Efficiency and Energy in Neural Information Retrieval

Harry Scells
Alexander von Humboldt Research Fellow
Leipzig University
https://scells.me

TH Köln · April 24, 2024

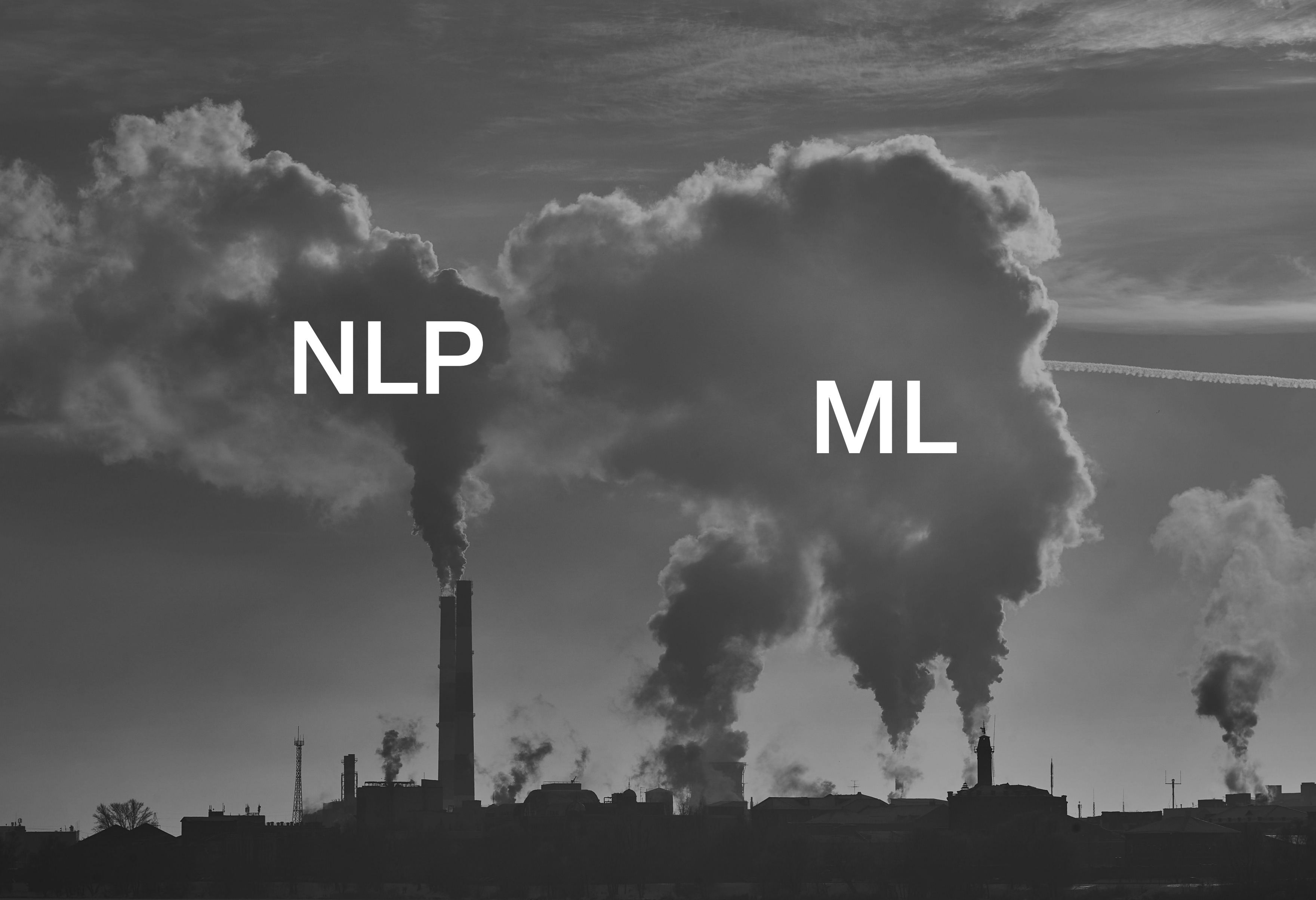
1. Green IR

   [Scells et al. 2022]

2. Efficient Listwise Neural Search

   [Schlatt et. al 2024]

3. Estimating Cost of IR (discussion)



[1] Strubell, E. et al. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*

Green IR

Why?

Large (pre-trained) neural language models

- ❑ 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 [Zuccon et al. 2023])

[1] Strubell, E. et al. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*

Green IR

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

[1] Strubell, E. et al. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*

Green IR

Retrieval Efficiency

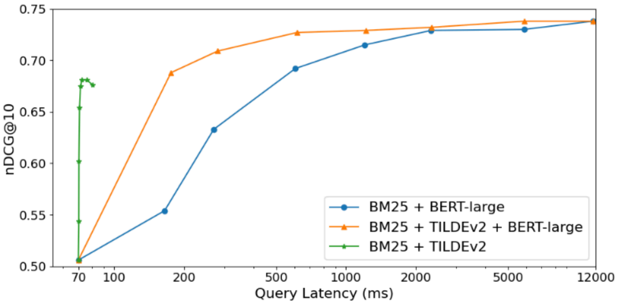
Speed a system can retrieve relevant information in response to a query.

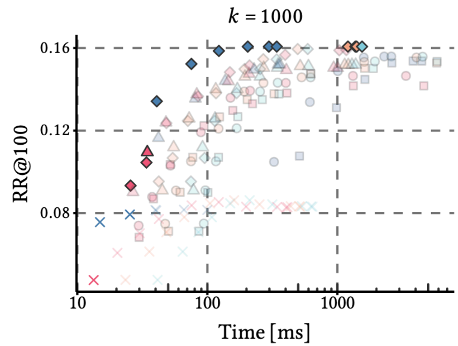
Factors that can impact retrieval efficiency include:

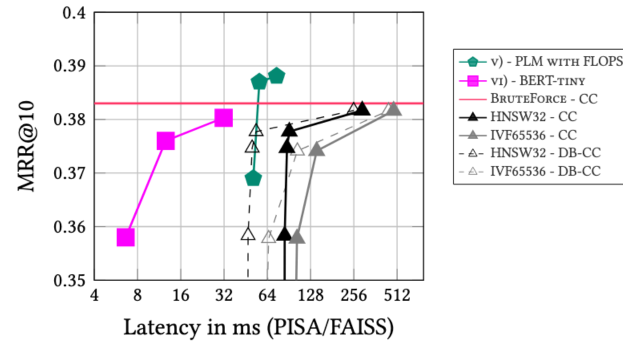
- **Size and complexity** of the corpus being searched
- Effectiveness of the **retrieval models** or techniques being used
- Efficiency of the **hardware and infrastructure** used

