Green Information Retrieval Research

Harry Scells
Leipzig University, Germany

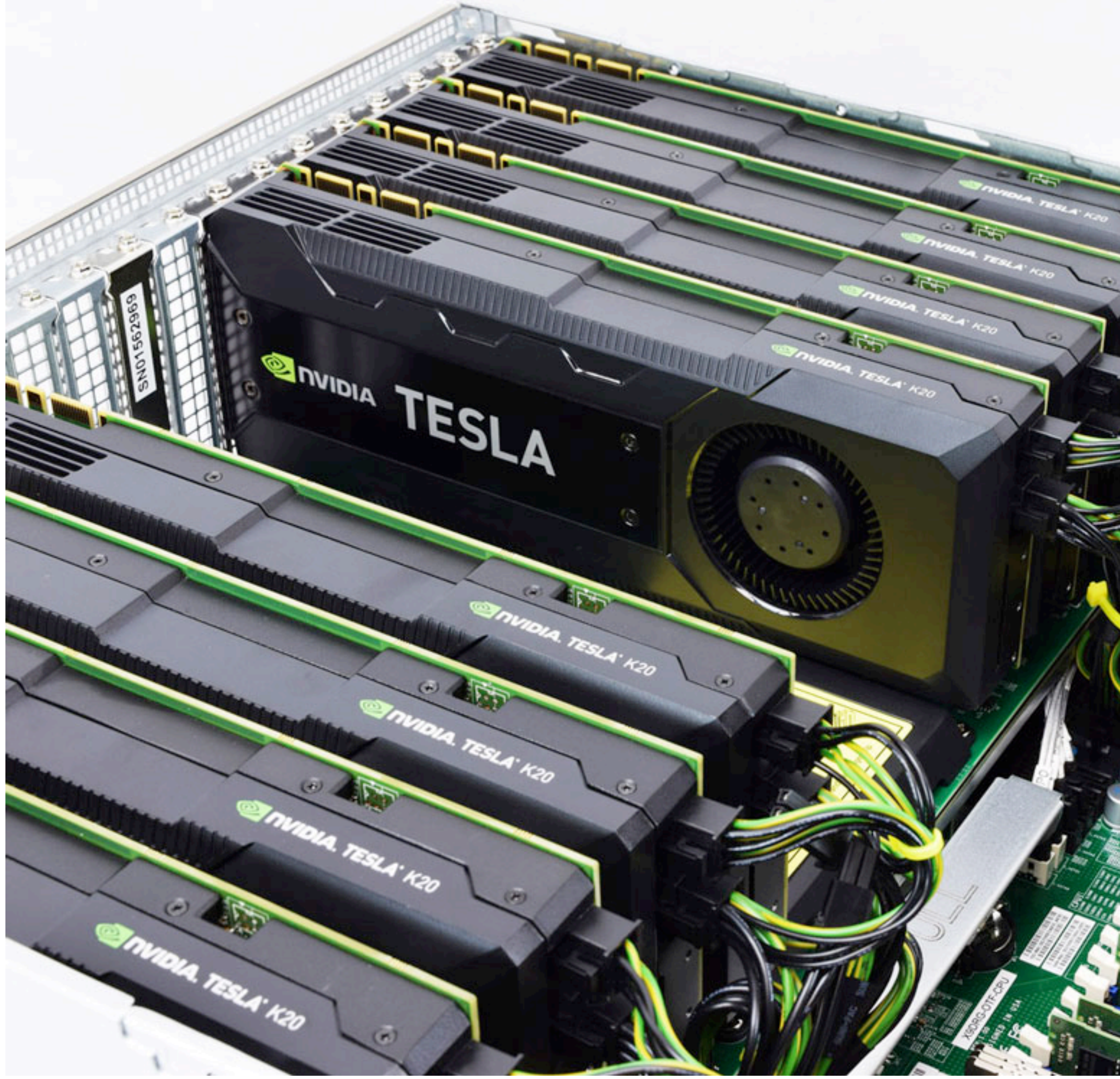
PART I

Context

[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*

Why?

- Large (pre-trained) neural language models
  - Expend high energy for training and inference (comared 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



[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*

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 for 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 for 1 hour

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

NLP

ML

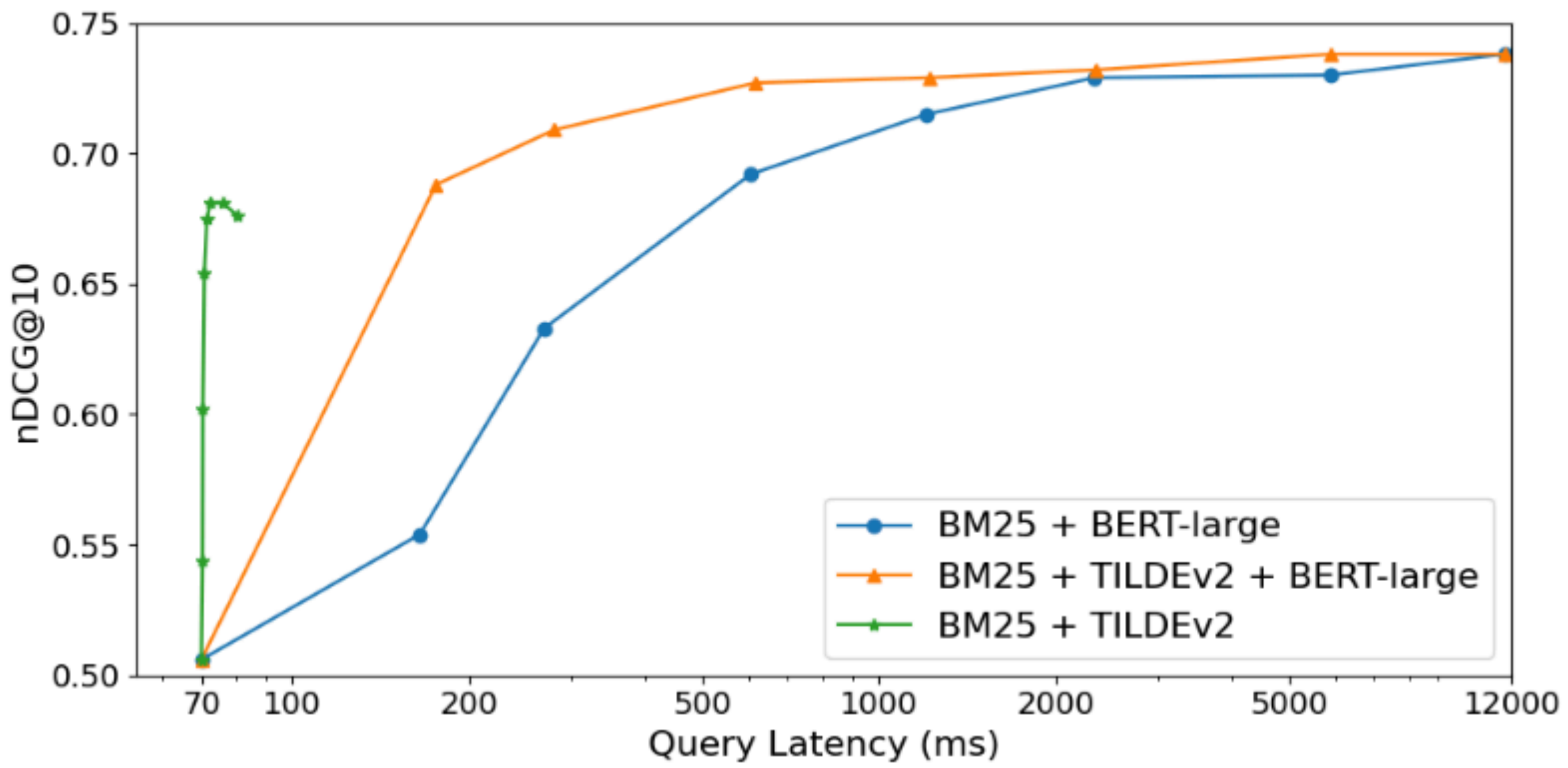
What about IR research?

Isn't this just retrieval efficiency?

