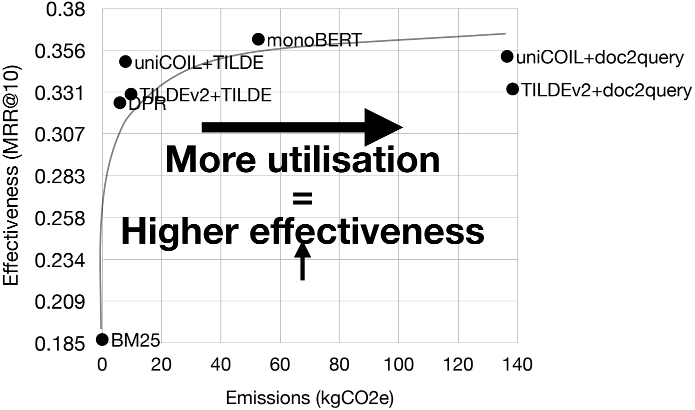
How many emissions produced to obtain a single result?



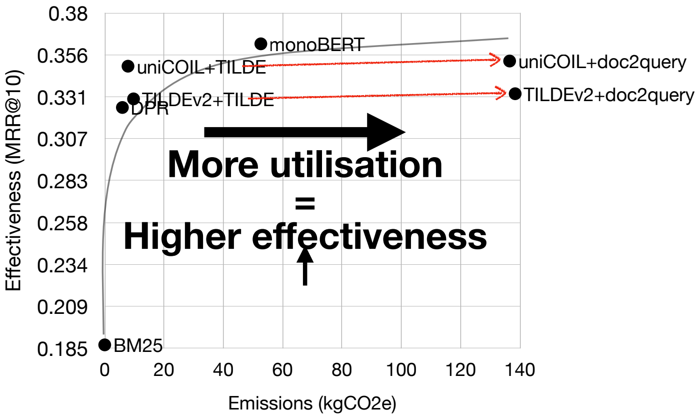
How many emissions produced to obtain a single result?



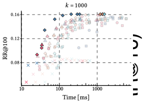
How many emissions produced to obtain a single result?

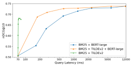


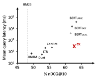
How many emissions produced to obtain a single result?

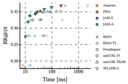


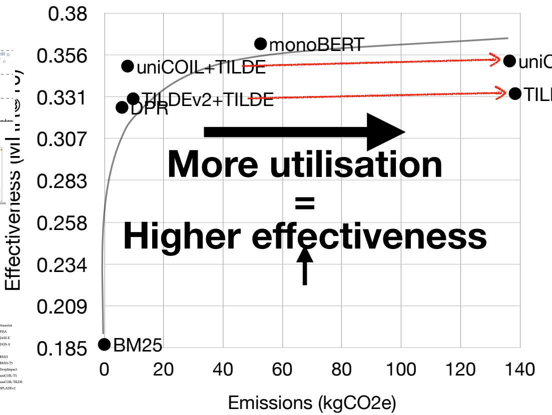
How many emissions produced to obtain a single result?