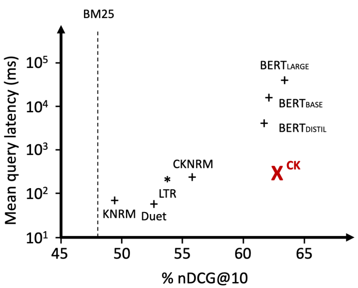
Green IR

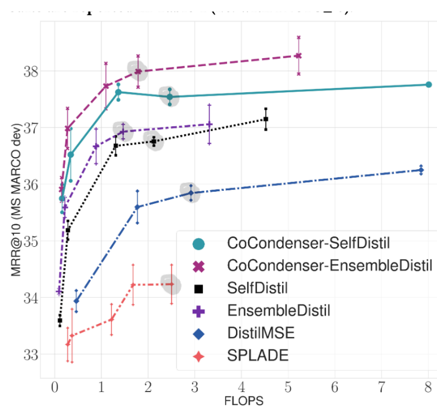
Retrieval Efficiency

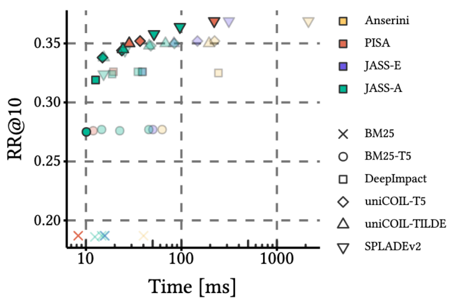


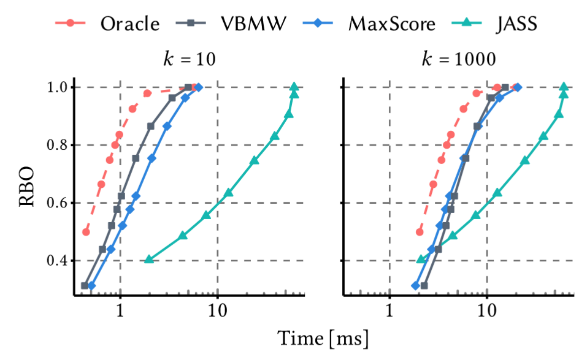






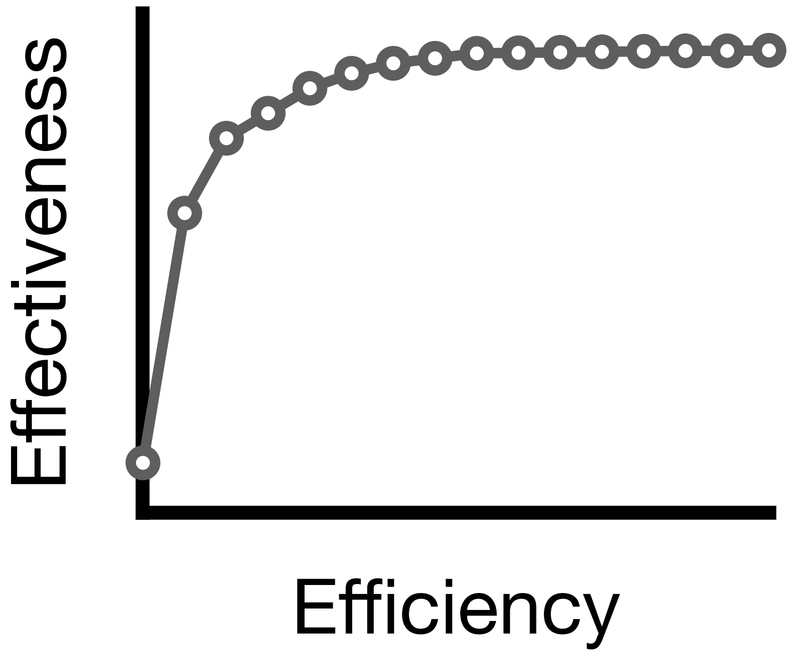






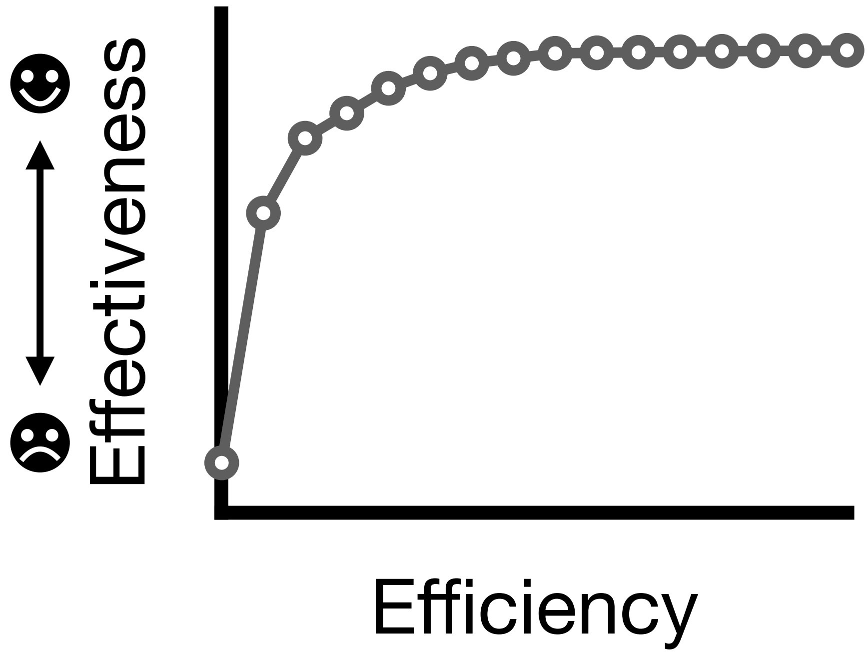
Green IR

Retrieval Efficiency



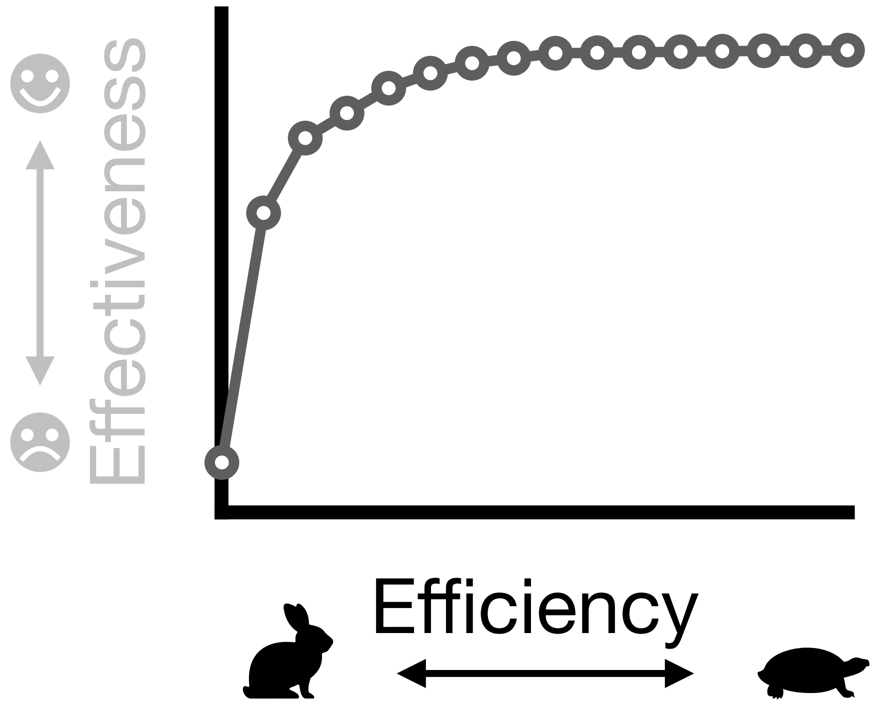
Green IR

Retrieval Efficiency



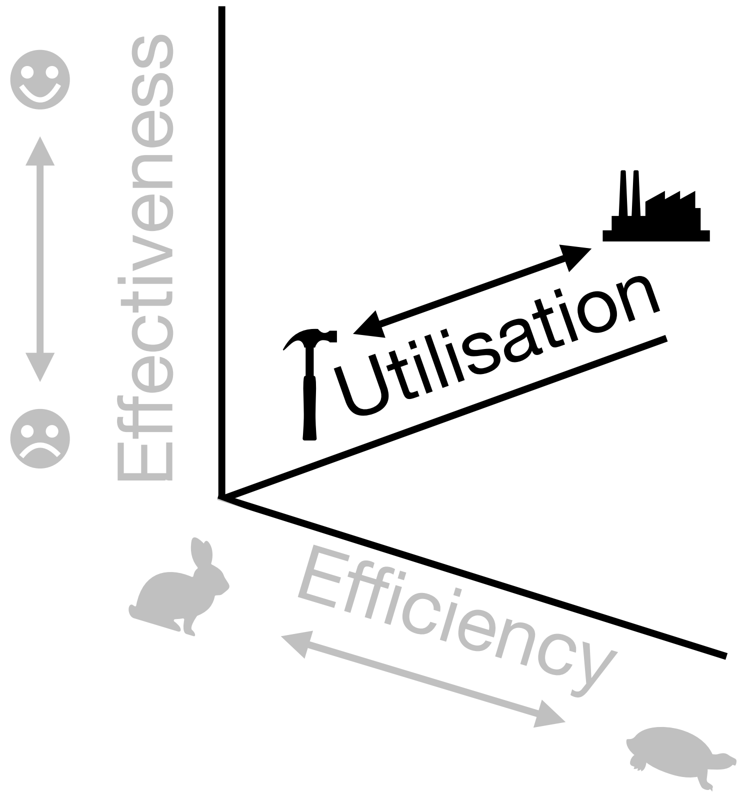
Green IR

Retrieval Efficiency



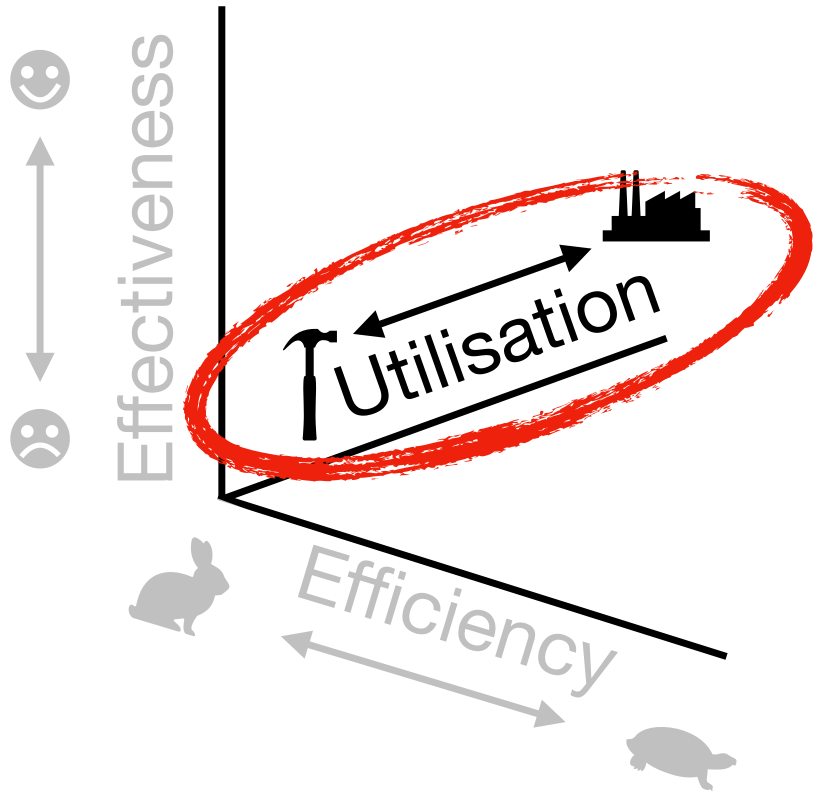
Green IR

Retrieval Efficiency



Green IR

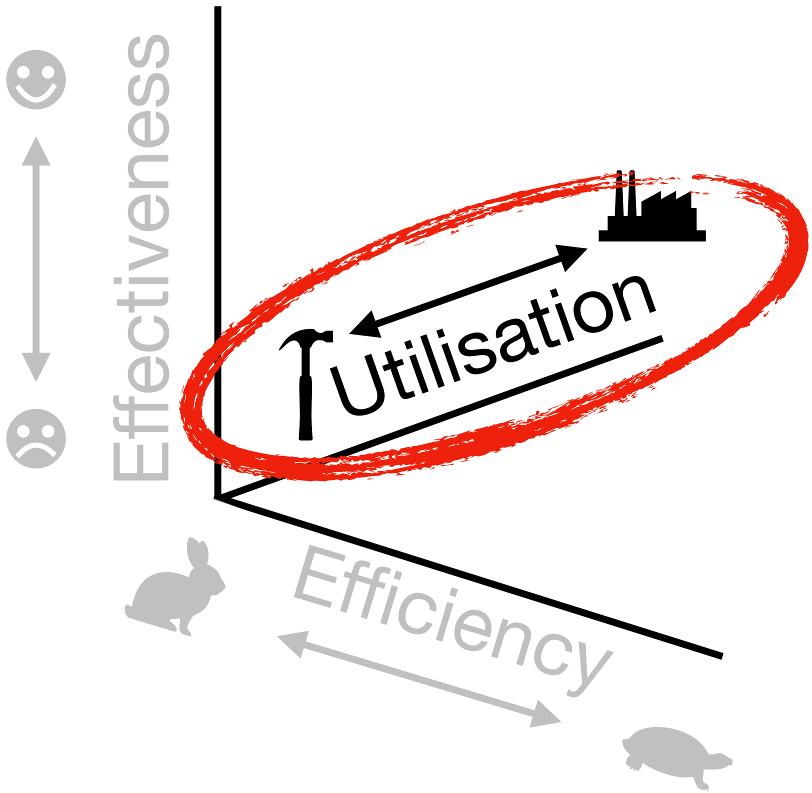
Retrieval Efficiency



Green IR

Retrieval Efficiency

Okay, so what does this mean for IR?



Green 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"

(Schwartz, R. et al. 2020. Green AI. Communications of the ACM)

Green 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"

(Schwartz, R. et al. 2020. Green AI. Communications of the ACM)

Neural methods require pre-trained LMs

- **Expensive** to create
- Becoming even more expensive (see: DSI and friends)

Green 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"

(Schwartz, R. et al. 2020. Green AI. Communications of the ACM)

Neural methods require pre-trained LMs

- ❑ **Expensive** to create
- ❑ Becoming even more expensive (see: DSI and friends)

Pre-trained LMs come at a high power and emissions cost

Green 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"

(Schwartz, R. et al. 2020. Green AI. Communications of the ACM)

Neural methods require pre-trained LMs

- **Expensive** to create
- Becoming even more expensive (see: DSI and friends)

Pre-trained LMs come at a high power and emissions cost

Missing dimension of IR evaluation

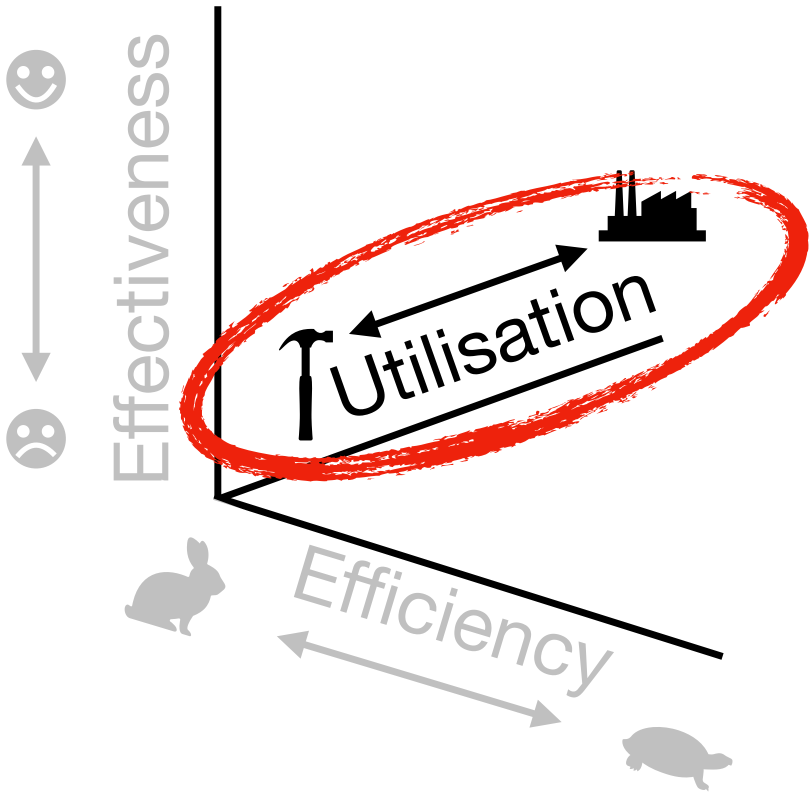
- Effectiveness
- Efficiency
- **Utilisation**

Green IR

Utilisation and Green IR

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

Okay, so how can I measure this?



First, measure power consumption:

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

Green IR

Measuring Emissions

First, measure power consumption:

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

Green IR

Measuring Emissions

First, measure power consumption:

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

Green IR

Measuring Emissions

First, measure power consumption:

PUE

Running Time

watts

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

Green IR

Measuring Emissions

First, measure power consumption:

PUE

Running Time

CPU, RAM, GPU power draw

watts

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

Green IR

Measuring Emissions

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:

Green IR

Measuring Emissions

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:

$$\underbrace{\textbf{kgCO}_2\textbf{e}}_{\text{emissions}} = \theta \cdot p_t$$

Green IR

Measuring Emissions

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:

emissions

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

Power consumption of experiments

Green IR

Measuring Emissions

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

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

Power consumption of experiments

Green IR

Measuring Emissions

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

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

Power consumption of experiments

Emissions of my search engine:

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

Green IR

Measuring Emissions

First, measure power consumption:

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

Next, measure emissions:

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

Emissions of my search engine:

$$\mathbf{kgCO_2e} = \theta \cdot \Delta_q \cdot \underbrace{p_q}_{\text{Power consumption of a single query}}$$

Green IR

Measuring Emissions

First, measure power consumption:

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

Next, measure emissions:

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

Emissions of my search engine:

$$\textbf{kgCO}_2\textbf{e} = \theta \cdot \underbrace{\Delta_q}_{\text{No. queries issued per unit time}} \cdot \underbrace{p_q}_{\text{Power consumption of a single query}}$$

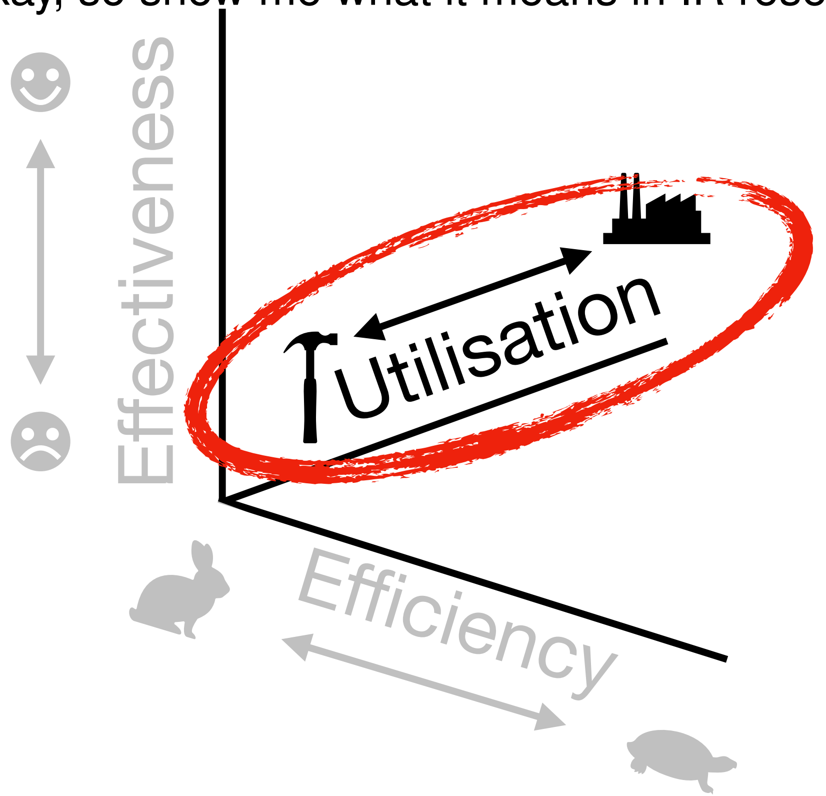
Green IR

Utilisation and Green IR

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

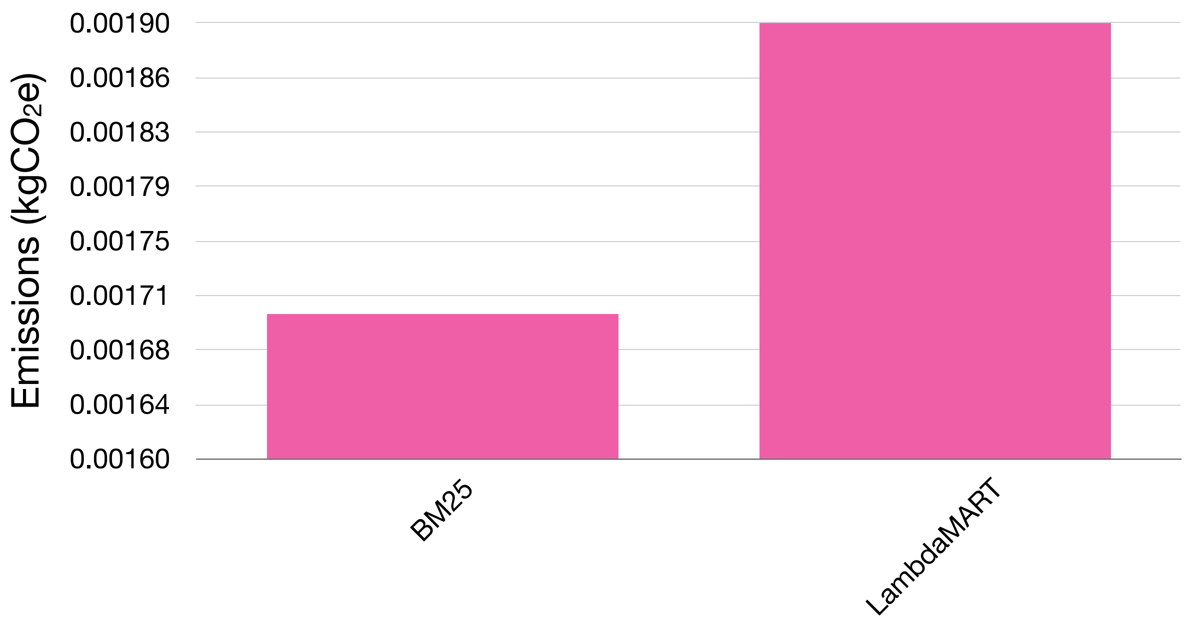
~~Okay, so how can I measure this?~~

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



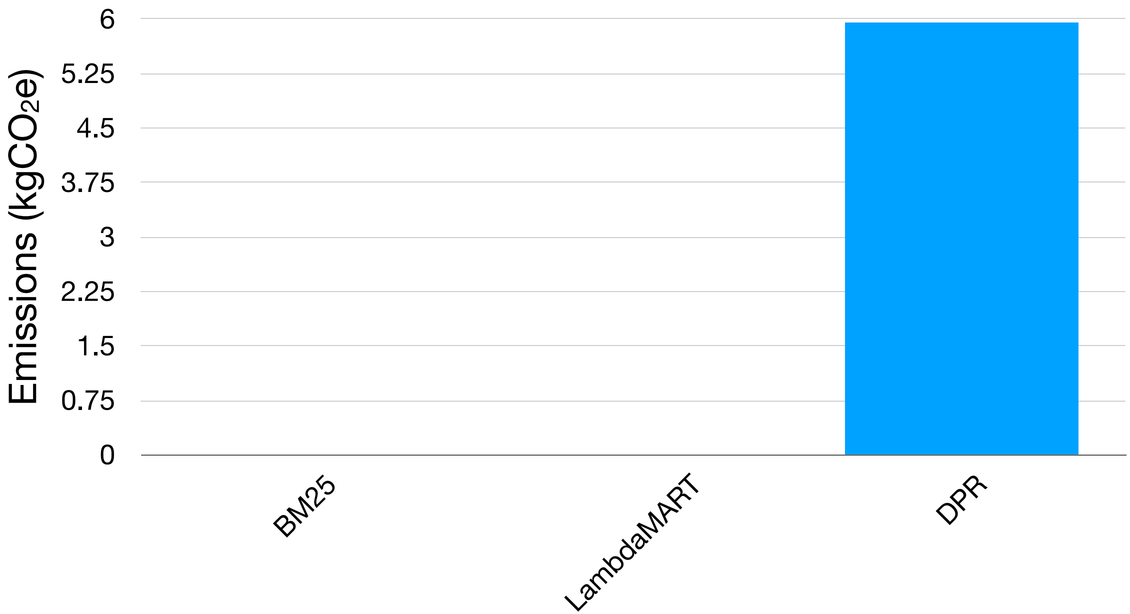
Green IR

How many emissions produced to obtain a single result?