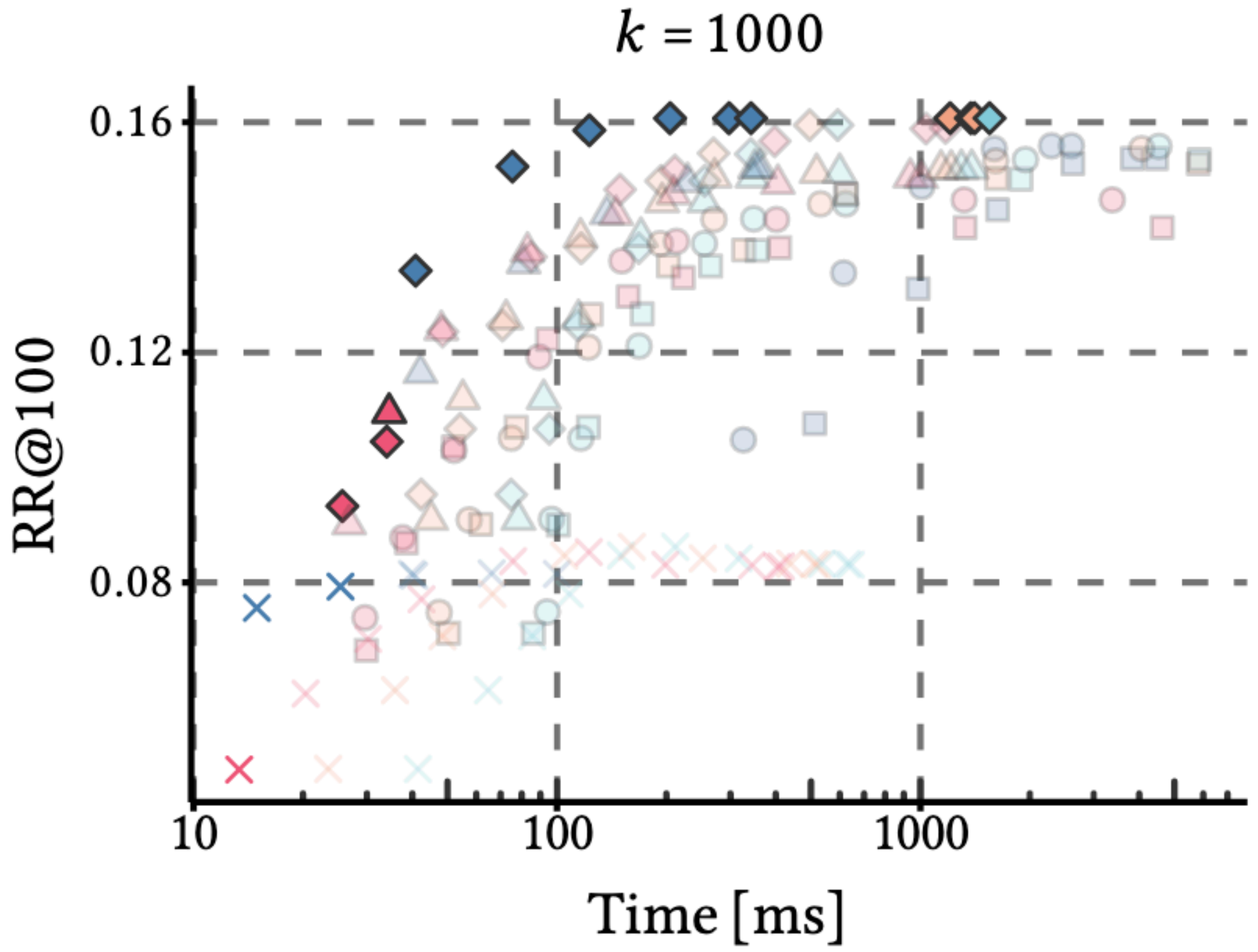
[1] Strubell, E. et al. 2019. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*

Retrieval Efficiency

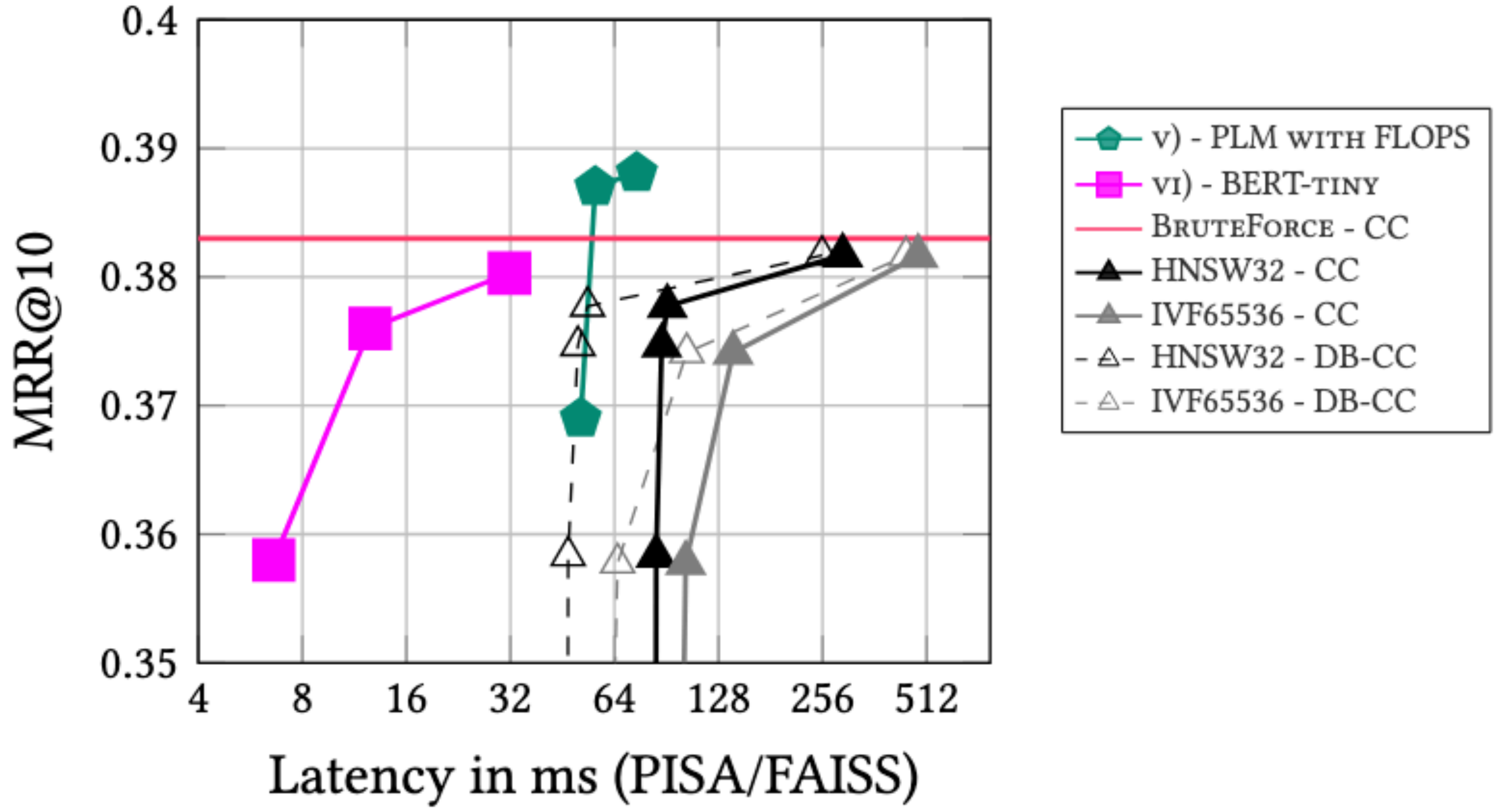
- **Speed** a system is able to retrieve relevant documents or information in response to a query.
- Factors that can impact retrieval efficiency include:
  - **Size and complexity of the corpus** being searched
  - Effectiveness of the **retrieval models** or techniques being used
  - Efficiency of the **hardware and infrastructure** used