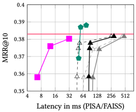


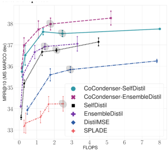




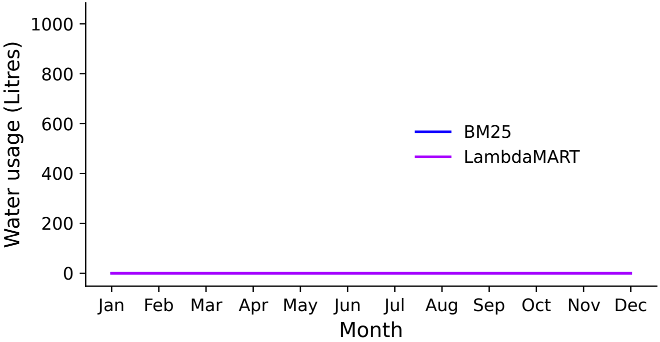




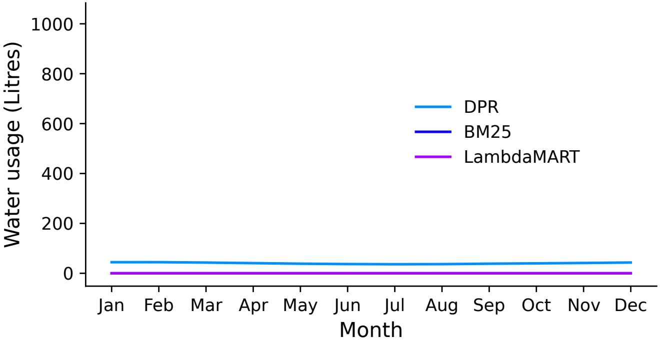




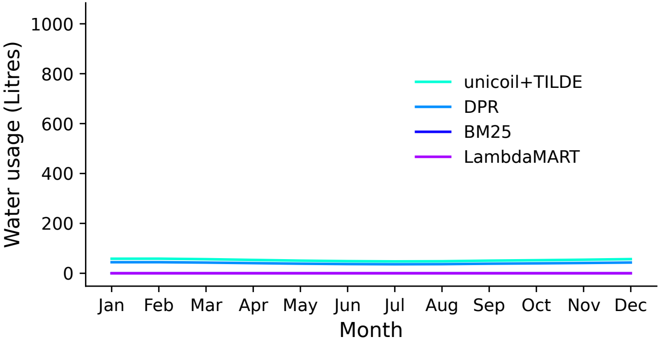
How much water used to produced to obtain a single result?



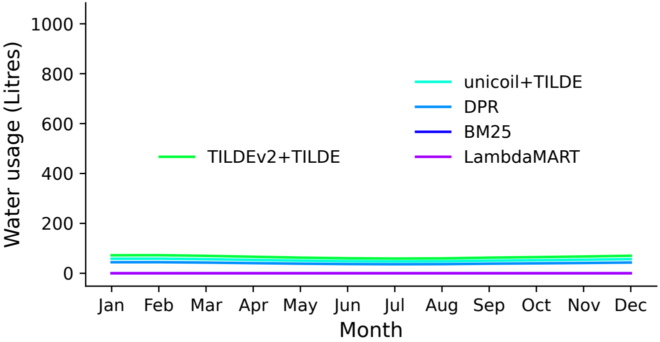
How much water used to produced to obtain a single result?



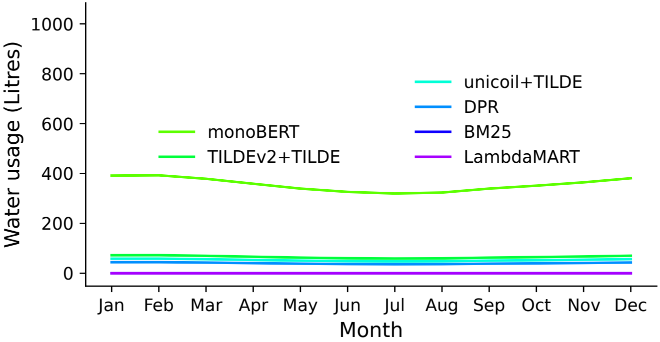
How much water used to produced to obtain a single result?



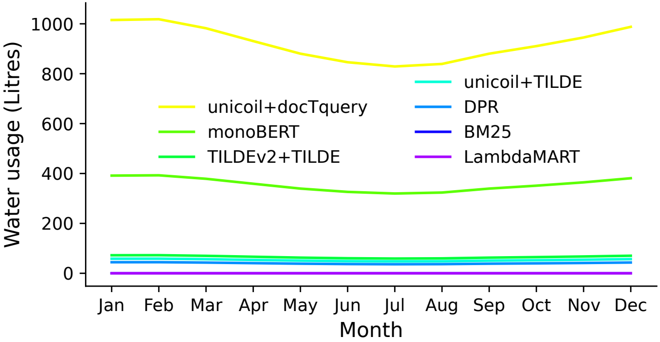
How much water used to produced to obtain a single result?