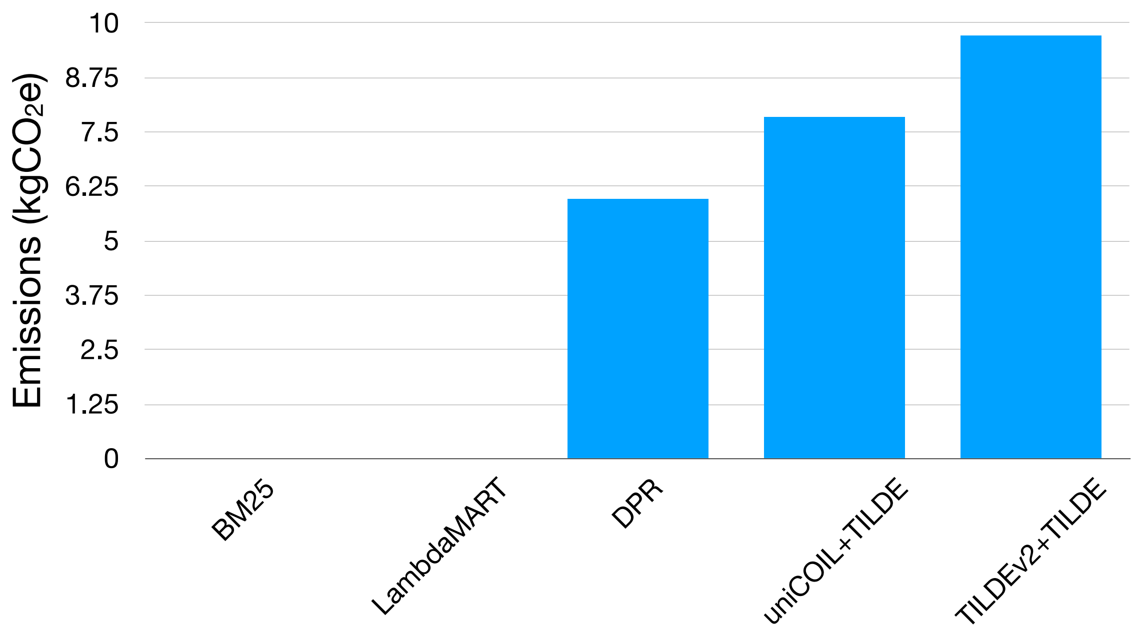
Green IR

How many emissions produced to obtain a single result?



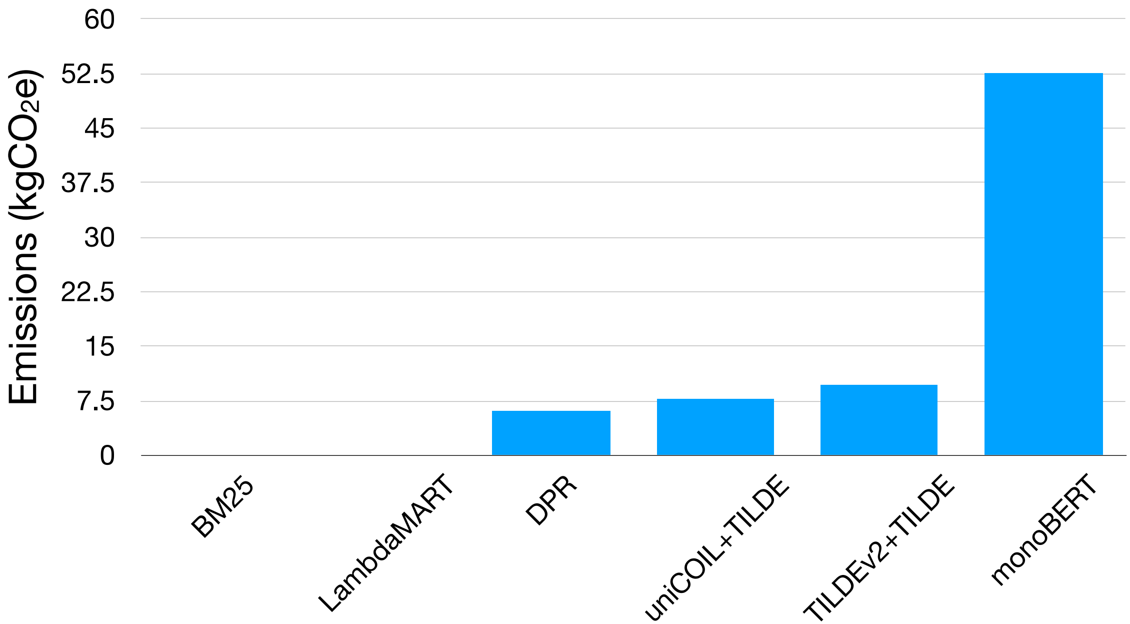
Green IR

How many emissions produced to obtain a single result?



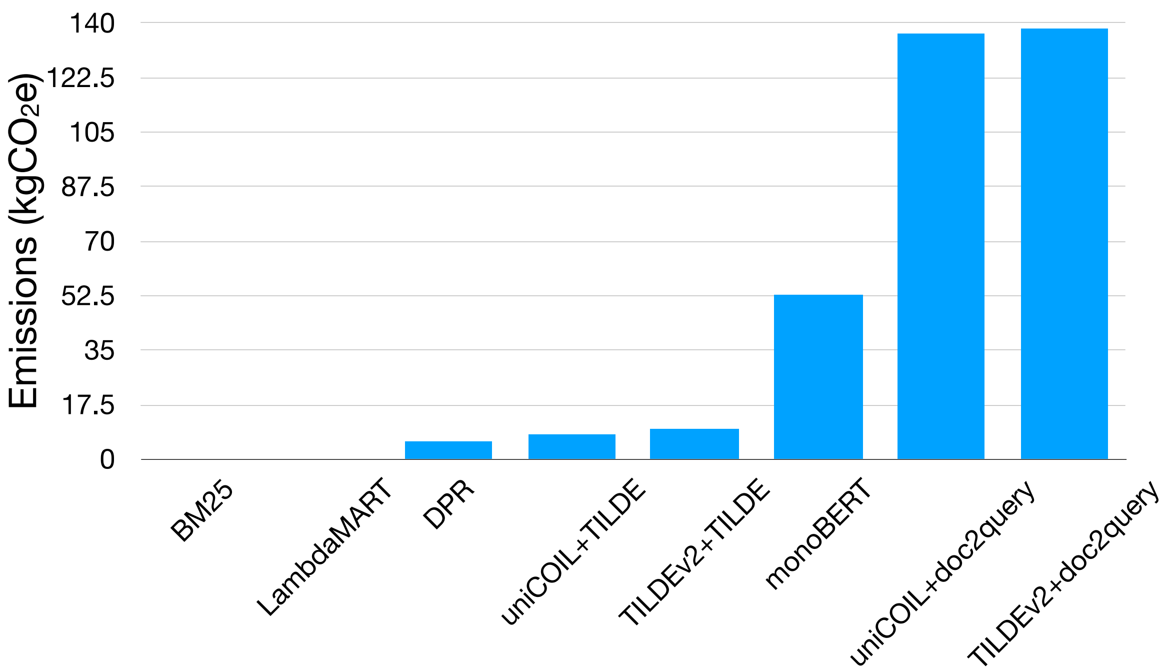
Green IR

How many emissions produced to obtain a single result?



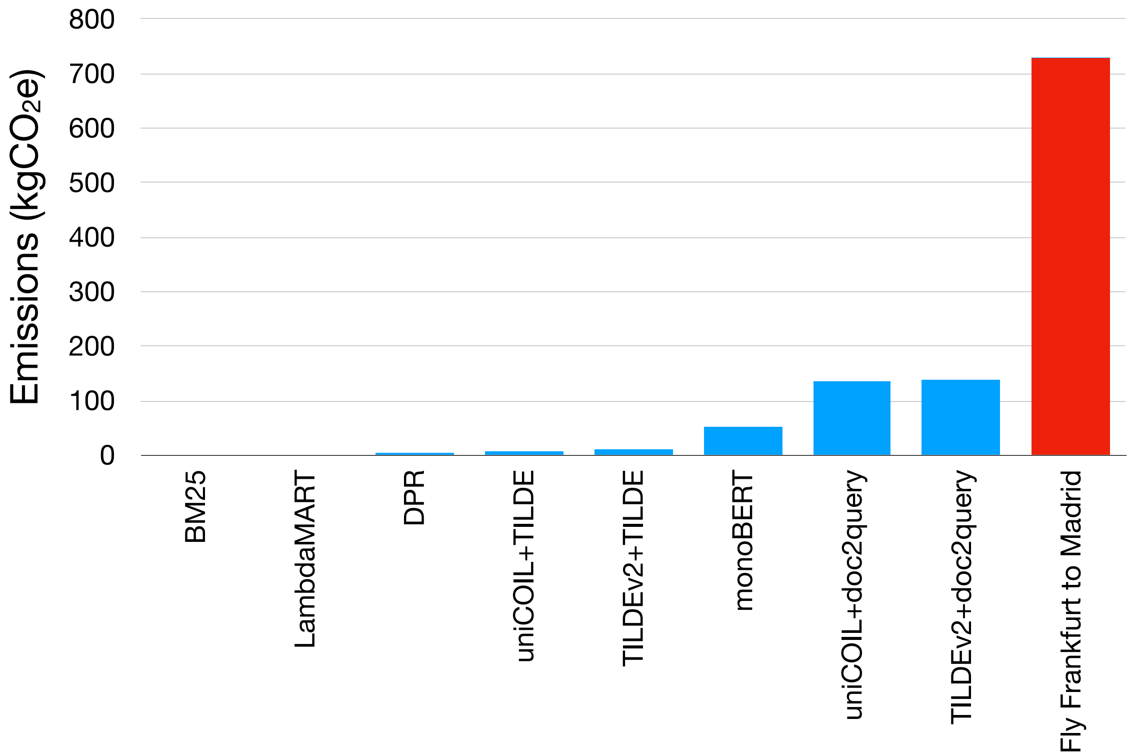
Green IR

How many emissions produced to obtain a single result?



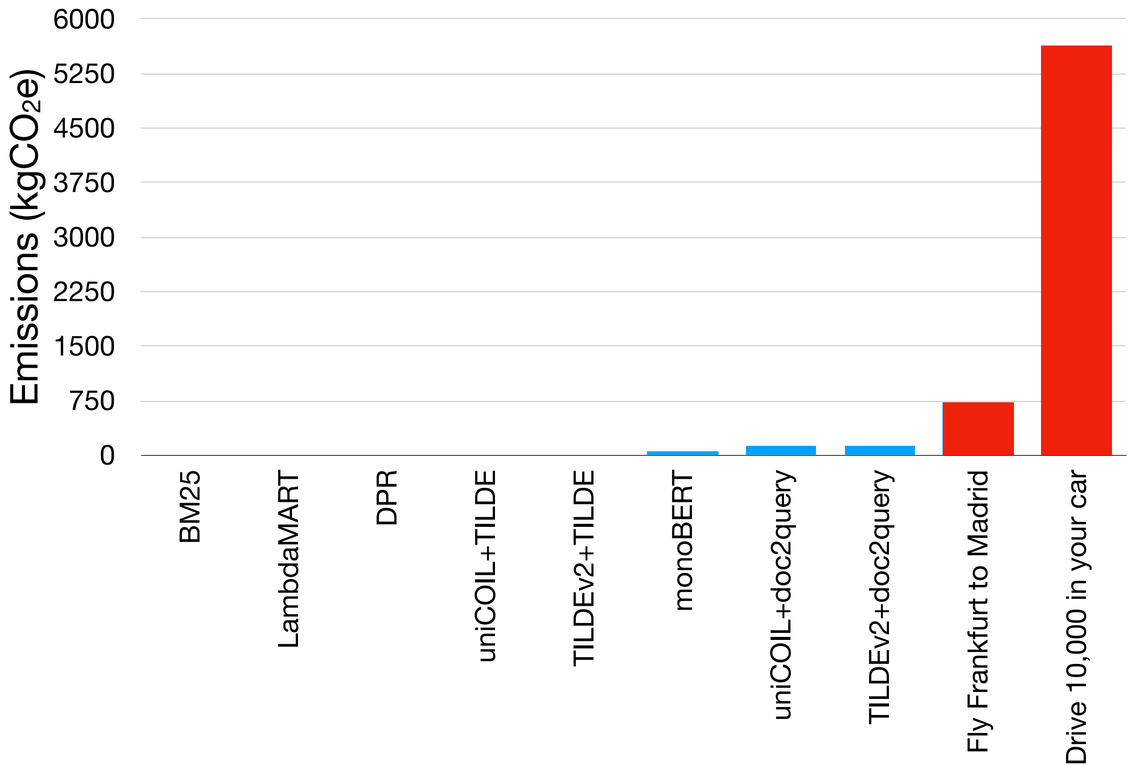
Green IR

How many emissions produced to obtain a single result?



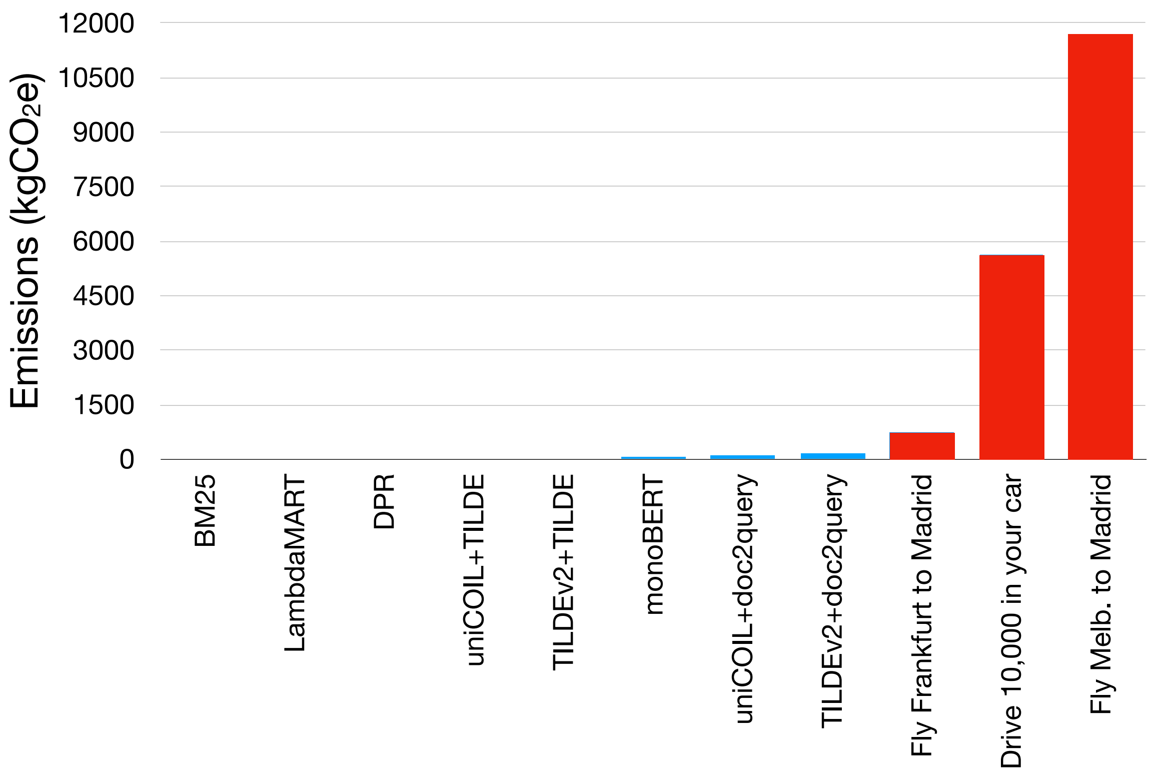
Green IR

How many emissions produced to obtain a single result?