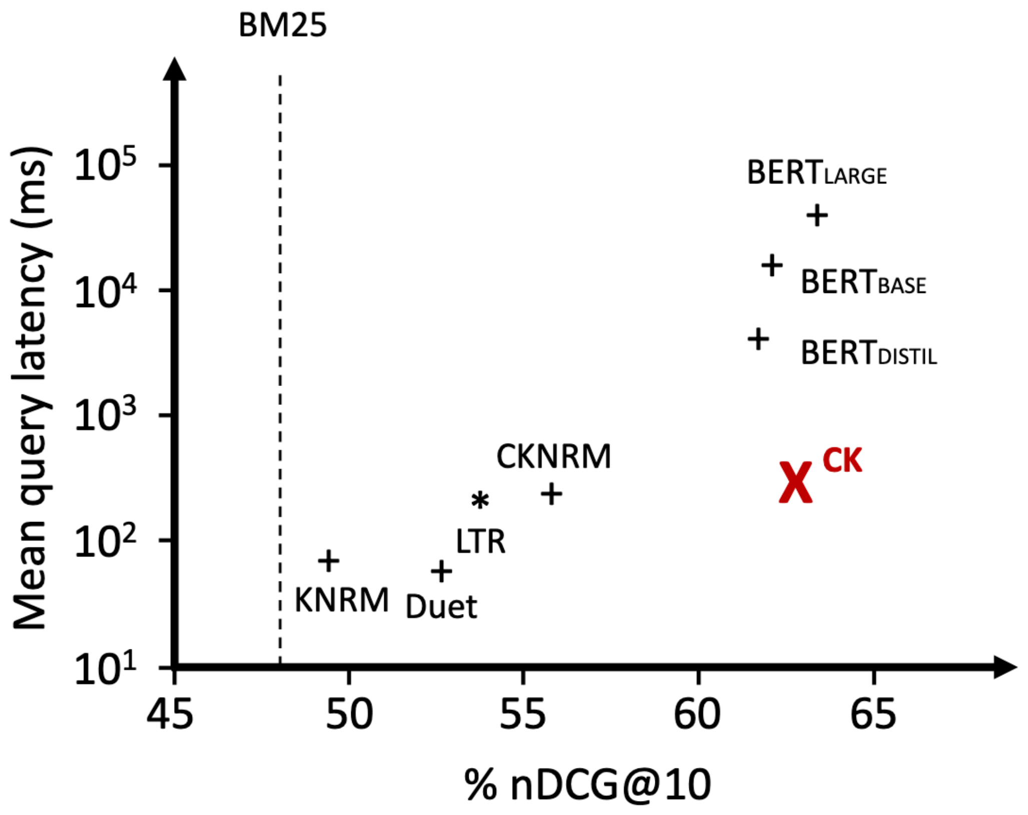
nDCG@10
Query Latency (ms)
BM25 + BERT-large
BM25 + TILDEv2 + BERT-large
BM25 + TILDEv2



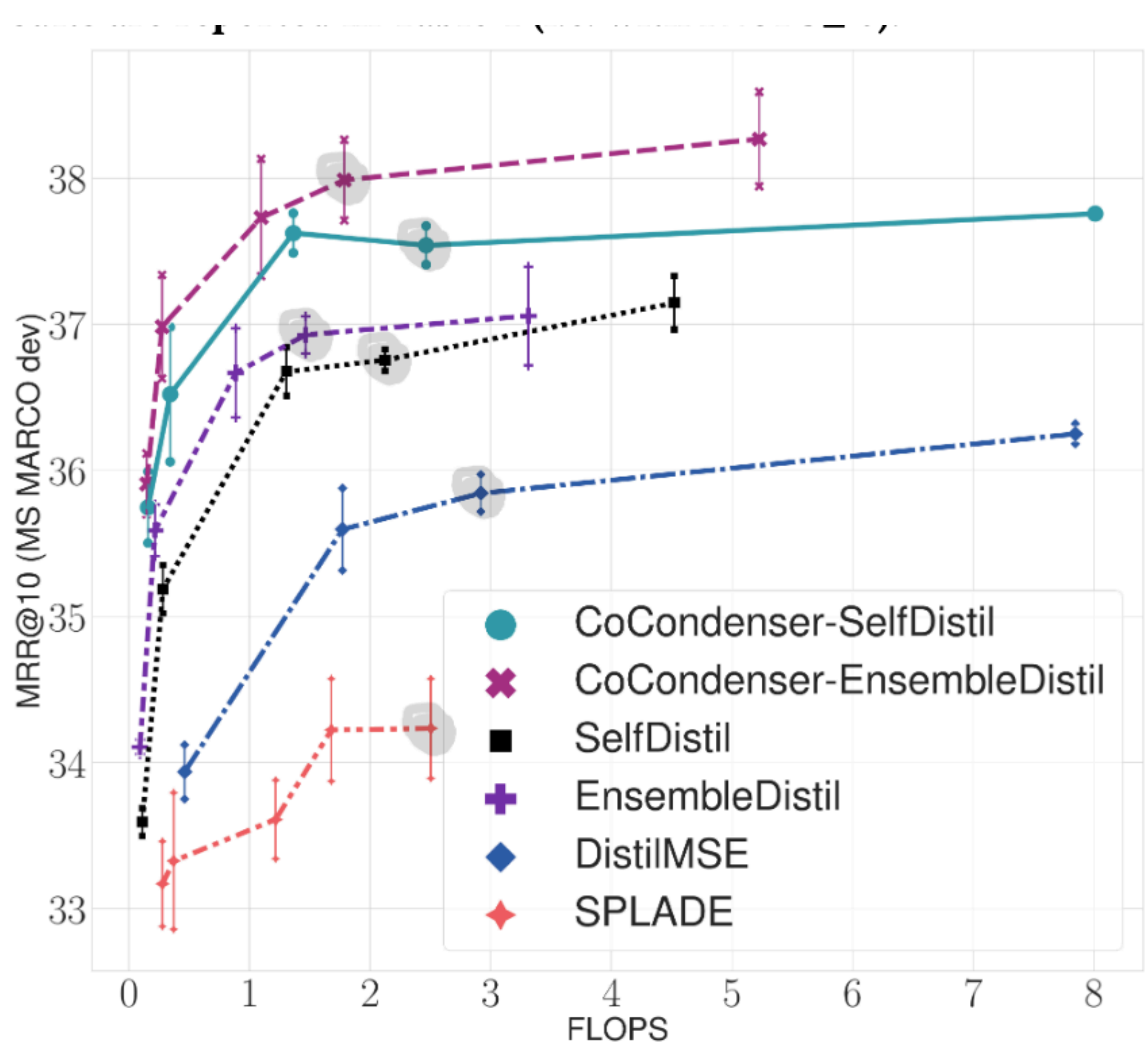
$k = 1000$
RR@100
Time [ms]



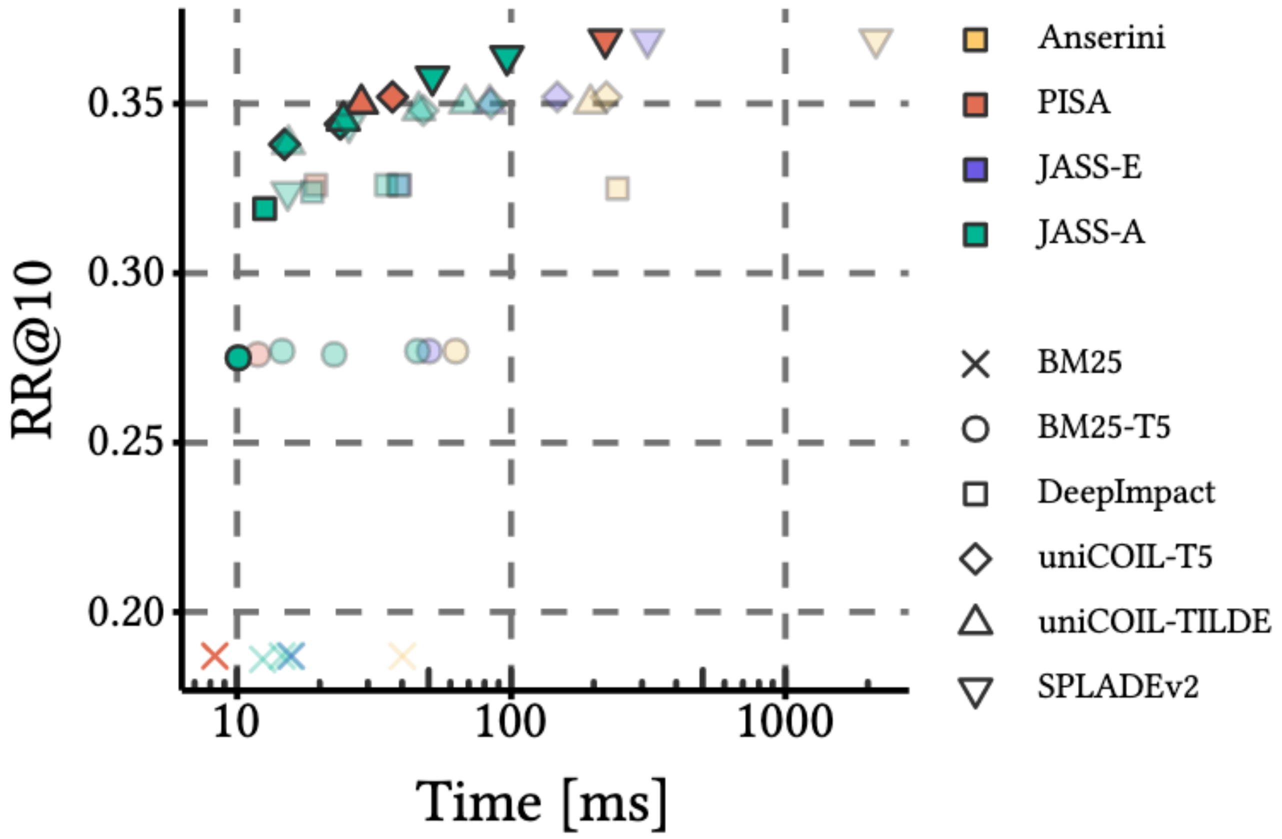
MRR@10
Latency in ms (PISA/FAISS)
v) - PLM with FLOPS
vi) - BERT-tiny
BruteForce - CC
HNSW32 - CC
IVF65536 - CC
HNSW32 - DB-CC
IVF65536 - DB-CC