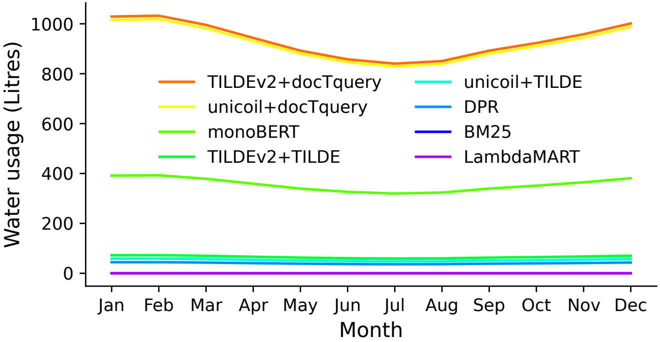
How much water used to produced to obtain a single result?



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



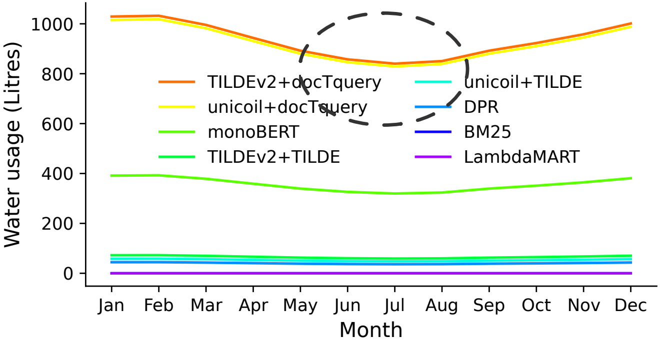
How much water used to produced to obtain a single result?



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

Time of year is important to how much water is used

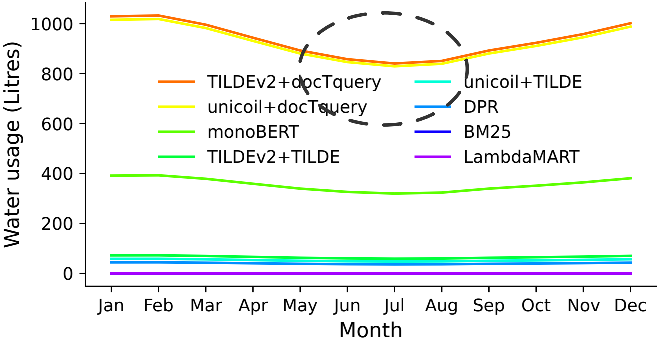
experiments performed in Australia



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

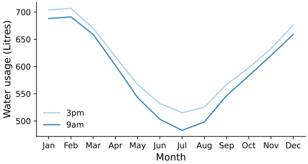
Time of year is important to how much water is used

experiments performed in Australia

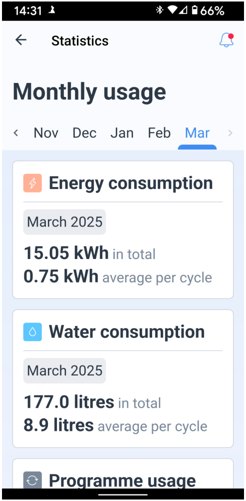


Time of day is equally important

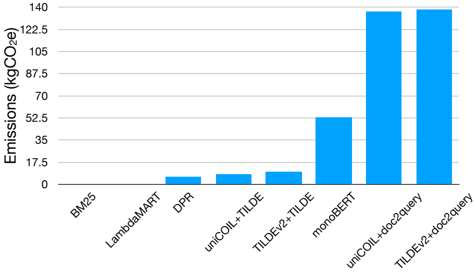
TILDEv2+docTquery

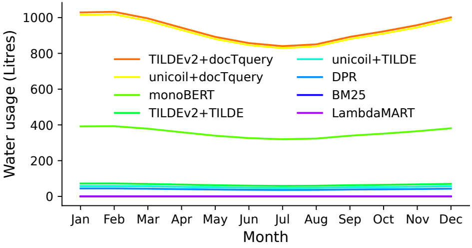


Is your model better than a dishwasher?



ca. 11.53kgCO_2e





Overview of Green IR

Measuring Utilisation

Corpus Subsampling

Retrieval Effectiveness

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

Retrieval Effectiveness

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

Problem: Our evaluation will always give us some number

- Is this number meaningful?