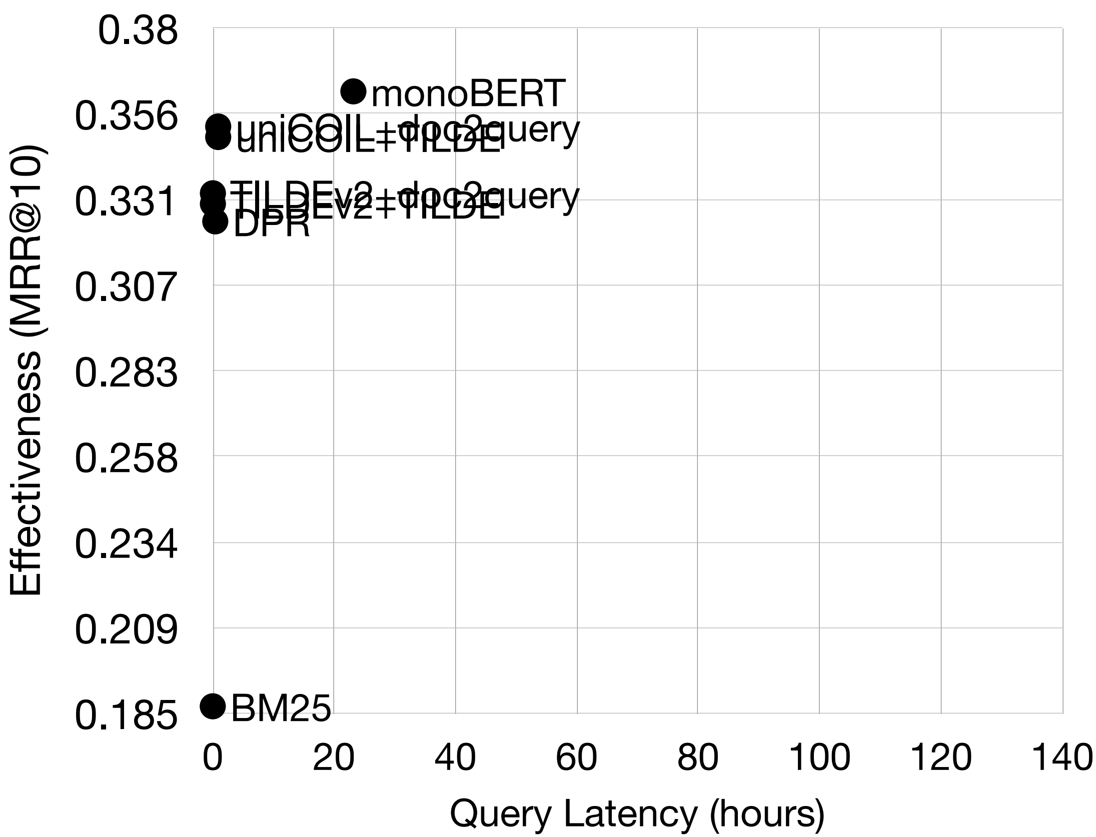
Green IR

How many emissions produced to obtain a single result?



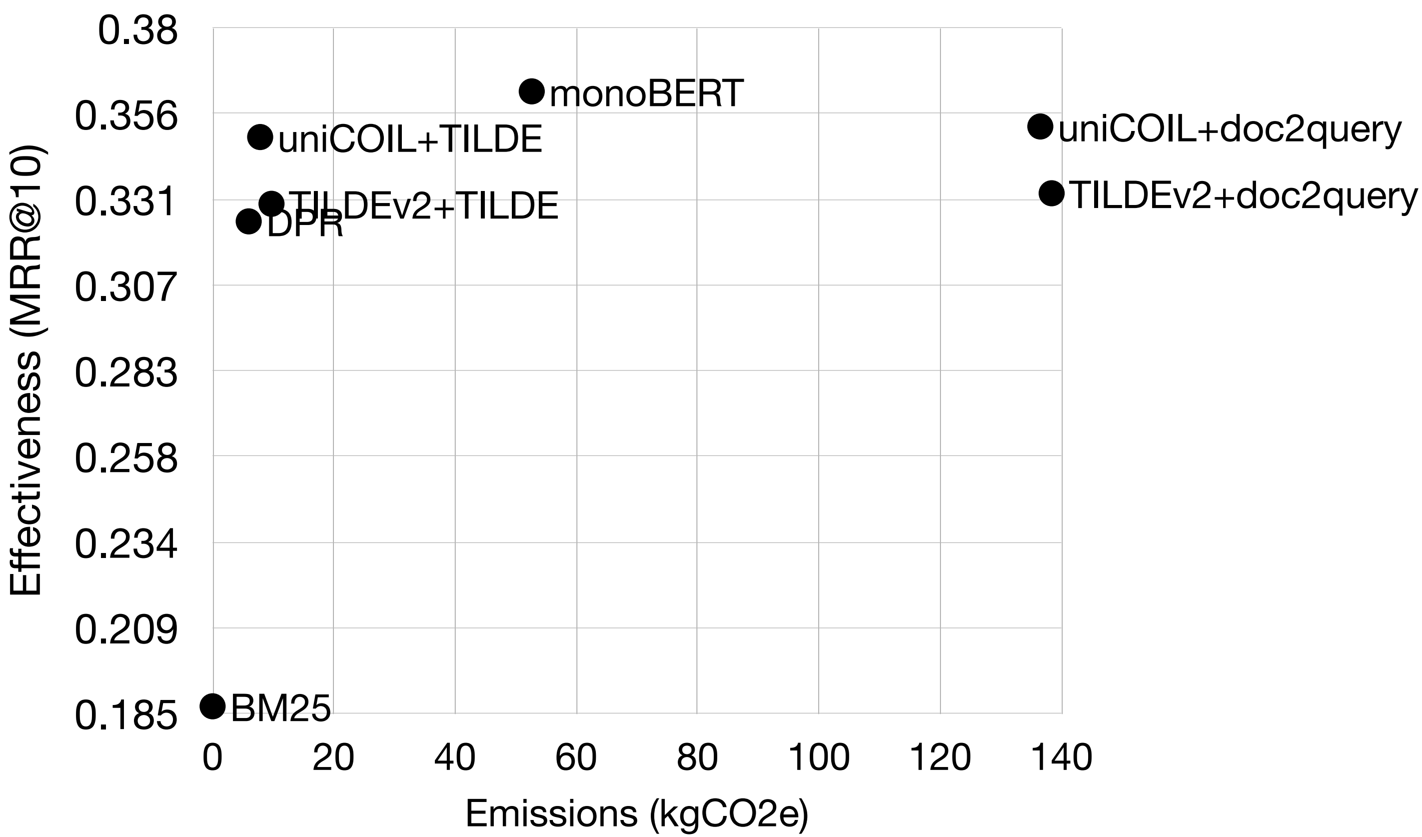
Green IR

What are the effectiveness-utilisation trade-offs of these methods?



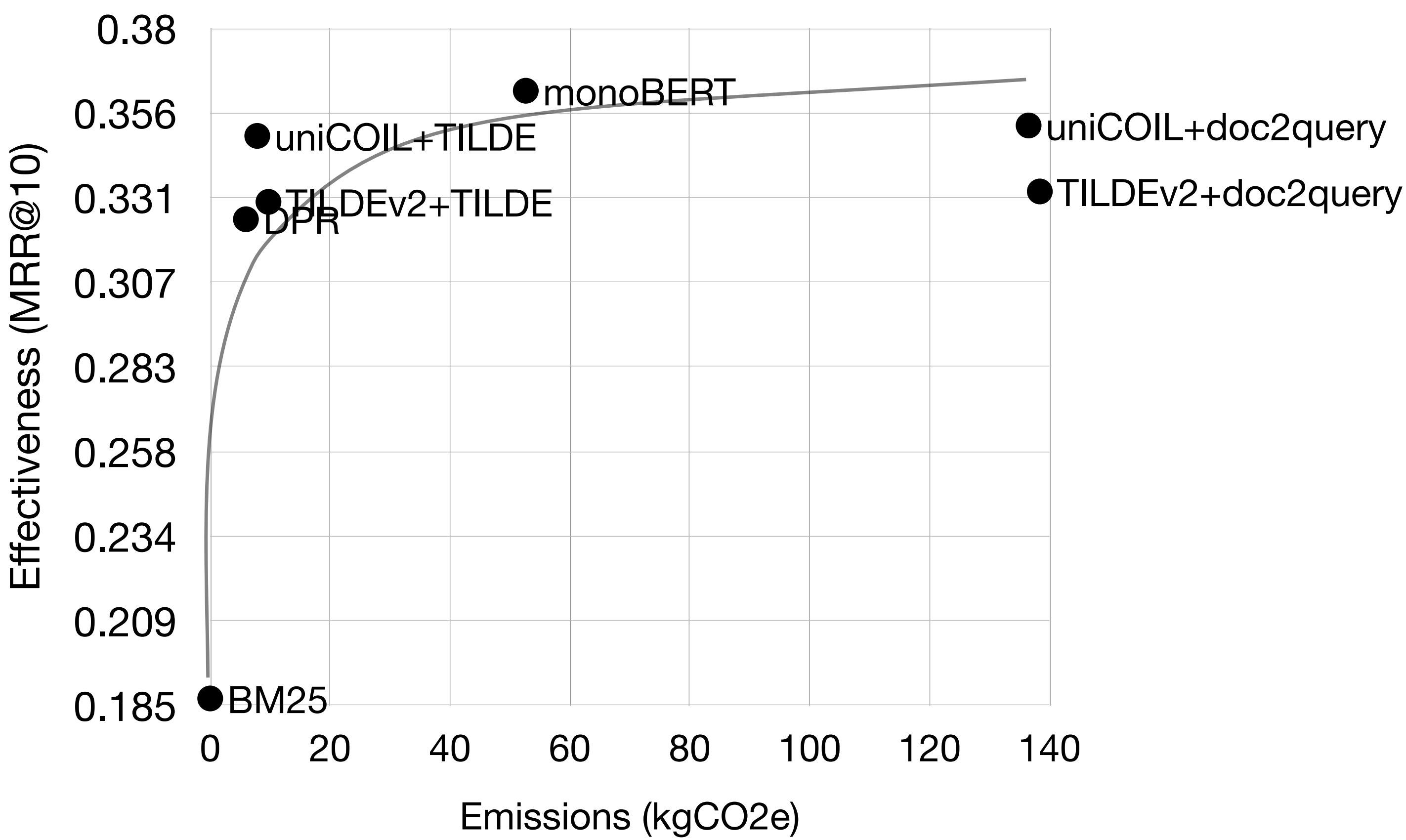
Green IR

What are the effectiveness-utilisation trade-offs of these methods?



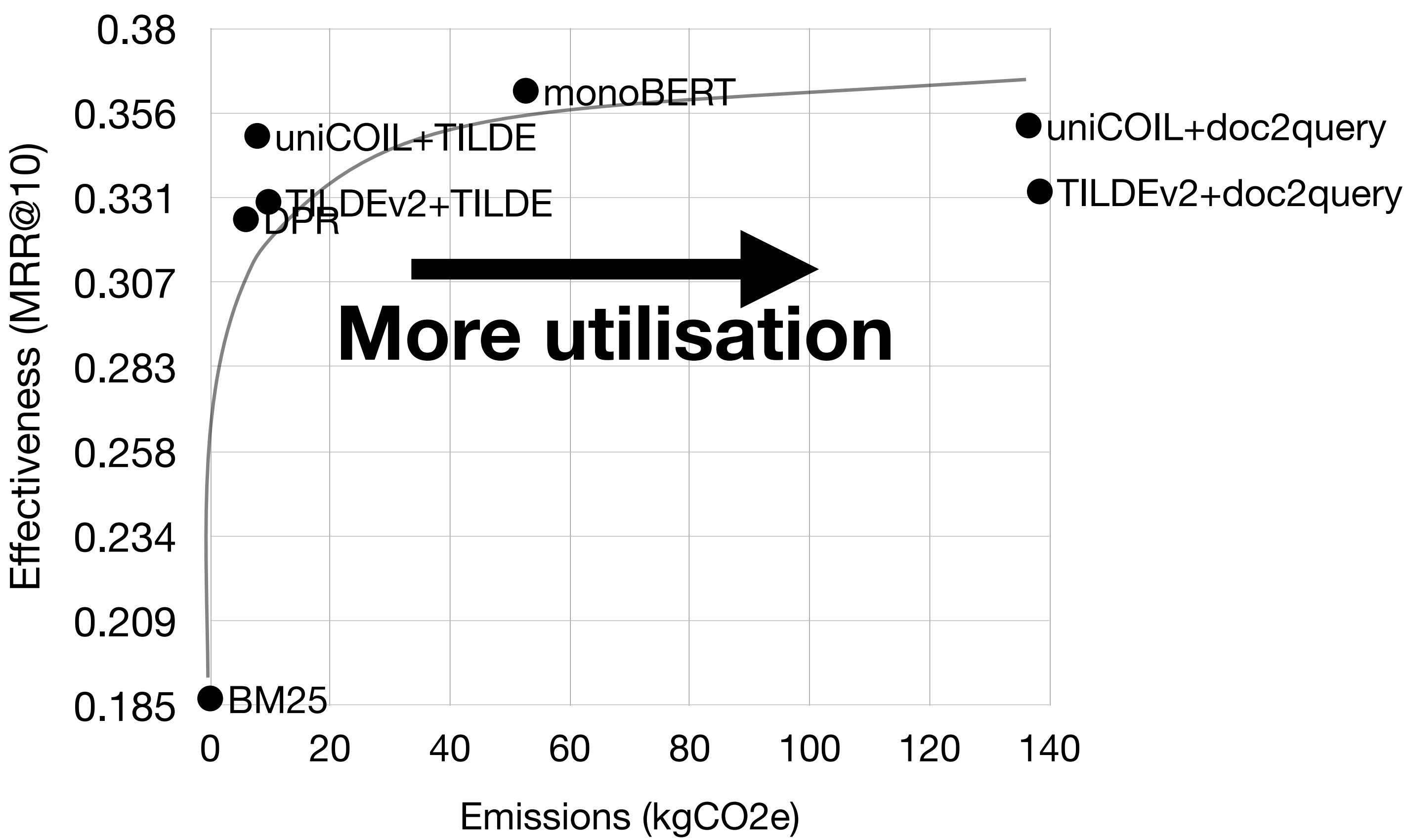
Green IR

What are the effectiveness-utilisation trade-offs of these methods?



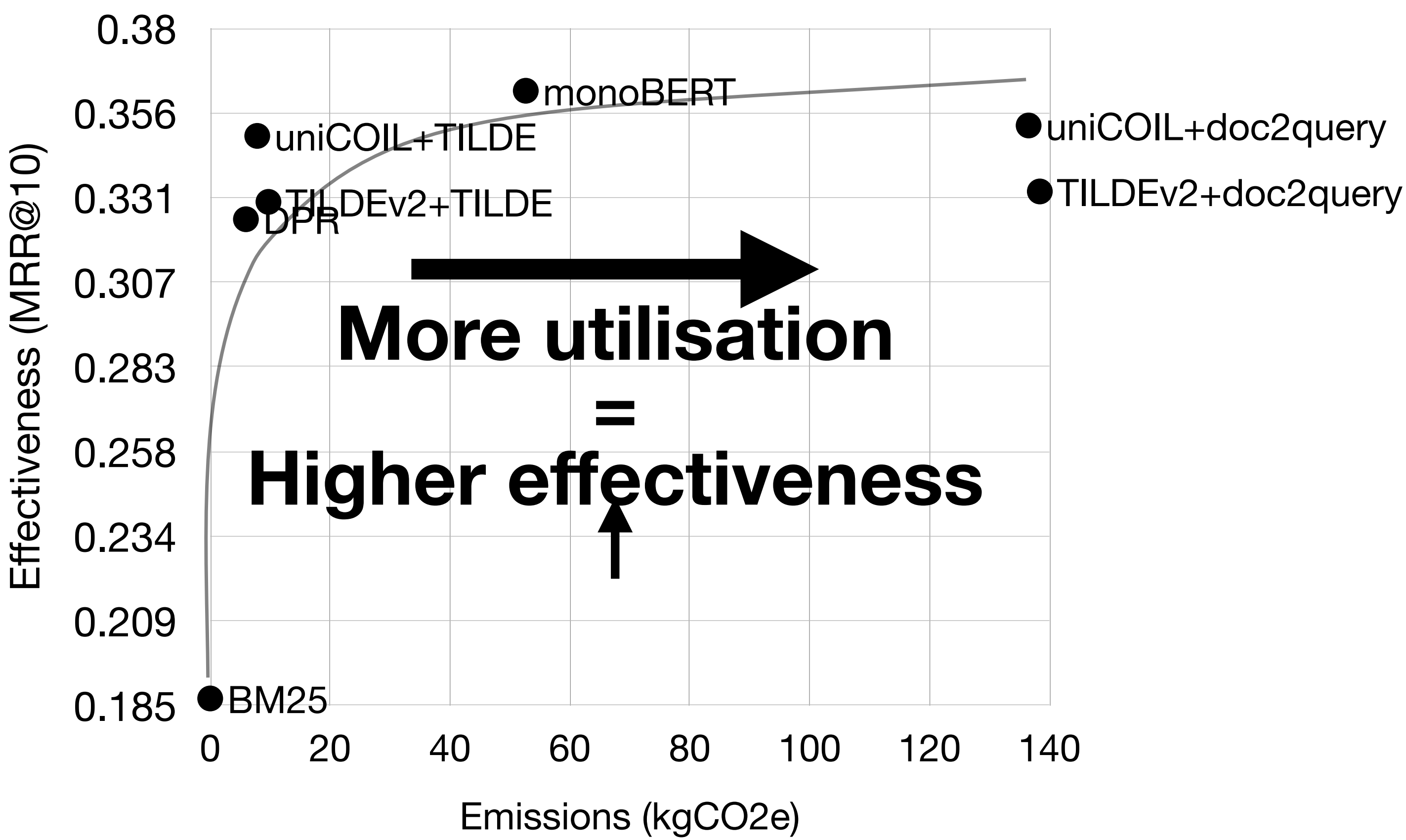
Green IR

What are the effectiveness-utilisation trade-offs of these methods?