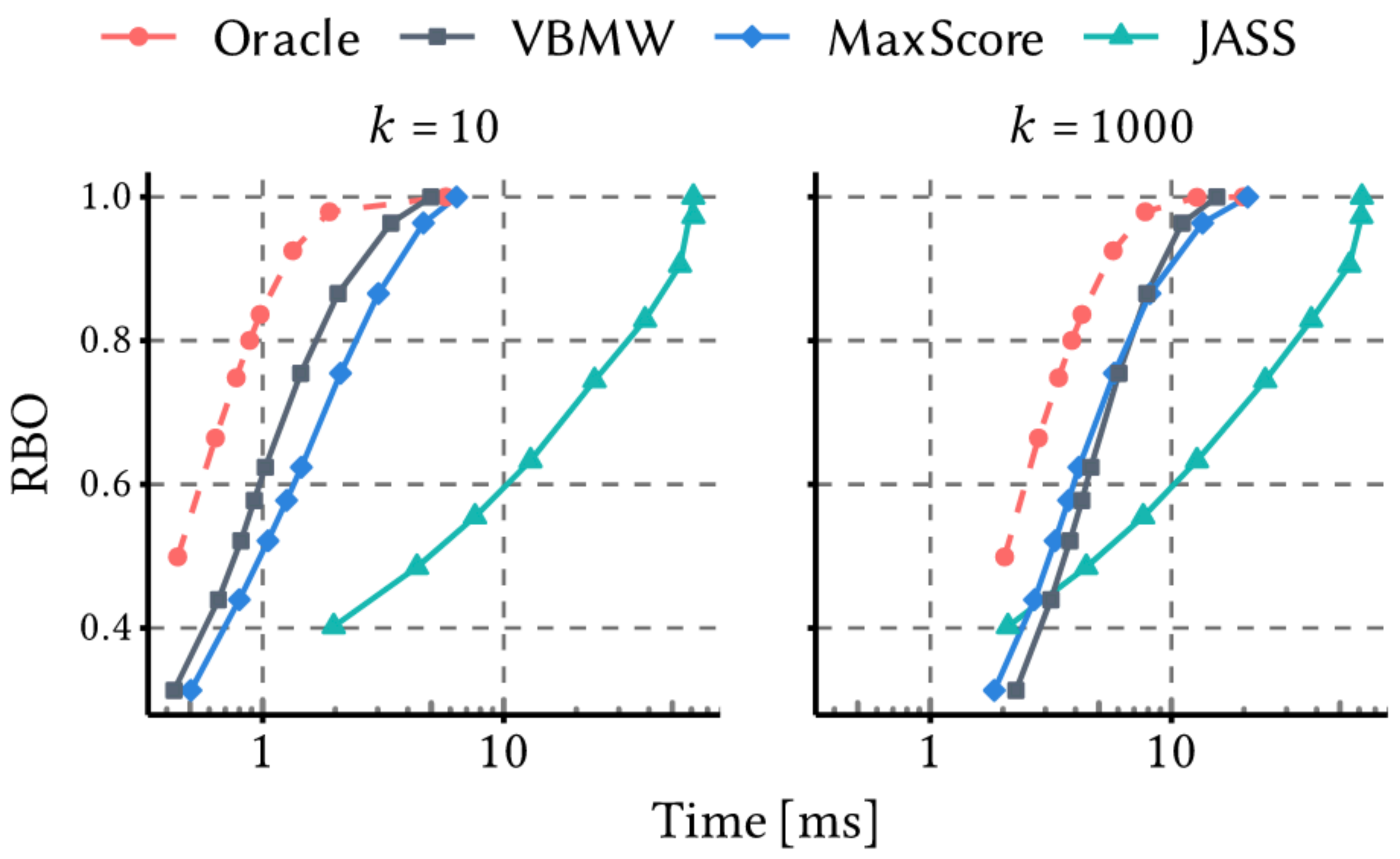
BM25
Mean query latency (ms)
% nDCG@10
BERTLARGE
BERTBASE
BERTDISTIL
CKNRM
LTR
KNRM
Duet
X CK



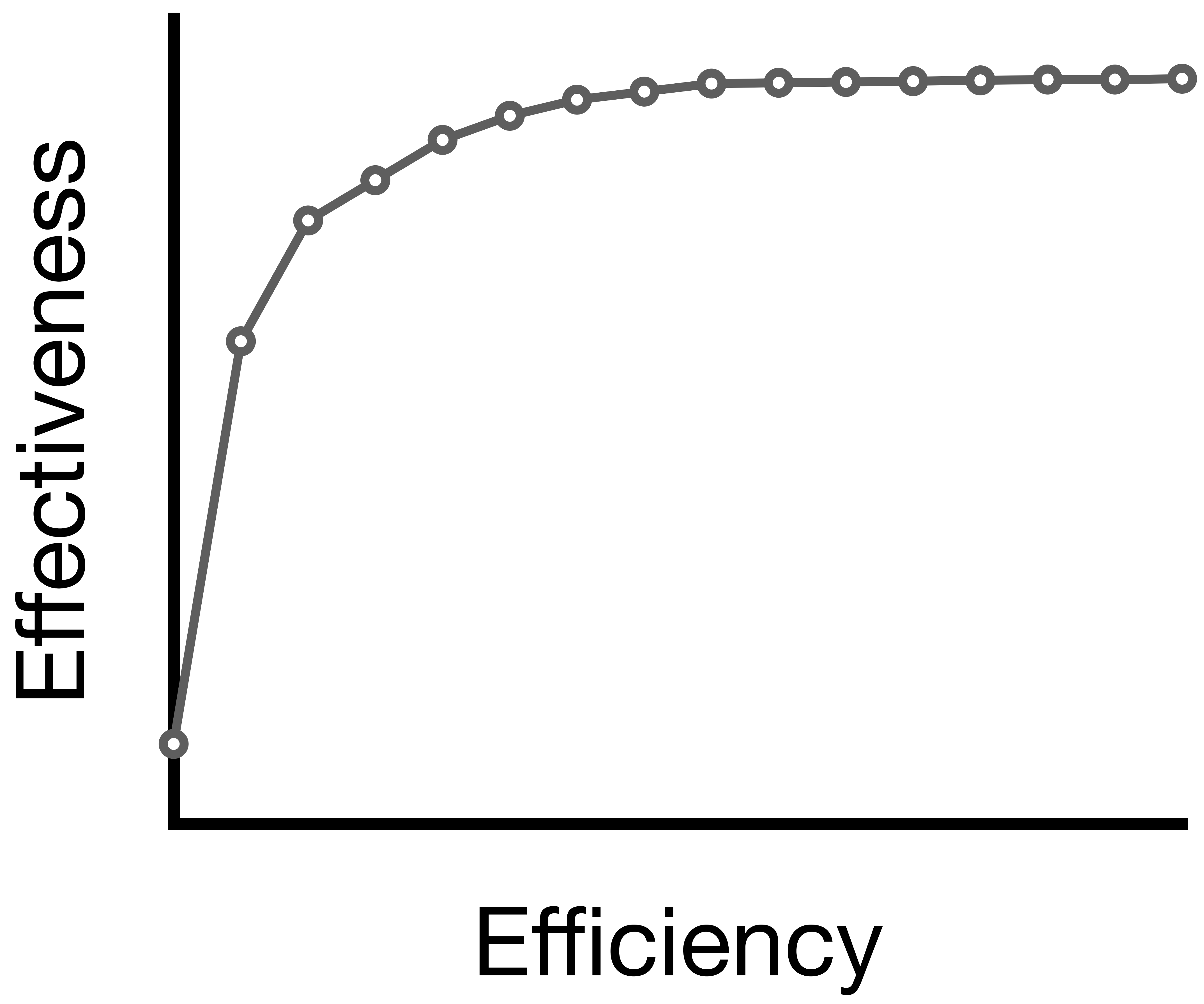
MRR@10 (MS MARCO dev)
FLOPS
CoCondenser-SelfDistil
CoCondenser-EnsembleDistil
SelfDistil
EnsembleDistil
DistilMSE
SPLADE