Retrieval Effectiveness

Evaluate how good 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 good 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
[Voorhees'19]

- Subjectiveness of relevance judgments
- Incompleteness of relevance judgments

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

hydrogen liquid at what temperature?

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

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

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

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

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

CHAT NOIR

hydrogen liquid at what temperature?

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

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

VS.

Liquid hydrogen

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

Retrieval Effectiveness

Problem (1): Relevance judgments are highly subjective

[Burgin'92; Lesk'68; Voorhees'00]

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

VS.

Impact of disagreement on system rankings:

- Human relevance assessors disagree substantially
- Impact on system rankings is negligible

Retrieval Effectiveness

Problem (2): Incompleteness of relevance judgments

Retrieval Effectiveness

Problem (2): Incompleteness of relevance judgments

CHAT NOIR

hydrogen liquid at what temperature?

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

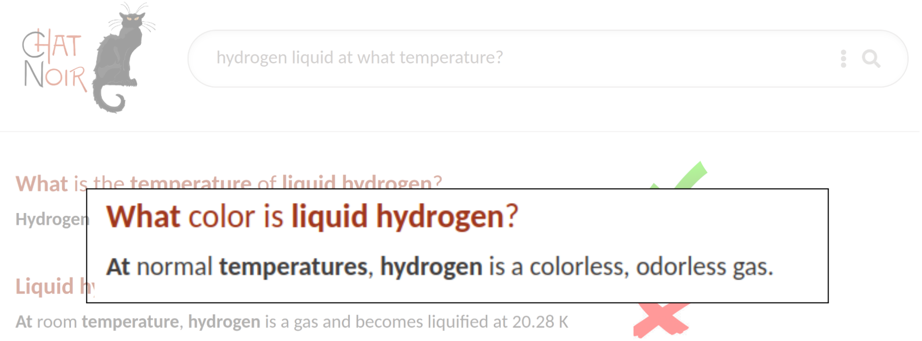
Hydrogen becomes **liquid** at −252.87 °C

Liquid hydrogen ✗

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

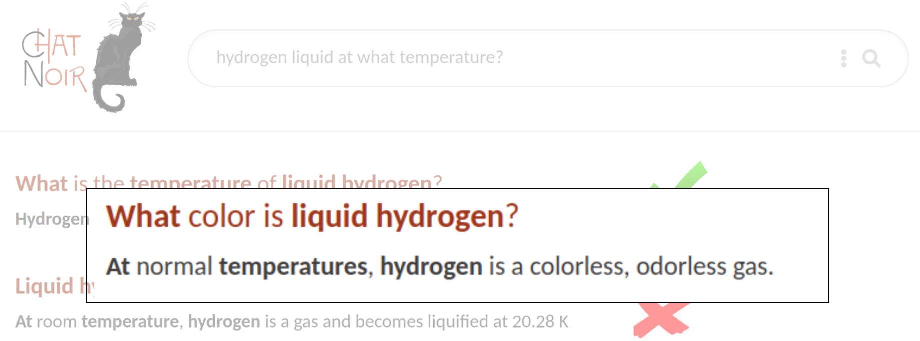
Retrieval Effectiveness

Problem (2): Incompleteness of relevance judgments



Retrieval Effectiveness

Problem (2): Incompleteness of relevance judgments



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

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

Measure Reliability of Experiments [Breuer'20]

Ranking correlations can confirm the reliability of evaluations

Measure 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 > Sytem B > System C > System D