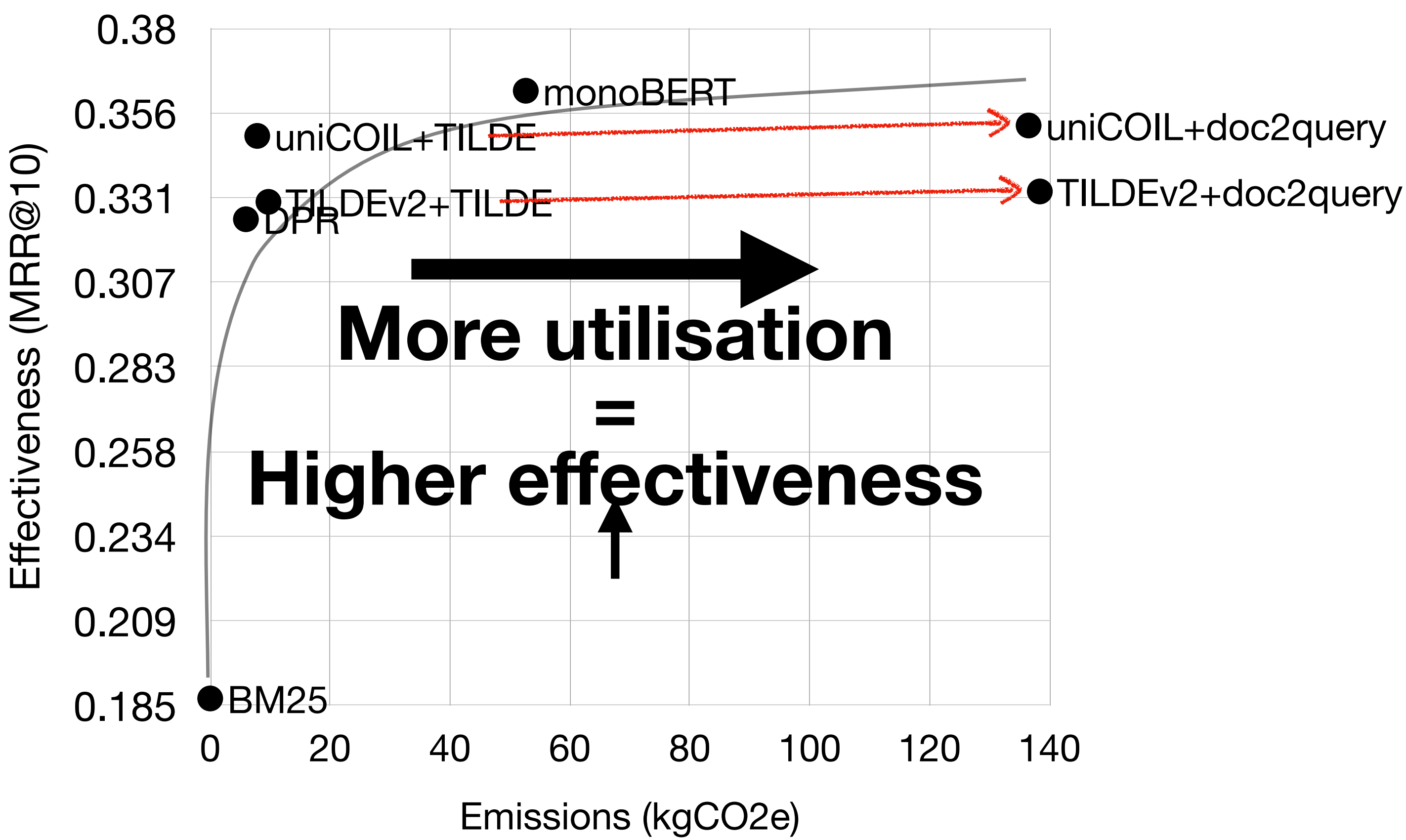
Green IR

What are the effectiveness-utilisation trade-offs of these methods?



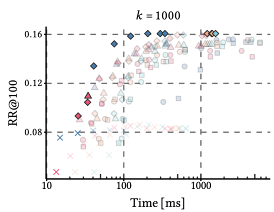
Green IR

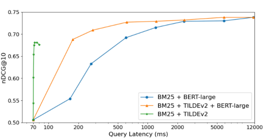
What are the effectiveness-utilisation trade-offs of these methods?

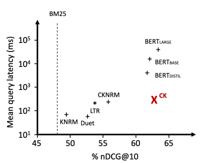


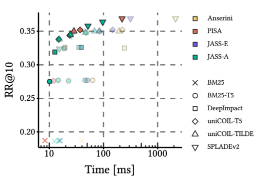
Green IR

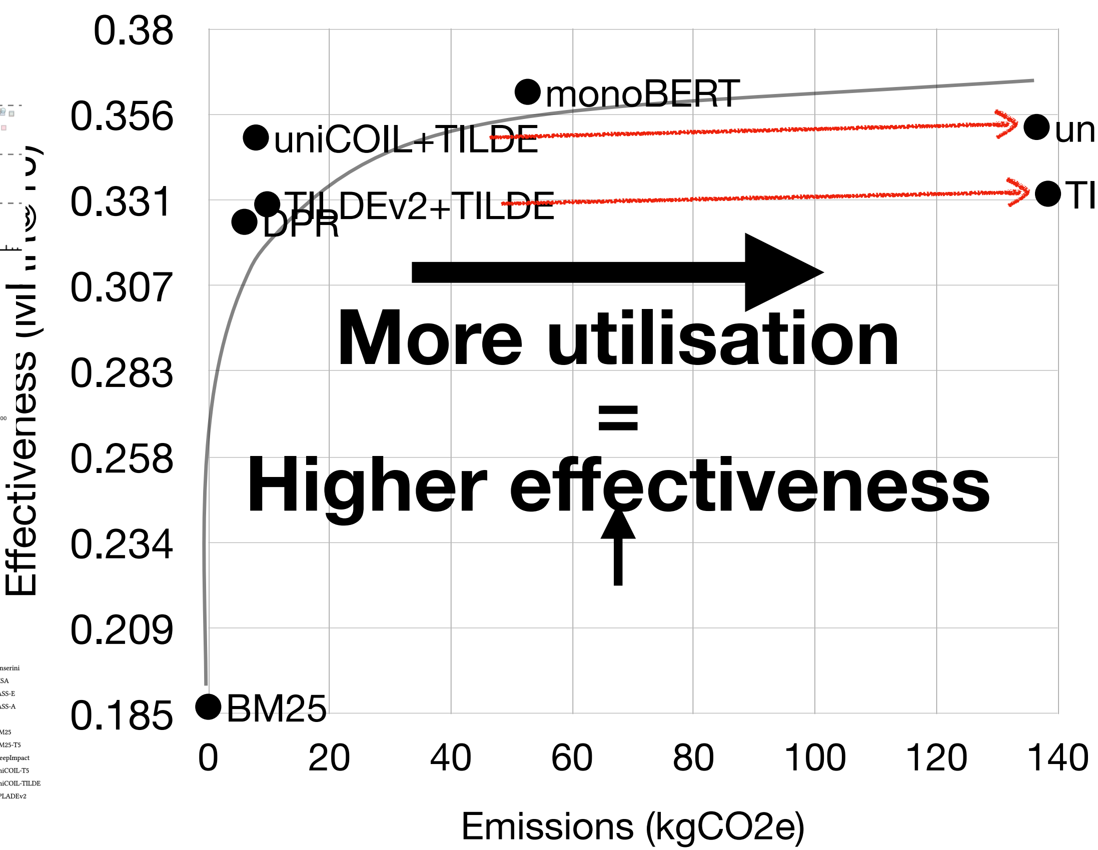
What are the effectiveness-utilisation trade-offs of these methods?

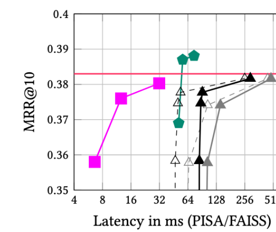


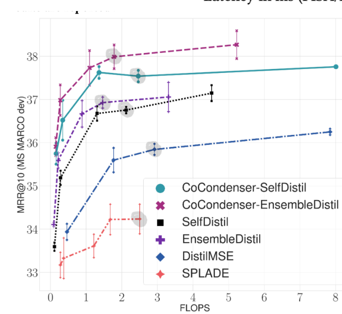












Green IR

Reduce, Reuse, Recycle

Reduce ➜ expend fewer resources

- Straightforward: simply reduce the number of experiments
- Limit expensive computations, e.g., use CPU, FPGAs over GPU
- Prior to starting any research or experiments, ask: How can I perform research with fewer resources?

Green IR

Reduce, Reuse, Recycle

Reduce ➜ expend fewer resources

- ❑ Straightforward: simply reduce the number of experiments
- ❑ Limit expensive computations, e.g., use CPU, FPGAs over GPU
- ❑ Prior to starting any research or experiments, ask: How can I perform research with fewer resources?

Reuse ➜ repurpose resources intended for one task to the same task

- ❑ Reuse existing software artefacts such as data, code, or models
- ❑ Take something existing and repurpose it for the same task it was devised for
- ❑ Prior to starting any research or experiments, ask: How can I repurpose data or code meant for one task to the same task?

Green IR

Reduce, Reuse, Recycle

Reduce ➜ expend fewer resources

- Straightforward: simply reduce the number of experiments
- Limit expensive computations, e.g., use CPU, FPGAs over GPU
- Prior to starting any research or experiments, ask: How can I perform research with fewer resources?

Reuse ➜ repurpose resources intended for one task to the same task

- Reuse existing software artefacts such as data, code, or models
- Take something existing and repurpose it for the same task it was devised for
- Prior to starting any research or experiments, ask: How can I repurpose data or code meant for one task to the same task?

Recycle ➜ repurpose resources intended for one task to a different task

- Recycle existing software artefacts such as data, code, or models
- Repurposing an existing artefact for a task it was not originally intended for
- Prior to starting any research or experiments, ask: How can I repurpose existing data or code meant for one task to a different task?

1 Green IR

[Scells et al. 2022]

2 Efficient Listwise Neural Search

[Schlatt et. al 2024]

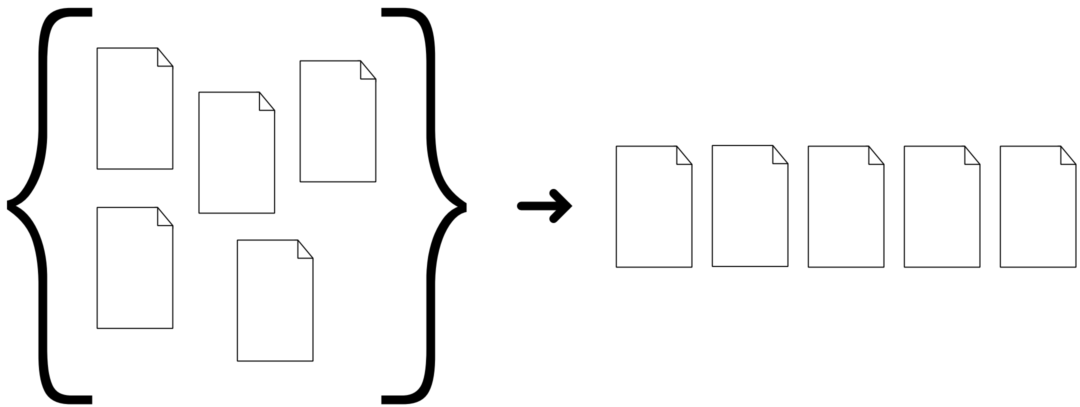
3 Estimating Cost of IR (discussion)

Efficient Listwise Neural Search

Motivation [Schlatt et. al 2024]

Learning task: Given a set of objects, rank them according to a ranking criterion

- Ranking of documents from a set of documents and a query
- Existing transformer architecture cannot model this task effectively
- Two properties: Permutation invariance and cross-document information

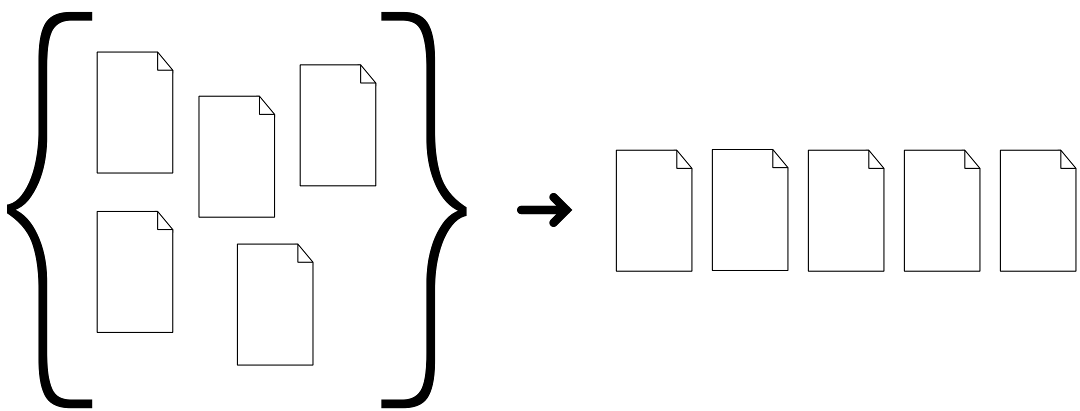


Efficient Listwise Neural Search

Motivation [Schlatt et. al 2024]

Learning task: Given a set of objects, rank them according to a ranking criterion

- ❑ Ranking of documents from a set of documents and a query
- ❑ Existing transformer architecture cannot model this task effectively
- ❑ Two properties: Permutation invariance and cross-document information



Existing architectures model either one of these properties **but never both**

- ❑ Trade off effective ranking for permutation invariance ➜ Pointwise
- ❑ Trade off efficient ranking for cross-document information ➜ Listwise

Efficient Listwise Neural Search

Model Architecture

Pointwise

- More efficient at the expense of effectiveness.
- Permutation-invariant, no cross-document information.
- Scaleable: each query-document pair is scored.

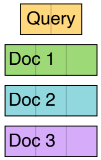
Listwise

- More effective at the expense of efficiency.
- Non-permutation-invariant, cross-document information.
- Unscaleable: All permutations of query-documents is scored.

Efficient Listwise Neural Search

Model Architecture

State of the Art

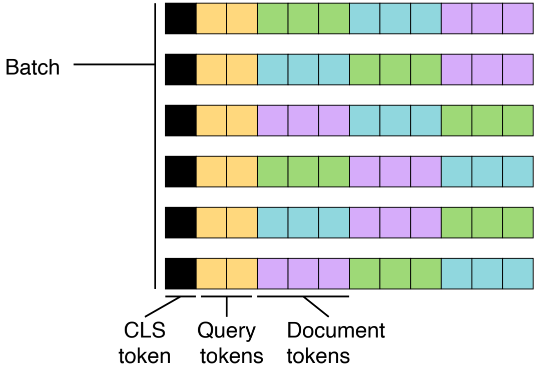


Pointwise

- More efficient at the expense of effectiveness.
- Permutation-invariant, no cross-document information.
- Scaleable: each query-document pair is scored.