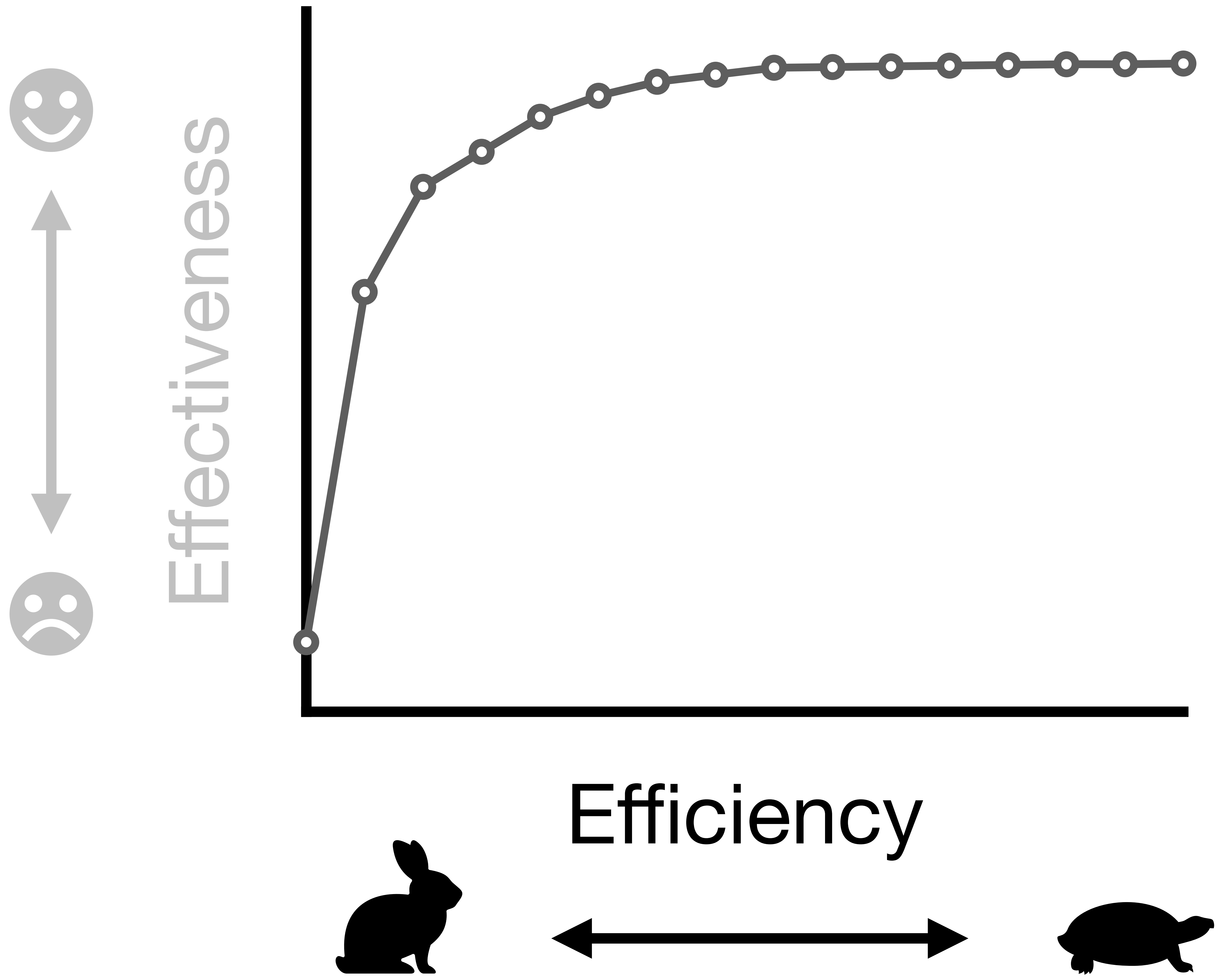
RR@10
Time [ms]
Anserini
PISA
JASS-E
JASS-A
BM25
BM25-T5
DeepImpact
uniCOIL-T5
uniCOIL-TILDE
SPLADEv2



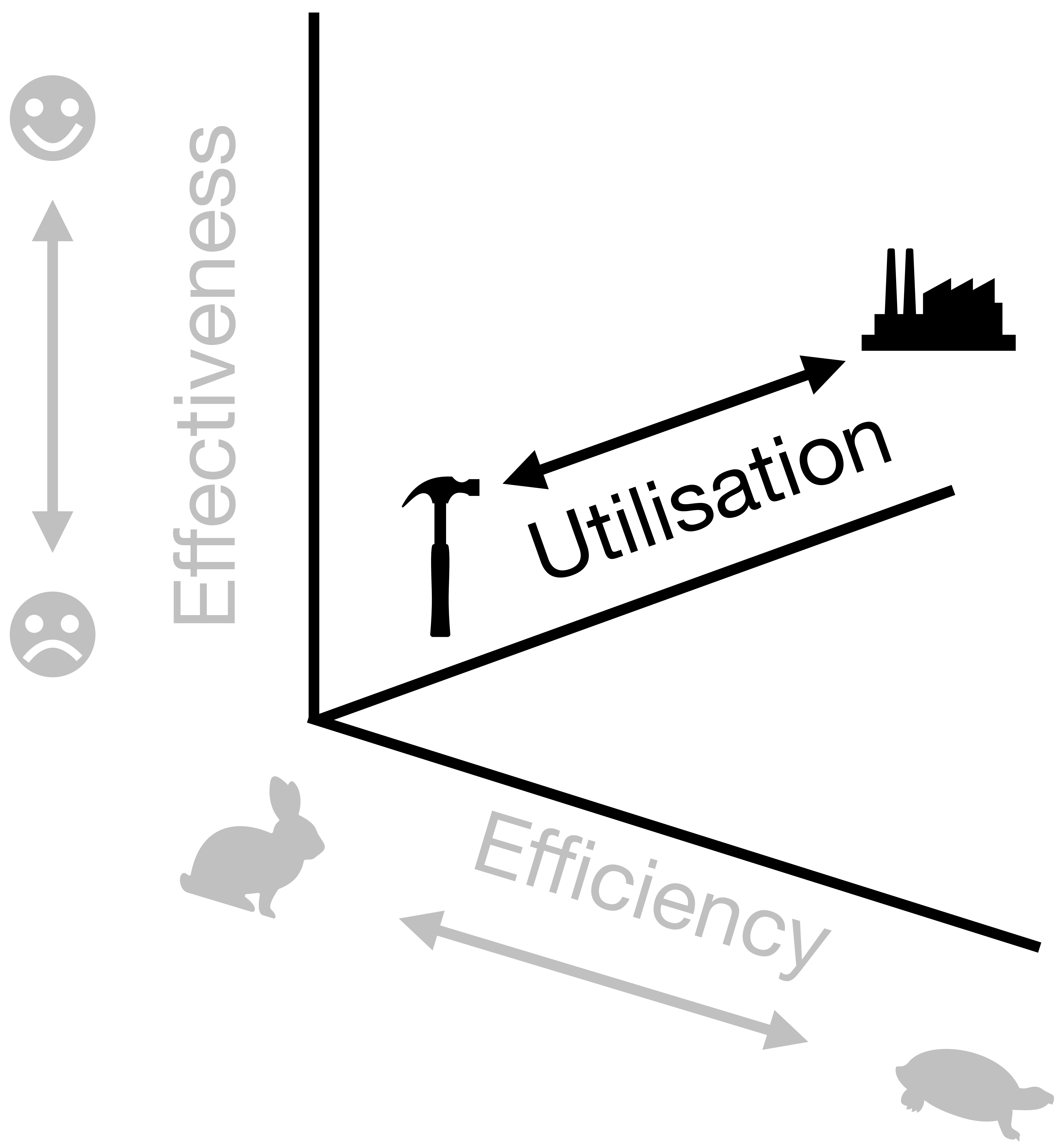
Oracle
VBMW
MaxScore
JASS
$k = 10$
$k = 1000$
RBO
Time [ms]