Measure 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 > Sytem B > System C > System D

Step 2: Repeat the experiment

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

Measure 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 > Sytem B > System C > System D

Step 2: Repeat the experiment

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

Example:

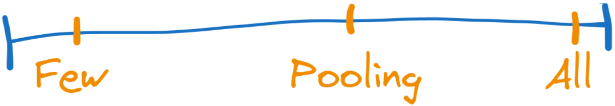
New System Ranking	τ
System A > Sytem B > System C > System D	1.0
System A > Sytem B > System D > System C	0.8
System D > Sytem C > System B > System A	-1.0

Goal: **Green** and **Reliable** IR Experiments

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

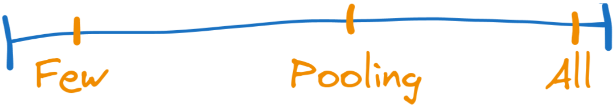
How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



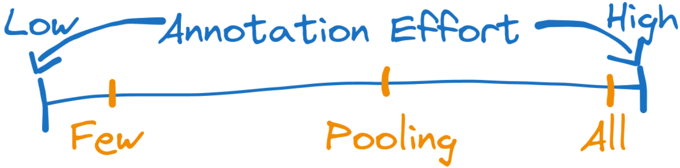
Few Judgments: E.g., one relevant document derived via click logs.

Top-k Pooling:

- Multiple teams develop retrieval systems independent of each other
- Judge the top-k results of each system (usually graded)

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



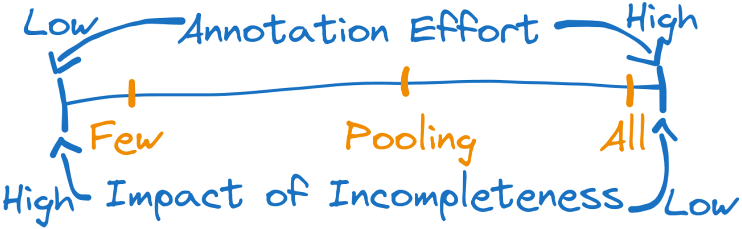
Few Judgments: E.g., one relevant document derived via click logs.

Top-k Pooling:

- Multiple teams develop retrieval systems independent of each other
- Judge the top-k results of each system (usually graded)

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



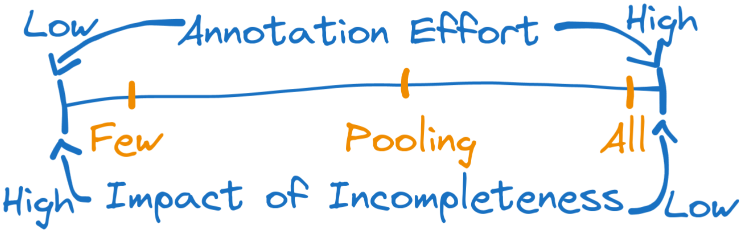
Few Judgments: E.g., one relevant document derived via click logs.

Top-k Pooling:

- Multiple teams develop retrieval systems independent of each other
- Judge the top-k results of each system (usually graded)

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



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

Top-k Pooling:

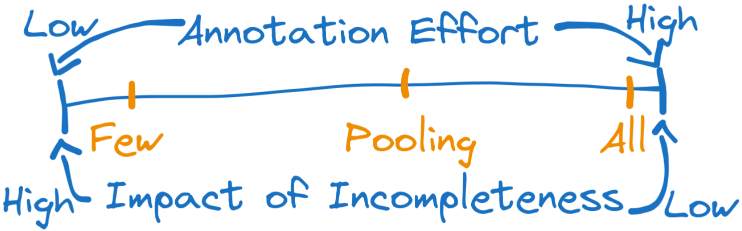
- Multiple teams develop retrieval systems independent of each other
- Judge the top-*k* results of each system (usually graded)

How many different rankings?