Listwise

- More effective at the expense of efficiency.
- Non-permutation-invariant, cross-document information.
- Unscaleable: All permutations of query-documents is scored.



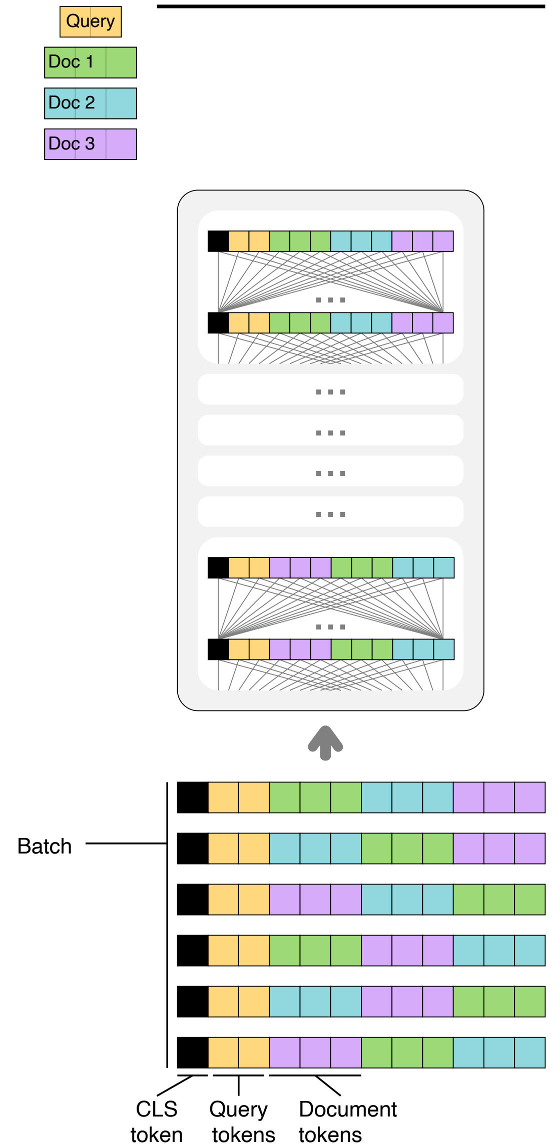
Efficient Listwise Neural Search

Model Architecture

Document scoring:

- Each permutation of documents is fed into model.
- Reason: Transformer is sequence modeller; order of documents biases the score.

State of the Art

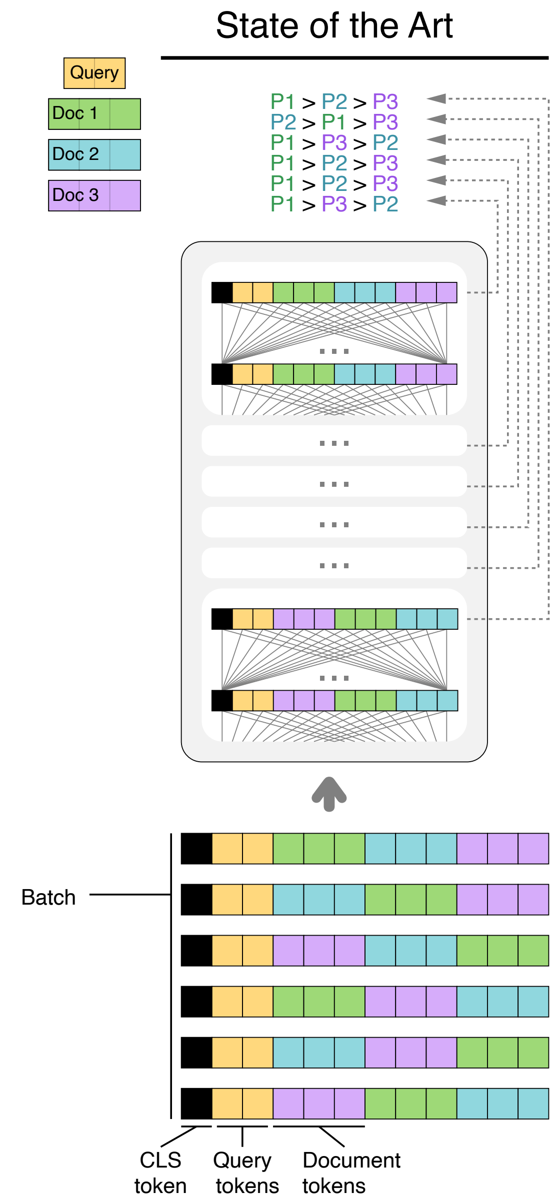


Efficient Listwise Neural Search

Model Architecture

Document scoring:

- Each permutation of documents is fed into model.
- Reason: Transformer is sequence modeller; order of documents biases the score.
- Task: Predict ordering preference of documents given query.

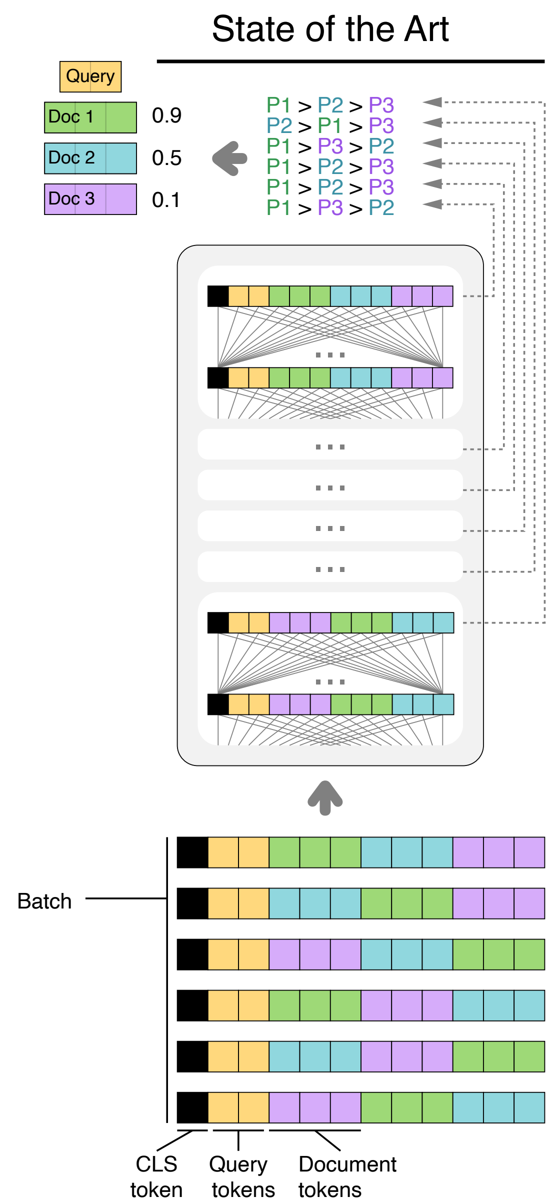


Efficient Listwise Neural Search

Model Architecture

Document scoring:

- ❑ Each permutation of documents is fed into model.
- ❑ Reason: Transformer is sequence modeller; order of documents biases the score.
- ❑ Task: Predict ordering preference of documents given query.
- ❑ Score computed by aggregating preferences.

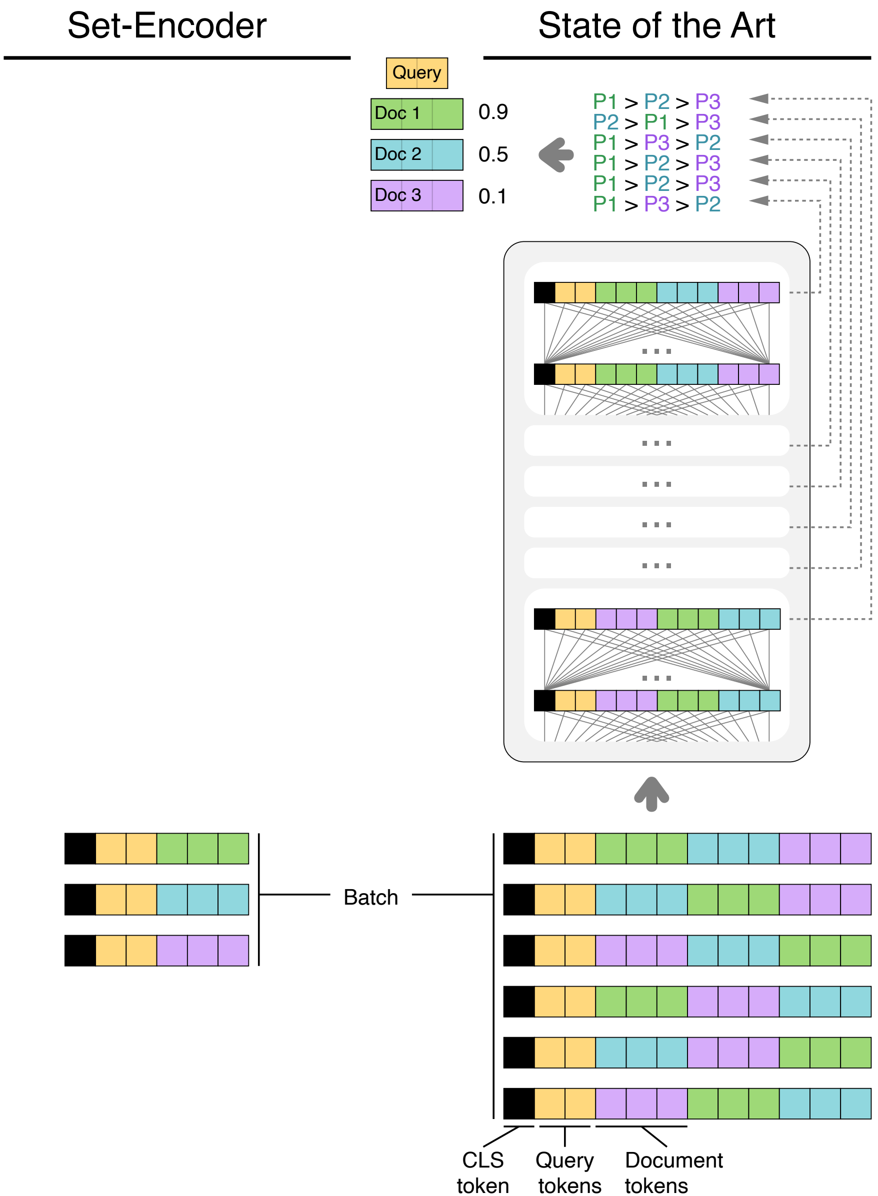


Efficient Listwise Neural Search

Model Architecture

Set-Encoder document scoring:

- ❑ Each query-document pair only needs to be scored once.

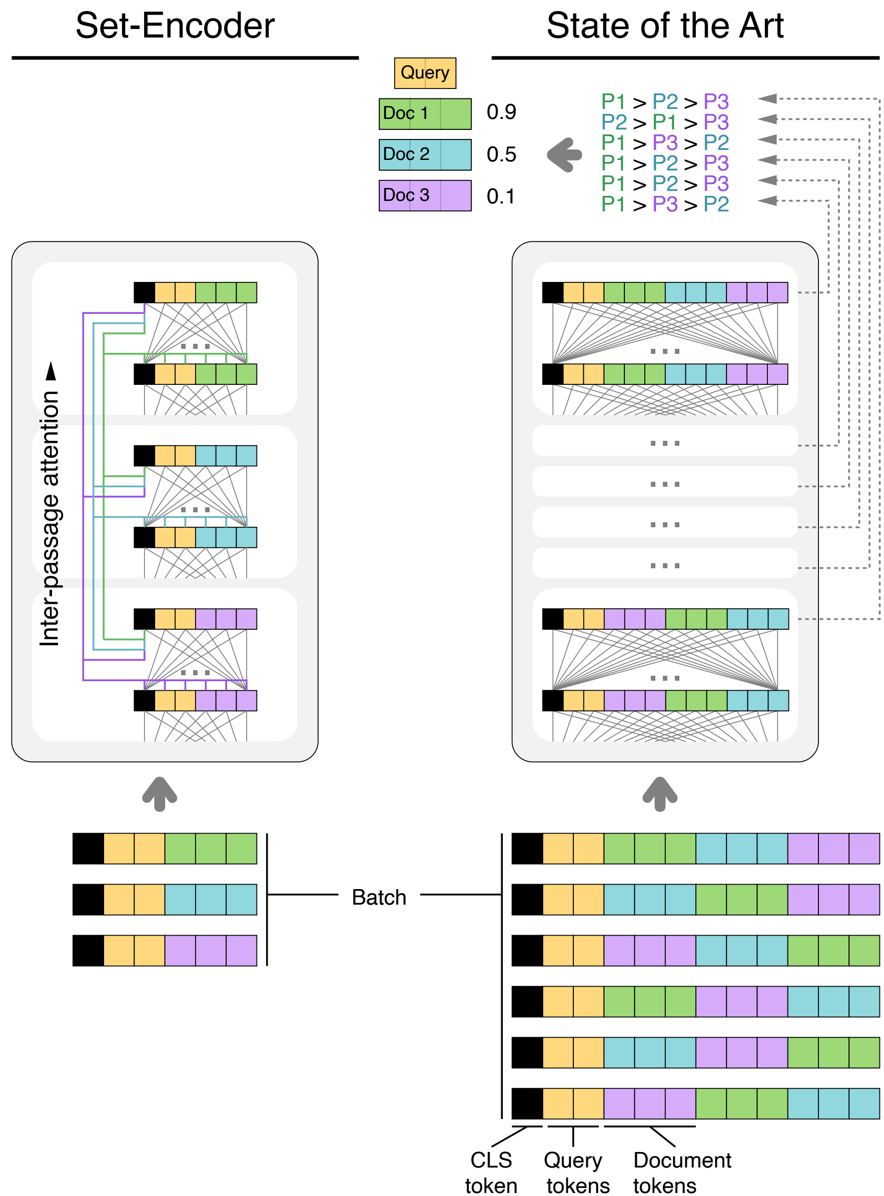


Efficient Listwise Neural Search

Model Architecture

Set-Encoder document scoring:

- Each query-document pair only needs to be scored once.
- Share cross-document information through attention mechanism.
- Reset positional information to make scores permutation invariant.

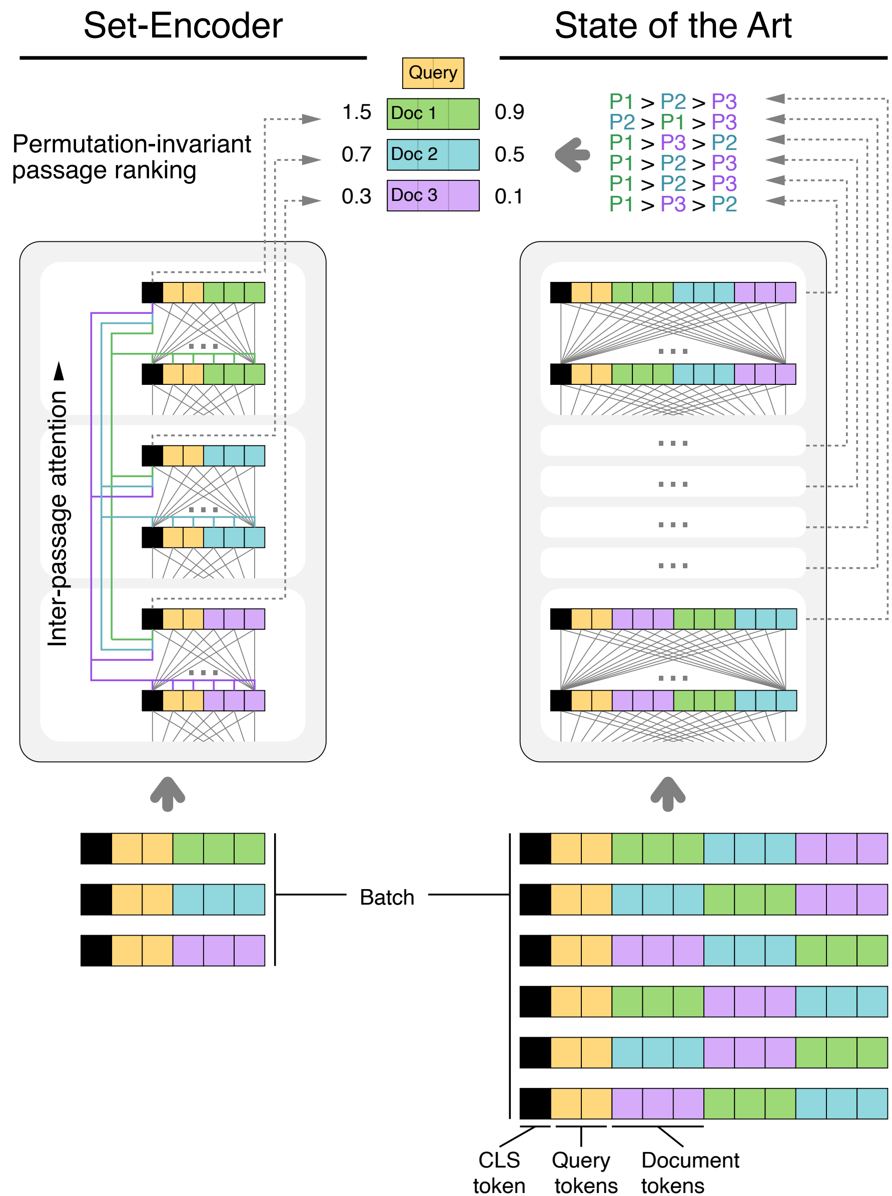


Efficient Listwise Neural Search

Model Architecture

Set-Encoder document scoring:

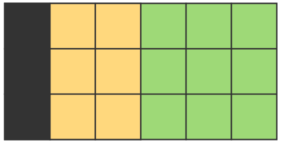
- Each query-document pair only needs to be scored once.
- Share cross-document information through attention mechanism.
- Reset positional information to make scores permutation invariant.
- Score computed directly for all query-document pairs.