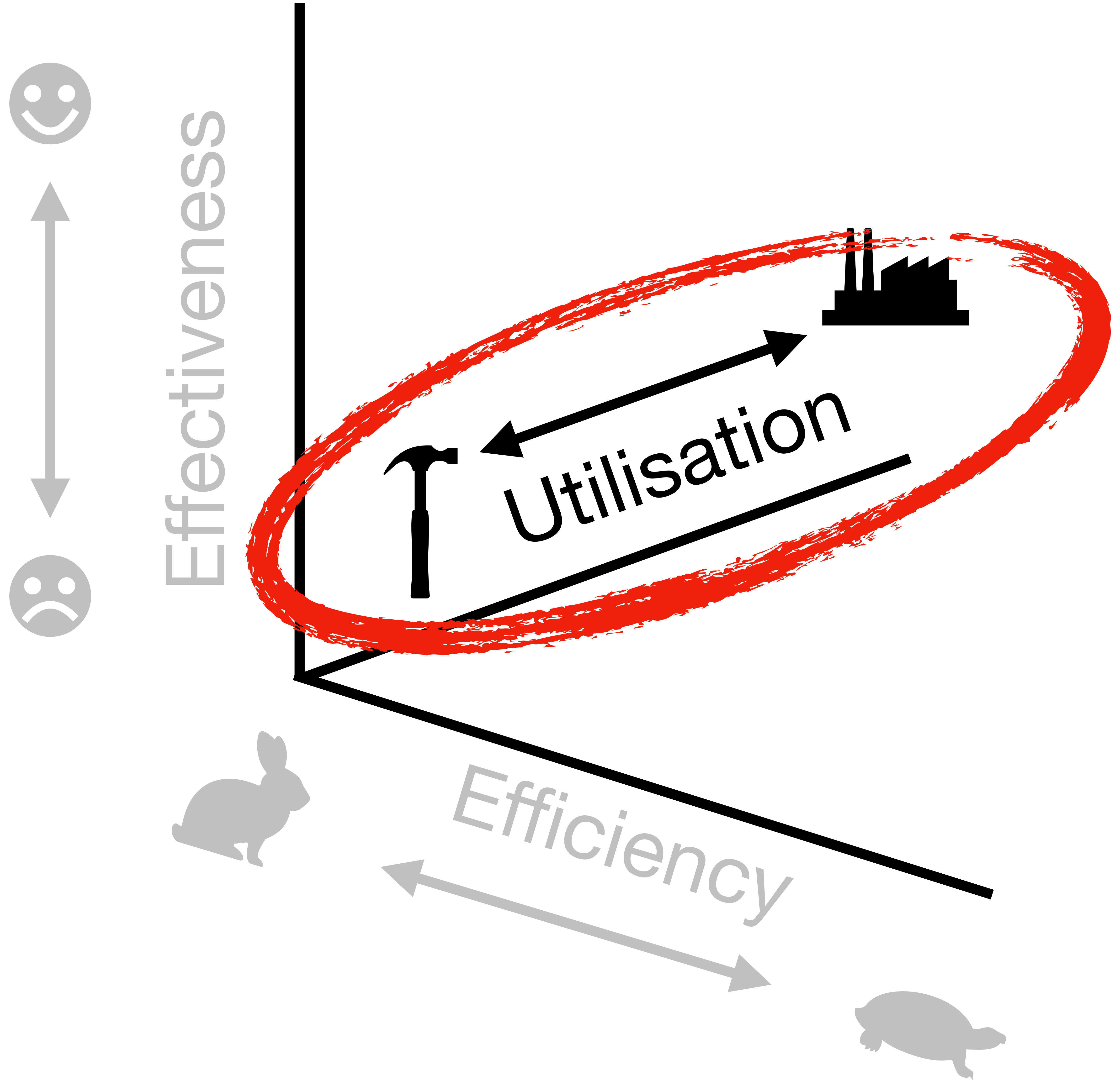

Effectiveness
Efficiency


Effectiveness
Efficiency



Effectiveness
Efficiency


Effectiveness
Utilisation
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*" [2]

[2] *Schwartz, R. et al. 2020. Green AI. Communications of the ACM.*

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*" [2]

- Neural methods require pre-trained LMs
  - **Expensive** to create
  - Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

[3] *Gao, L. and Callan, J. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*

[4] *Ma, X. et al. 2021. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. Proceedings of the 14th ACM International Conference on Web Search and Data Mining*

[5] *Tay, Y. et al. 2022. Transformer Memory as a Differentiable Search Index. arXiv preprint arXiv:2202.06991.*

[6] *Zhou, Y. et al. 2022. DynamicRetriever: A Pre-training Model-based IR System with Neither Sparse nor Dense Index. arXiv preprint*

[2] *Schwartz, R. et al. 2020. Green AI. Communications of the ACM.*

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*" [2]

- Neural methods require pre-trained LMs
  - **Expensive** to create
  - Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

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

[3] *Gao, L. and Callan, J. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*

[4] *Ma, X. et al. 2021. PROP: Pre-training with Representative Words Prediction for Ad-hoc Retrieval. Proceedings of the 14th ACM International Conference on Web Search and Data Mining*

[5] *Tay, Y. et al. 2022. Transformer Memory as a Differentiable Search Index. arXiv preprint arXiv:2202.06991.*

[6] *Zhou, Y. et al. 2022. DynamicRetriever: A Pre-training Model-based IR System with Neither Sparse nor Dense Index. arXiv preprint*