Labels				Top-10 Rankings
0	1	2	3	
∞	1	—	—	11
∞	10	10	10	$4^{10} > 1$ million

How build our Evaluation Dataset? Step 1: Queries

Many queries with few judgments or few queries with many judgments?



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

Top-k Pooling:

- Multiple teams develop retrieval systems independent of each other
- Judge the top-*k* results of each system (usually graded)

How many different rankings?

Pooling advantageous from Green IR Perspective

Labels				Top-10 Rankings
0	1	2	3	
∞	1	—	—	11
∞	10	10	10	$4^{10} > 1$ million

How build our Evaluation Dataset? 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 Evaluation Dataset? 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 [Mei'08]
- We do not need to include all relevant documents
- We only need a subset that allows reliable evaluations

How build our Evaluation Dataset? 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 [Mei'08]
- 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

Judgment Pool:

- Select all documents with a judgment. 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

Judgment Pool:

- Select all documents with a judgment. 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

Judgment Pool:

- Select all documents with a judgment. 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

Judgment Pool + Random

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

How build our Evaluation Dataset? Step 2: Documents

Judgment Pool:

- Select all documents with a judgment. 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

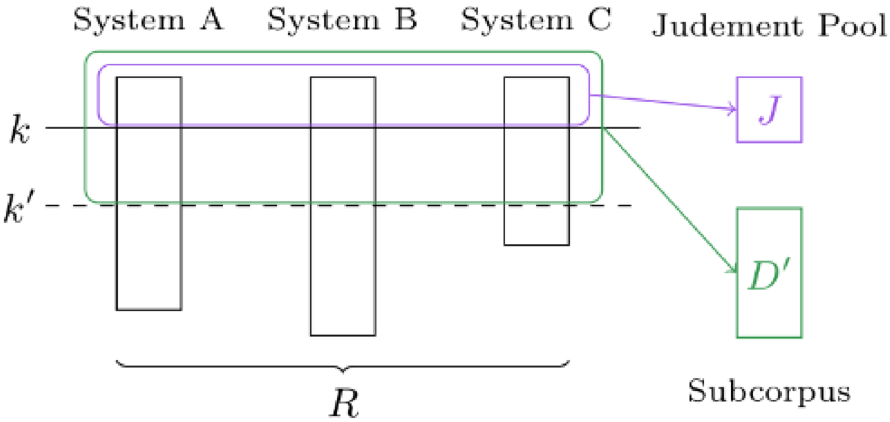
Judgment Pool + Random

- All documents with a judgment 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 judgment pool
- Advantage: Incorporates all query interpretations. Can use all above.

How build our Evaluation Dataset? Step 2: Documents



Evaluation of Corpus Subsampling

Reliability: What subsampling approaches yield robust observations?

Step 1: Create ground truth system rankings

- Use complete judgment pool to evaluate all systems from all teams

Evaluation of Corpus Subsampling

Reliability: What subsampling approaches yield robust observations?

Step 1: Create ground truth system rankings

- Use complete judgment pool to evaluate all systems from all teams

Step 2: Repeat Experiments with Leave-one-Group-out method

- For each team, assume all systems of the team did not participate
- Remove all documents from the judgment pool and corpus solely retrieved by the left-out team
- Yields incomplete judgment pool and incomplete corpus subsample
- Re-Evaluate all systems/teams with incomplete judgments/corpus
- How similar is the new system ranking with the ground-truth?
  - Best result: system rankings are identical, i.e., τ = 1.0

Evaluation of Corpus Subsampling

Experimental Setup:

- We run the subsampling approaches on 9 evaluation campaigns
  - 4 on ClueWeb09: 1.0 billion documents (4.0 TB)
  - 2 on ClueWeb12: 0.7 billion documents (4.5 TB)
  - 1 on Robust04: 0.5 million documents (0.6 GB)
  - 2 on MS MARCO: 8.8 million documents (2.9 GB)

Evaluation of Corpus Subsampling

Experimental Setup:

– We run the subsampling approaches on 9 evaluation campaigns
  - 4 on ClueWeb09: 1.0 billion documents (4.0 TB)
  - 2 on ClueWeb12: 0.7 billion documents (4.5 TB)
  - 1 on Robust04: 0.5 million documents (0.6 GB)
  - 2 on MS MARCO: 8.8 million documents (2.9 GB)

Results:

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

Evaluation of Corpus Subsampling

Experimental Setup:

- We run the subsampling approaches on 9 evaluation campaigns
  - 4 on ClueWeb09: 1.0 billion documents (4.0 TB)
  - 2 on ClueWeb12: 0.7 billion documents (4.5 TB)
  - 1 on Robust04: 0.5 million documents (0.6 GB)
  - 2 on MS MARCO: 8.8 million documents (2.9 GB)

Re-Pooling does not overestimate:

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

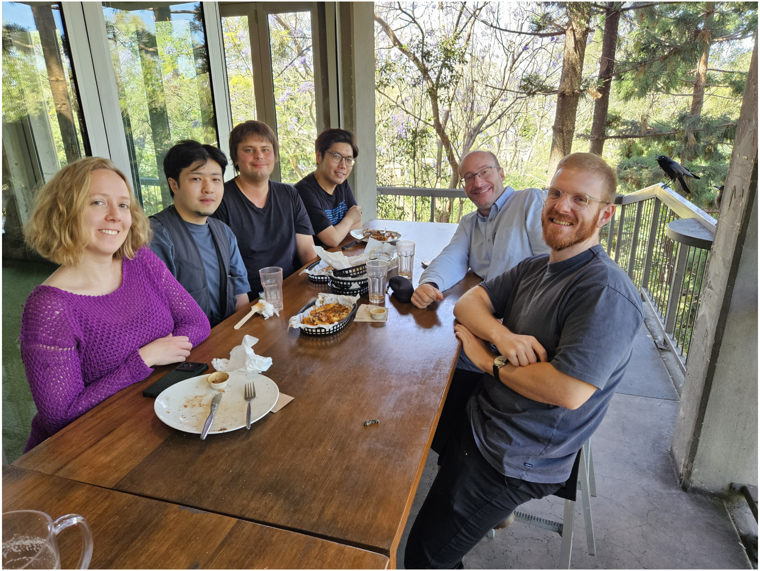
Corpus Subsampling

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

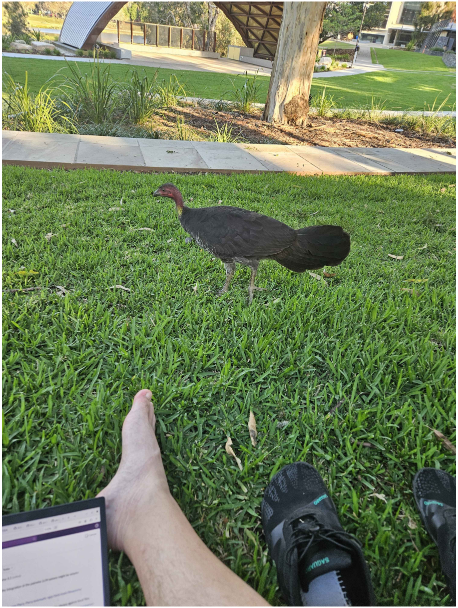
Some other Side-Aspects of the work :)

Some other Side - Aspects of the work :)



Some other Side - Aspects of the work :)

Some other Side - Aspects of the work :)



Conclusion and Future Work

- There are many diverse ways to measure efficiency and utilization
- The emissions of our experiments is not negligible
- Averages can hide many things: Is an evaluation reliable?
- From the perspective of GreenIR:
  - Many (pooled) judgments per query > one/few judgments per query
  - Corpus subsampling: reliable evaluation orders of magnitude fewer documents

Future Work

- Can corpus subsampling be incorporated into evaluation campaigns?
- How to do holistic evaluations that combine efficiency with effectiveness?
- Upcoming workshop on that: ReNeuIR 2025 at SIGIR