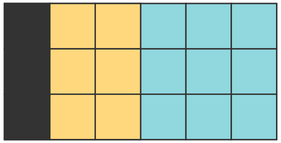
Efficient Listwise Neural Search

Modelling Cross-Document Interactions with Attention

q, d_1



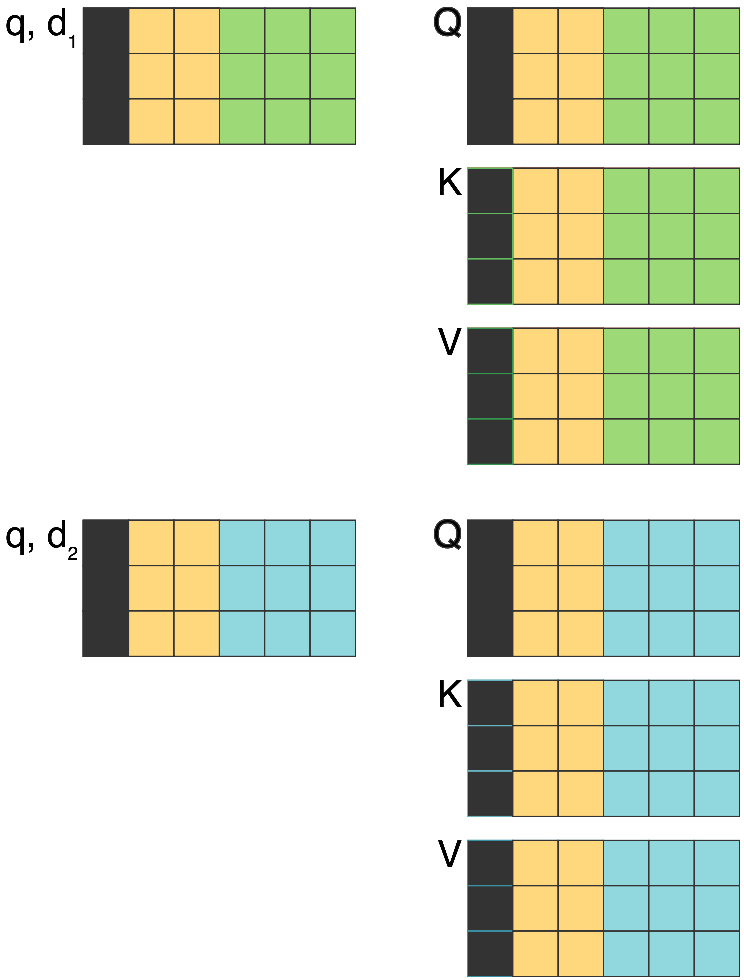
q, d_2



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{h}}\right)V$$

Efficient Listwise Neural Search

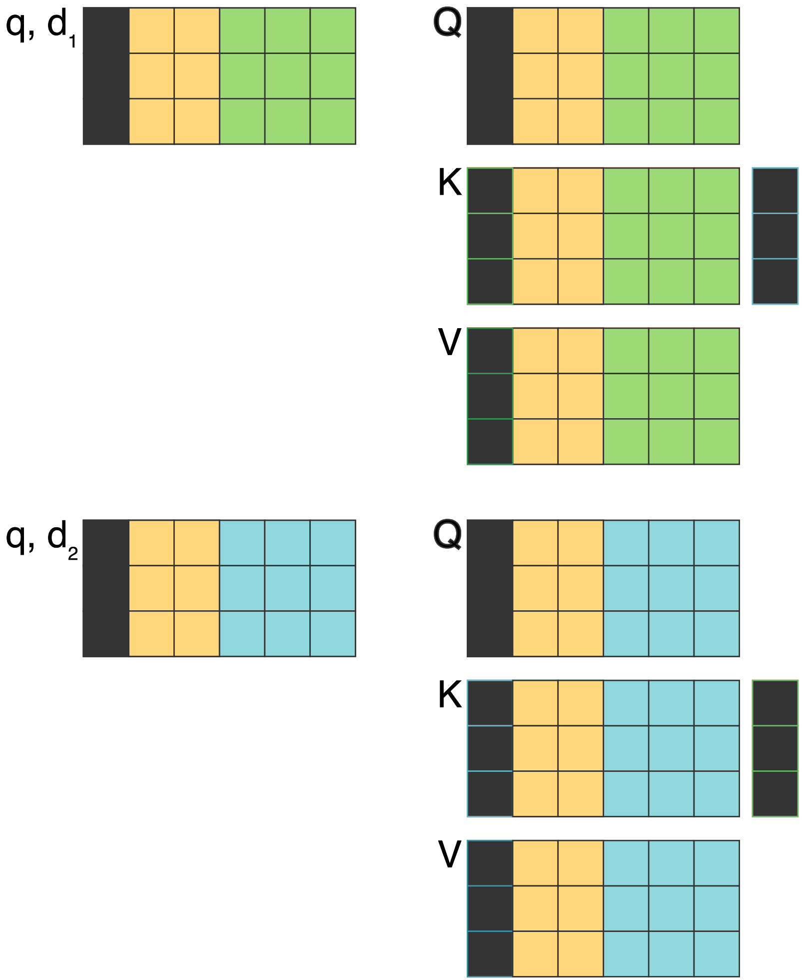
Modelling Cross-Document Interactions with Attention



$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{h}}\right)V$$

Efficient Listwise Neural Search

Modelling Cross-Document Interactions with Attention

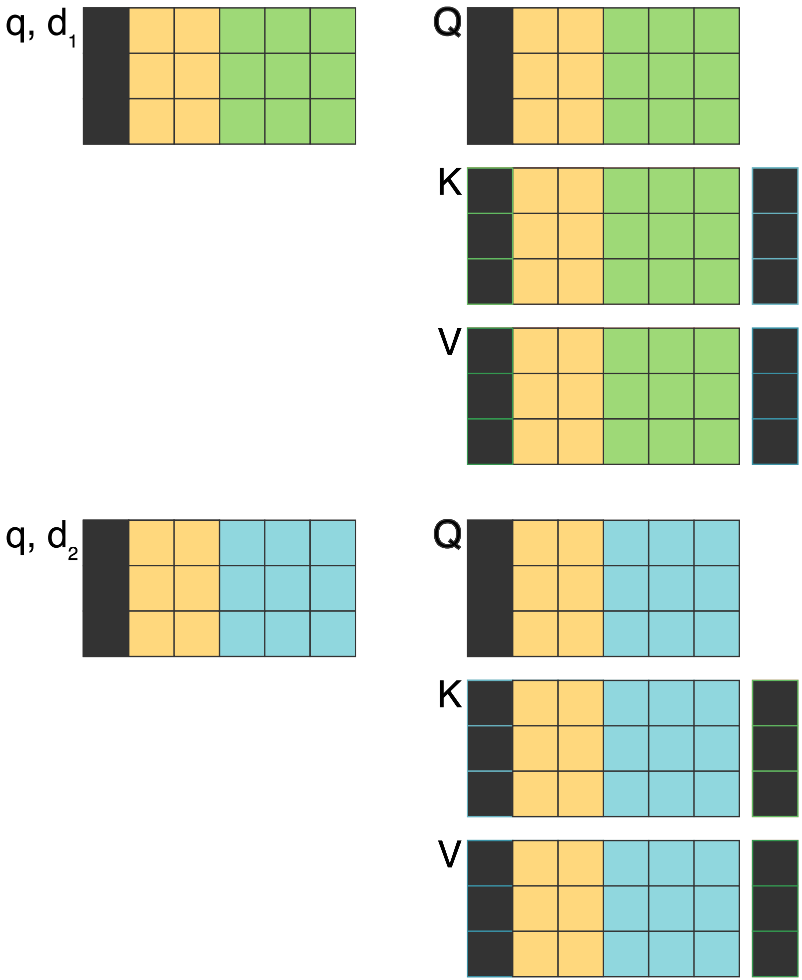


$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{h}}\right)V$$

For d$_i$, let $\bar{K}^i = [K_1^j : j \neq i]$

Efficient Listwise Neural Search

Modelling Cross-Document Interactions with Attention



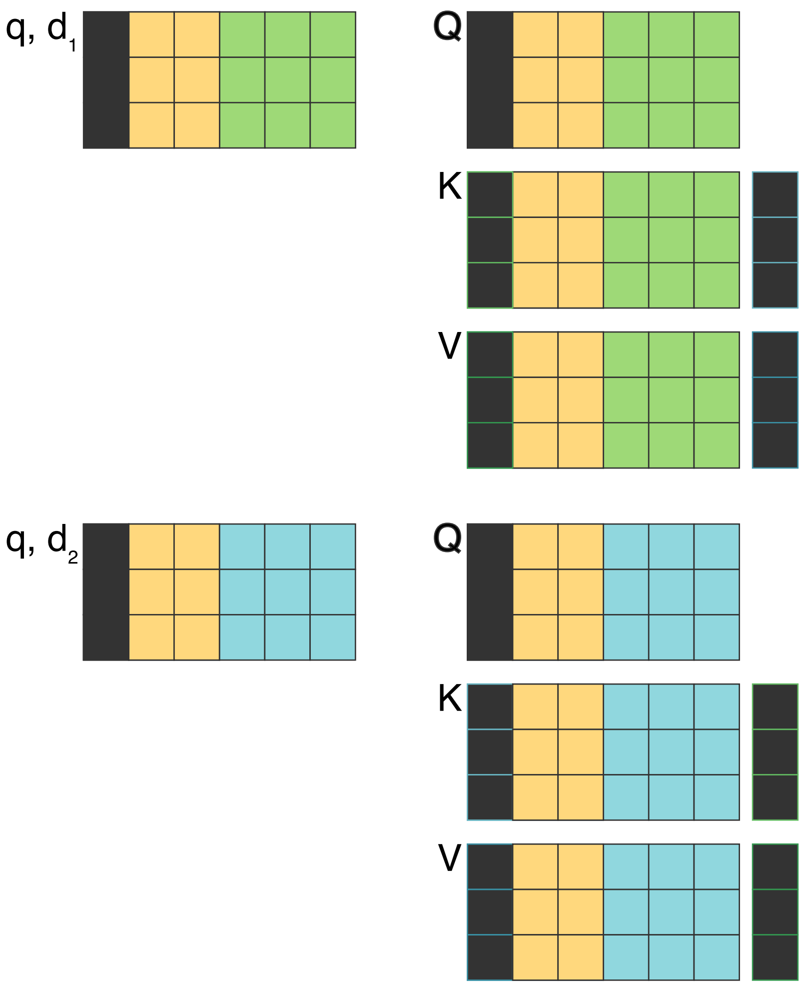
$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}(\frac{\text{QK}^T}{\sqrt{h}})V$$

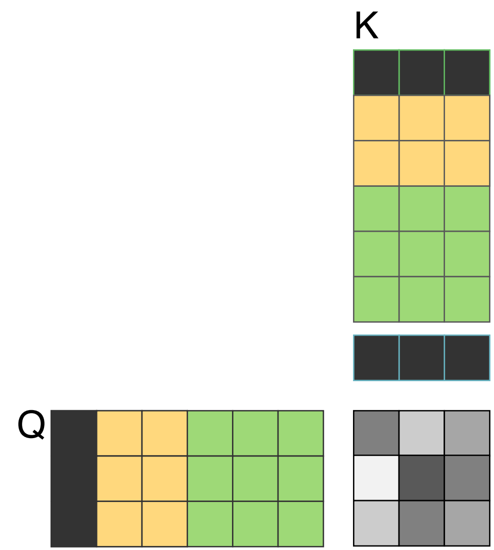
For d$_i$, let $\bar{K}^i = [K_1^j : j \neq i]$

For d$_i$, let $\bar{V}^i = [V_1^j : j \neq i]$

Efficient Listwise Neural Search

Modelling Cross-Document Interactions with Attention





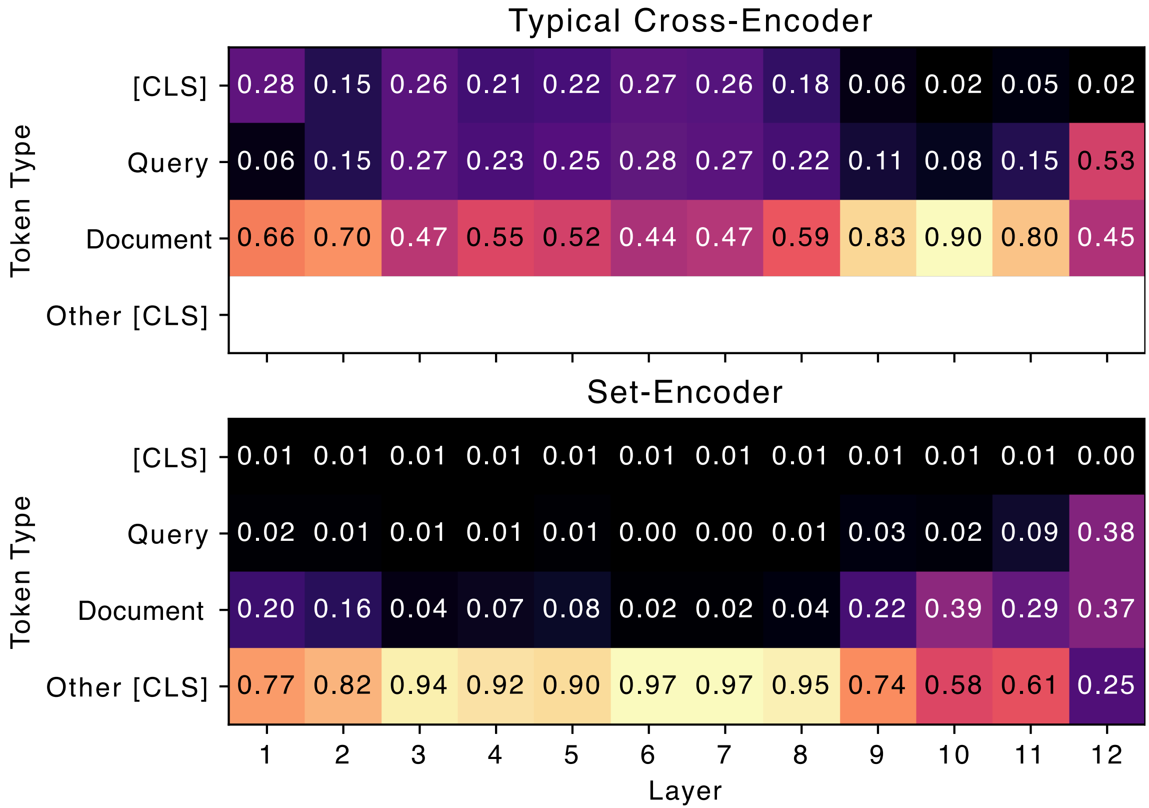
$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^T}{\sqrt{h}}\right)V$$