[2] *Schwartz, R. et al. 2020. Green AI. Communications of the ACM.*

Utilisation and Green IR

Green IR is…

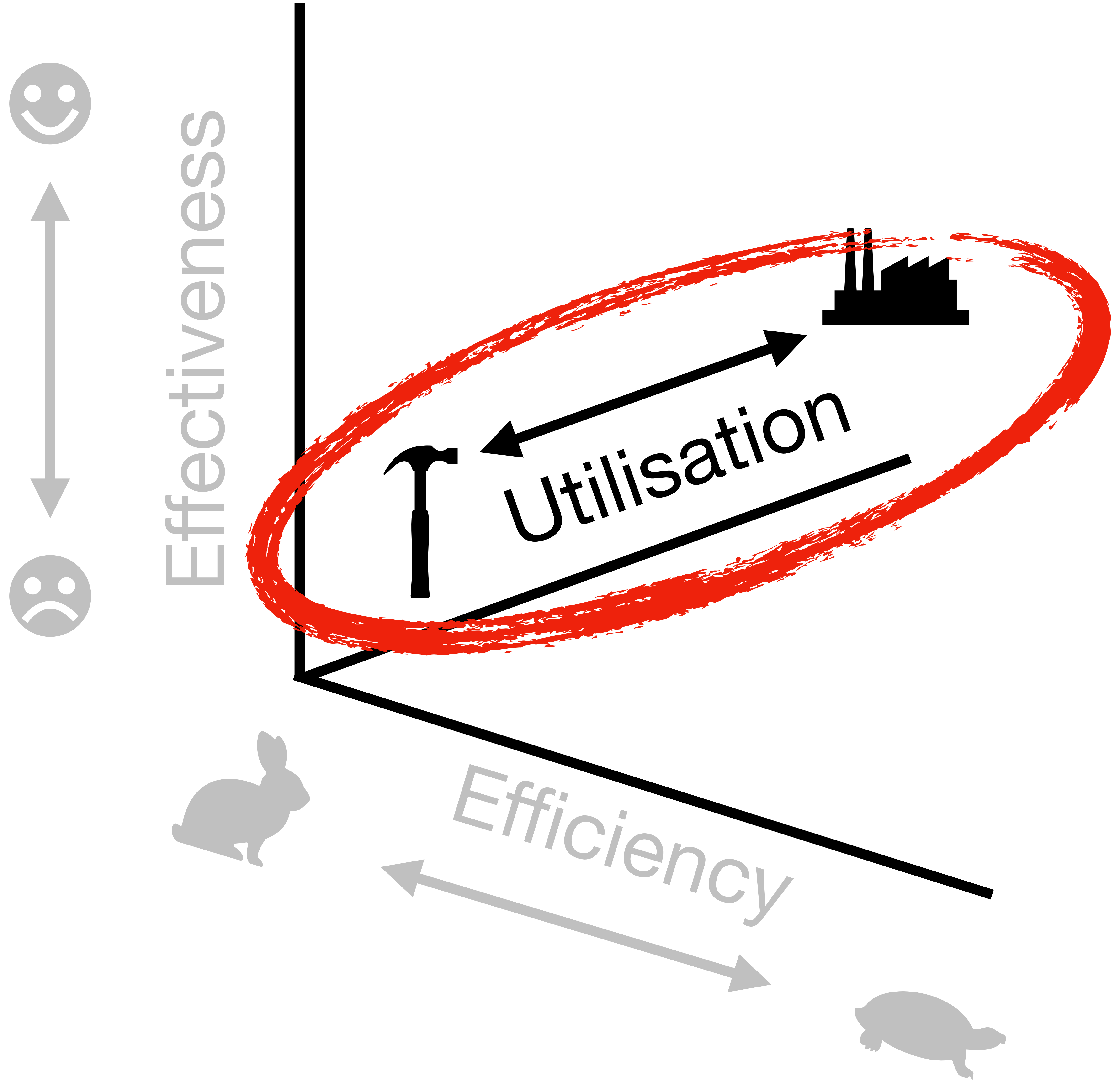
- "*research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent*" [2]

- Neural methods require pre-trained LMs
  - **Expensive** to create
  - Trend in IR towards creating **IR-specific** LMs [3,4,5,6]

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

- Missing dimension of IR evaluation
  - Effectiveness
  - Efficiency
  - **Utilisation**

[2] *Schwartz, R. et al. 2020. Green AI. Communications of the ACM.*



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

Okay, so how can I measure this?

Measuring emissions

- First, measure power consumption:

Measuring emissions

- First, measure power consumption:

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

Measuring emissions

- First, measure power consumption:

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

Measuring emissions

- First, measure power consumption:

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

Measuring emissions

- First, measure power consumption:

Running Time

PUE

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

watts

Measuring emissions

- First, measure power consumption:

Running Time

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

watts

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

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:

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:

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

Measuring emissions

- First, measure power consumption:

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

- Next, measure emissions:

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

Measuring 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

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

Power consumption
of experiments

Measuring emissions

- First, measure power consumption:

Running Time

PUE

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

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

Power consumption of experiments

Measuring emissions

- 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:

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

- Emissions of my search engine:

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

Measuring emissions

- First, measure power consumption:

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

- Next, measure emissions:

$$\underset{\text{emissions}}{\mathbf{kgCO_2e}} = \underset{\text{avg. } CO_2\text{e (kg) per kWh where experiments took place}}{\theta} \cdot \underset{\text{Power consumption of experiments}}{p_t}$$

- Emissions of my search engine:

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

Measuring emissions

- First, measure power consumption:

Running Time

PUE

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

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

Power consumption of experiments

- Emissions of my search engine:

No. queries issued per unit time

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

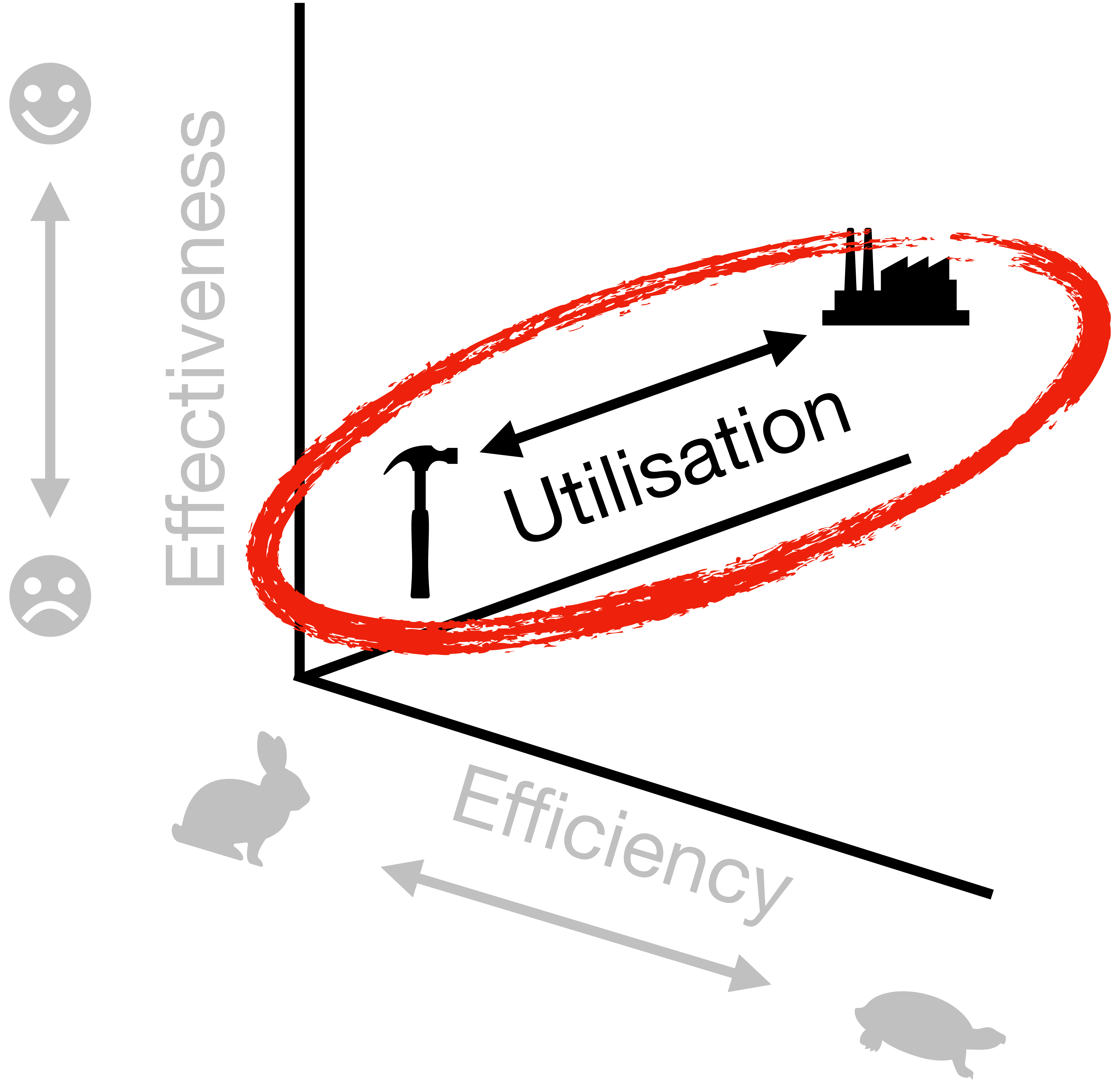
Power consumption of a single query

Measuring power and emissions in practice

Name	CPU	DRAM	GPU	Network	Repository
CodeCarbon [71]	✓	✓	✓	✗	https://github.com/mlco2/codecarbon
pyJoules	✓	✓	✓	✗	https://github.com/powerapi-ng/pyJoules
energyusage [47]	✓	✓	✓	✗	https://github.com/responsibleproblemsolving/energy-usage
Carbontracker [3]	✓	✗	✓	✗	https://github.com/lfwa/carbontracker
Experiment Impact Tracker [33]	✓	✗	✓	✗	https://github.com/Breakend/experiment-impact-tracker
Cumulator [81]	✓	✓	✓	✓	https://github.com/epfl-iglobalhealth/cumulator

```
from codecarbon import EmissionsTracker

tracker = EmissionsTracker()
tracker.start()
# Experiment code goes here
tracker.stop()
```



~~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!