For d$_i$, let $\bar{K}^i = [K_1^j : j \neq i]$
For d$_i$, let $\bar{V}^i = [V_1^j : j \neq i]$

Cross-document attention for d$_i$:
$\text{Attention}(\text{Q}^i, [\text{K}^i\bar{\text{K}}^i], [\text{V}^i\bar{\text{V}}^i])$

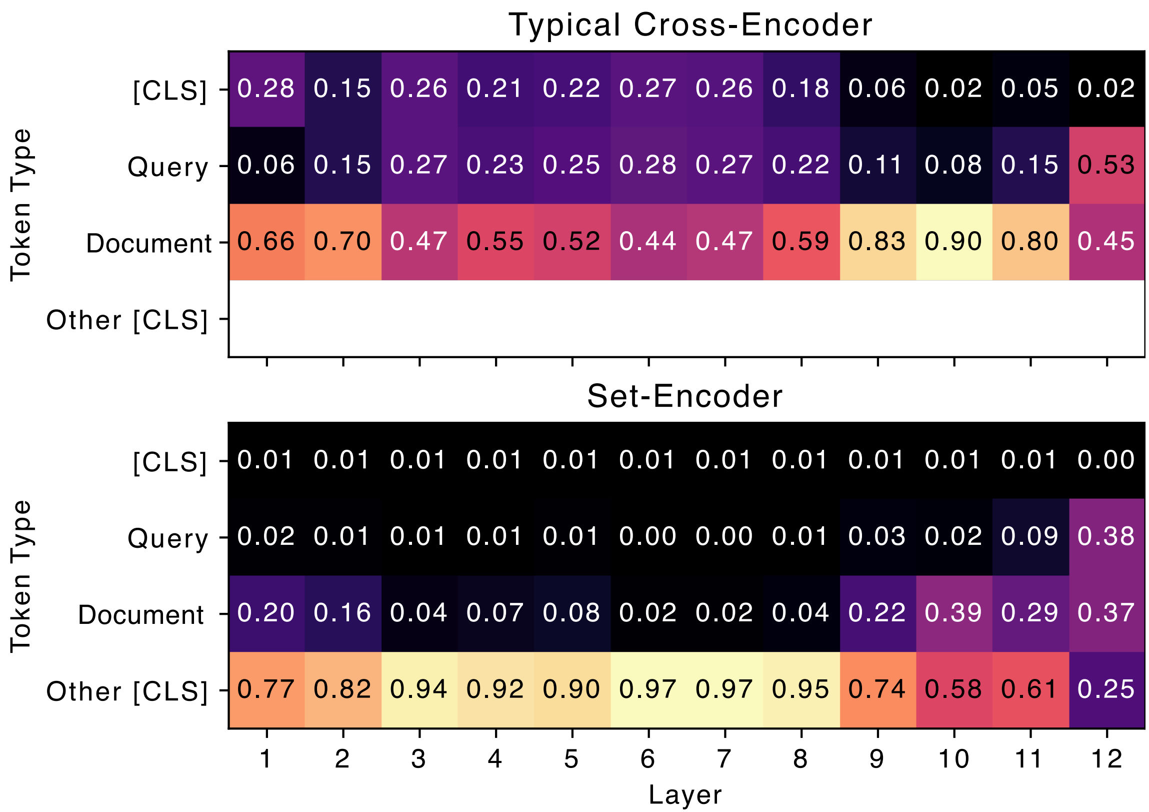
Efficient Listwise Neural Search

Attention Visualised



Efficient Listwise Neural Search

Attention Visualised



Set-Encoder attends to other documents in early layers, then the document to score in final layers.

Efficient Listwise Neural Search

Results: Ranking Effectiveness

Model	Parameters	Effectiveness (nDCG@10)
monoBERT base	110M	0.379
monoBERT large	340M	0.381
monoT5 base	220M	0.376
monoT5 large	3B	0.410
LiT5-Distill	220M	0.406
Set-Encoder	110M	0.406

Efficient Listwise Neural Search

Results: Ranking Effectiveness

Model	Parameters	Effectiveness (nDCG@10)
monoBERT base	110M	0.379
monoBERT large	340M	0.381
monoT5 base	220M	0.376
monoT5 large	3B	0.410
LiT5-Distill	220M	0.406
Set-Encoder	110M	0.406

Set-Encoder has same effectiveness of SOTA listwise model with half the parameters.

Efficient Listwise Neural Search

Results: Ranking Effectiveness

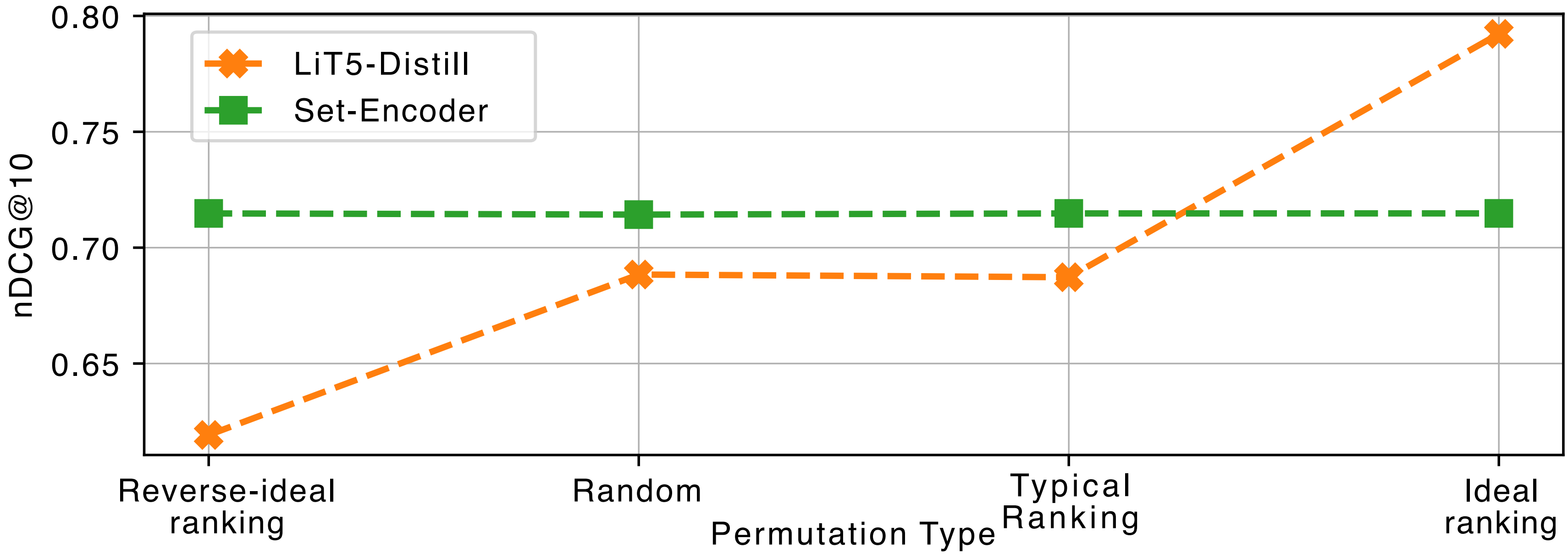
Model	Parameters	Effectiveness (nDCG@10)
monoBERT base	110M	0.379
monoBERT large	340M	0.381
monoT5 base	220M	0.376
monoT5 large	3B	0.410
LiT5-Distill	220M	0.406
Set-Encoder	110M	0.406

Set-Encoder has same effectiveness of SOTA listwise model with half the parameters.

Set-Encoder has similar effectiveness to SOTA pointwise model with 3B fewer parameters.

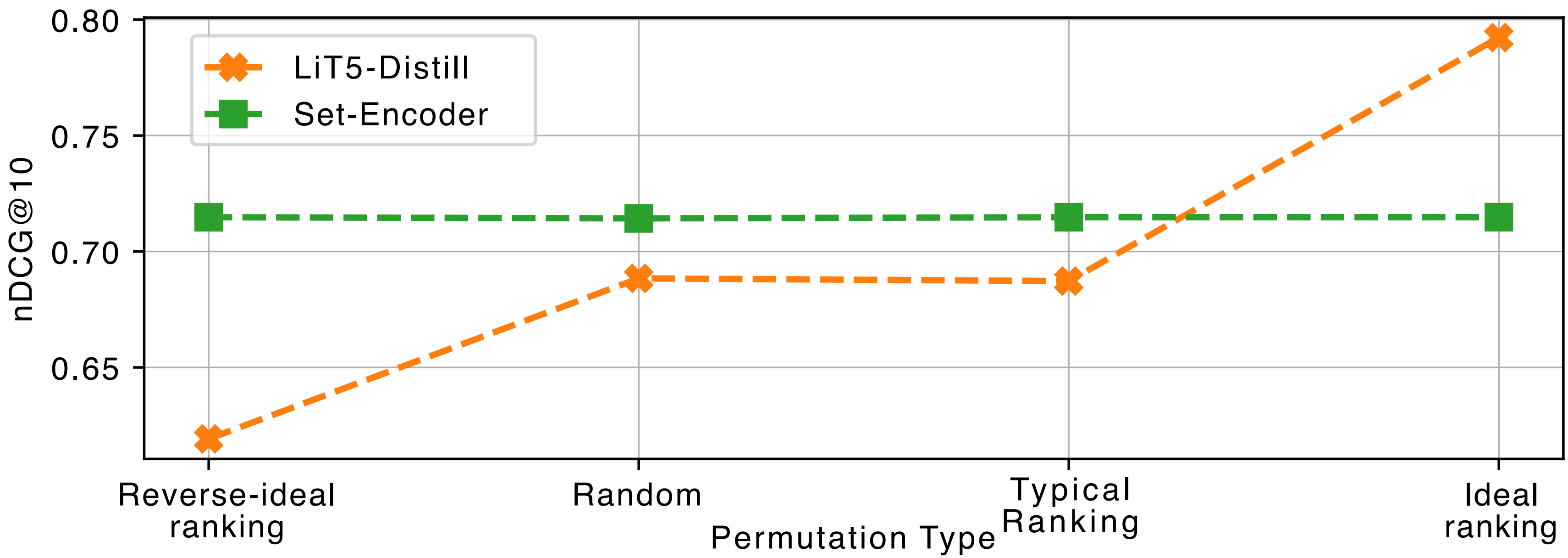
Efficient Listwise Neural Search

Robustness to Initial Ranking Permutations



Efficient Listwise Neural Search

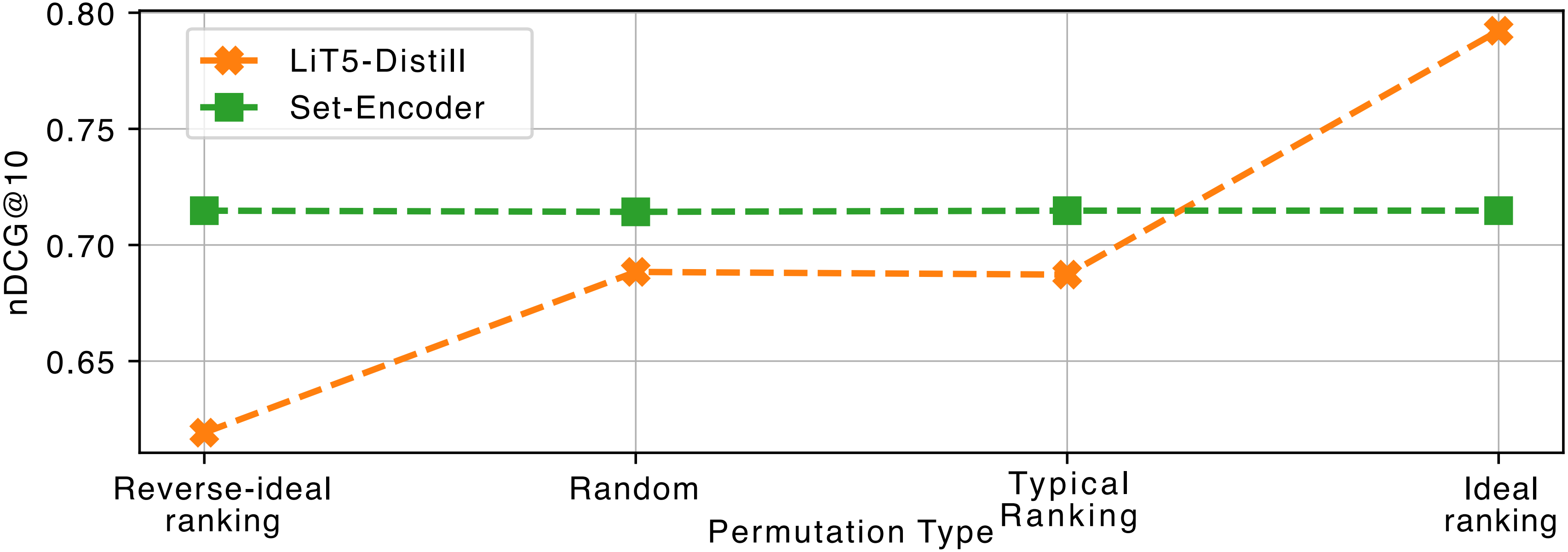
Robustness to Initial Ranking Permutations



Irrespective of initial document ranking, Set-Encoder has same effectiveness.

Efficient Listwise Neural Search

Robustness to Initial Ranking Permutations



Irrespective of initial document ranking, Set-Encoder has same effectiveness.

SOTA Listwise model makes document ranking worse when given ideal ranking.

1 Green IR

[Scells et al. 2022]

2 Efficient Listwise Neural Search

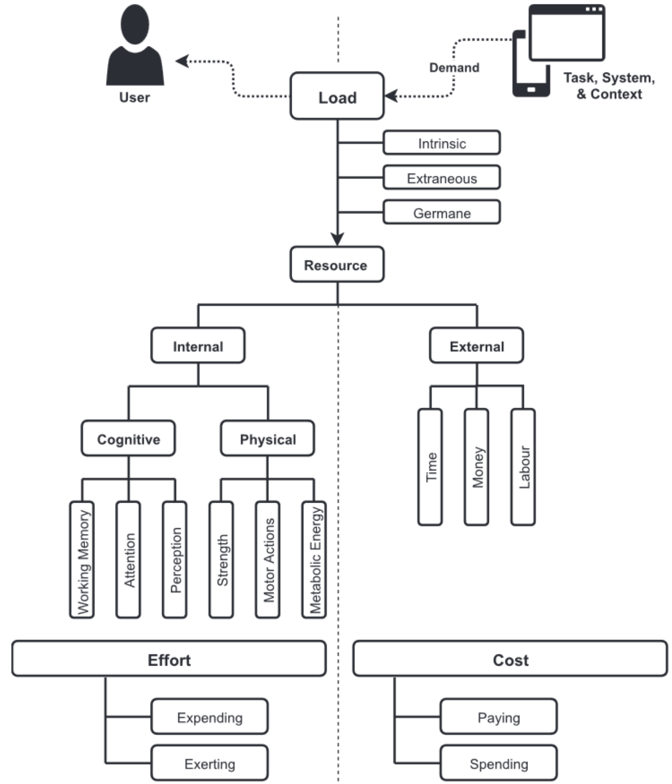
[Schlatt et. al 2024]

3 Estimating Cost of IR (discussion)

Estimating Cost of IR

Starting the Discussion

What do we mean by cost?



cost ➜ system (time, money, energy)

- ❑ training efficiency?
- ❑ inference efficiency?
- ❑ energy utilisation?

cost ➜ user (cost, effort, load) [McGregor et al. 2023]

- ❑ cognitive costs, fatigue, spend or conserve my resources to achieve goal?
- ❑ cognitive or physical effort, task complexity, total labour/time to achieve goal
- ❑ cognitive load, demands, properties of task that regulate exertion, overload