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

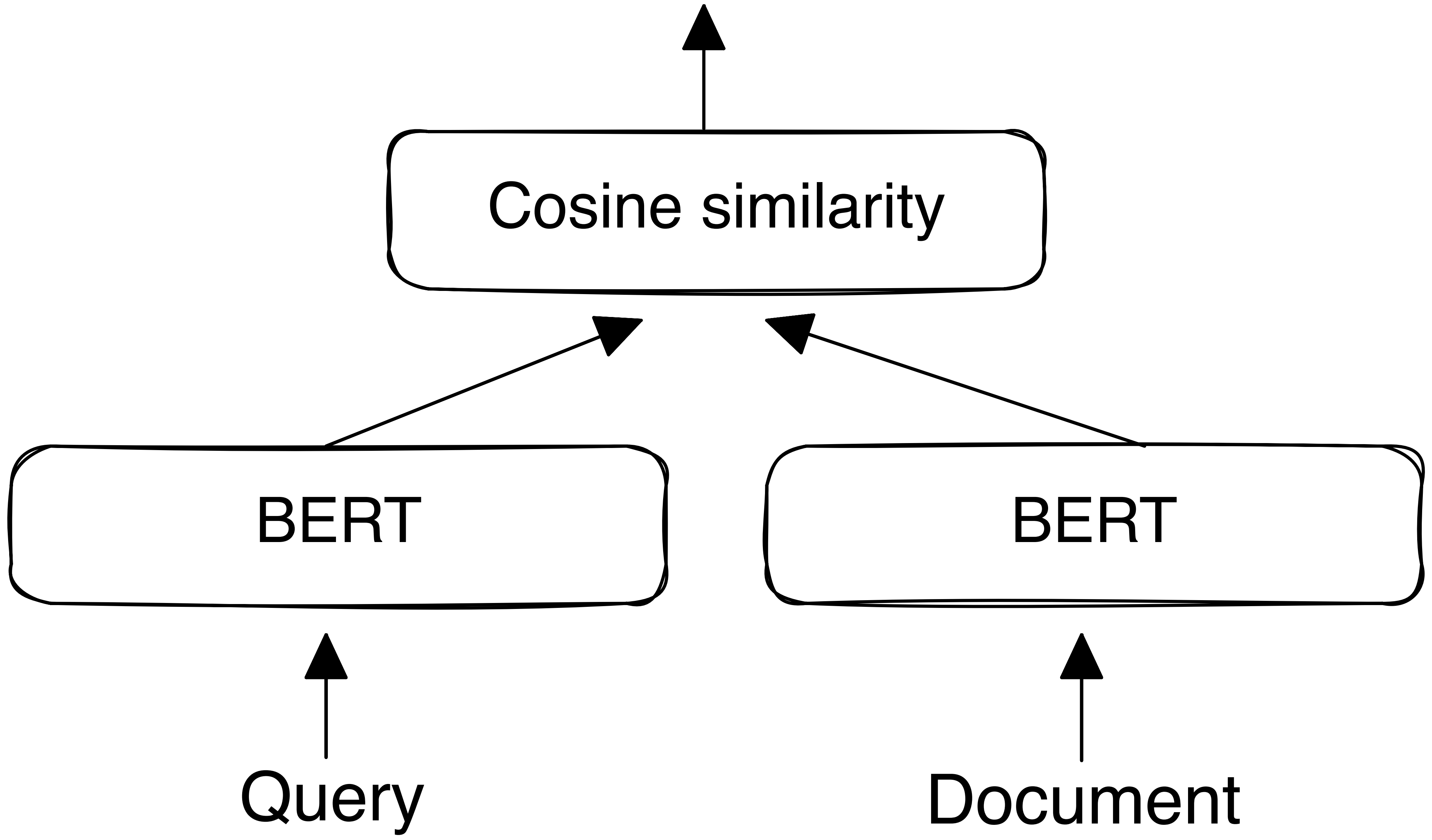
“Neural”

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Dense retriever (bi-encoder)



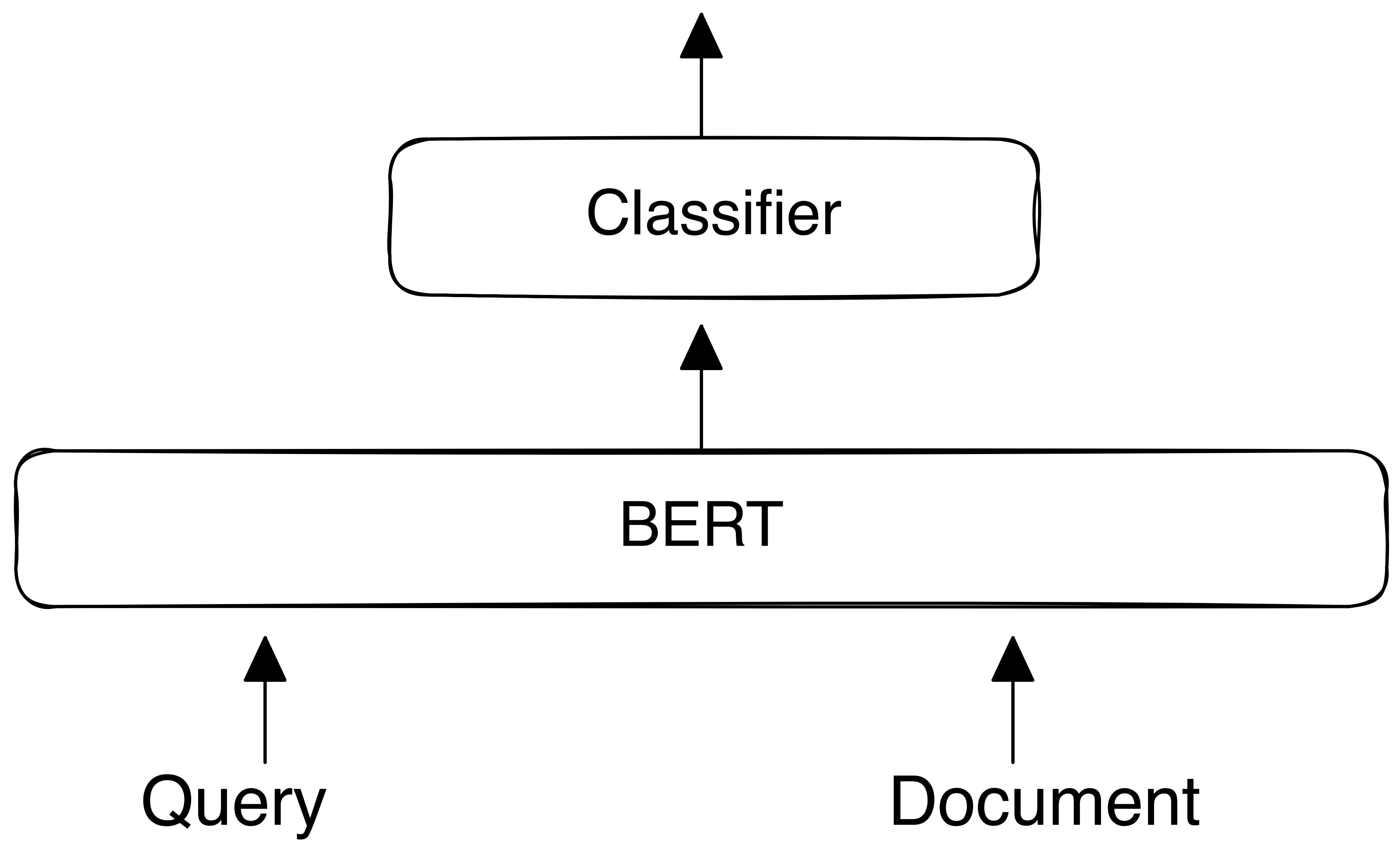
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Dense retriever (bi-encoder)

BERT (cross-encoder)



Experimental Setup Overview

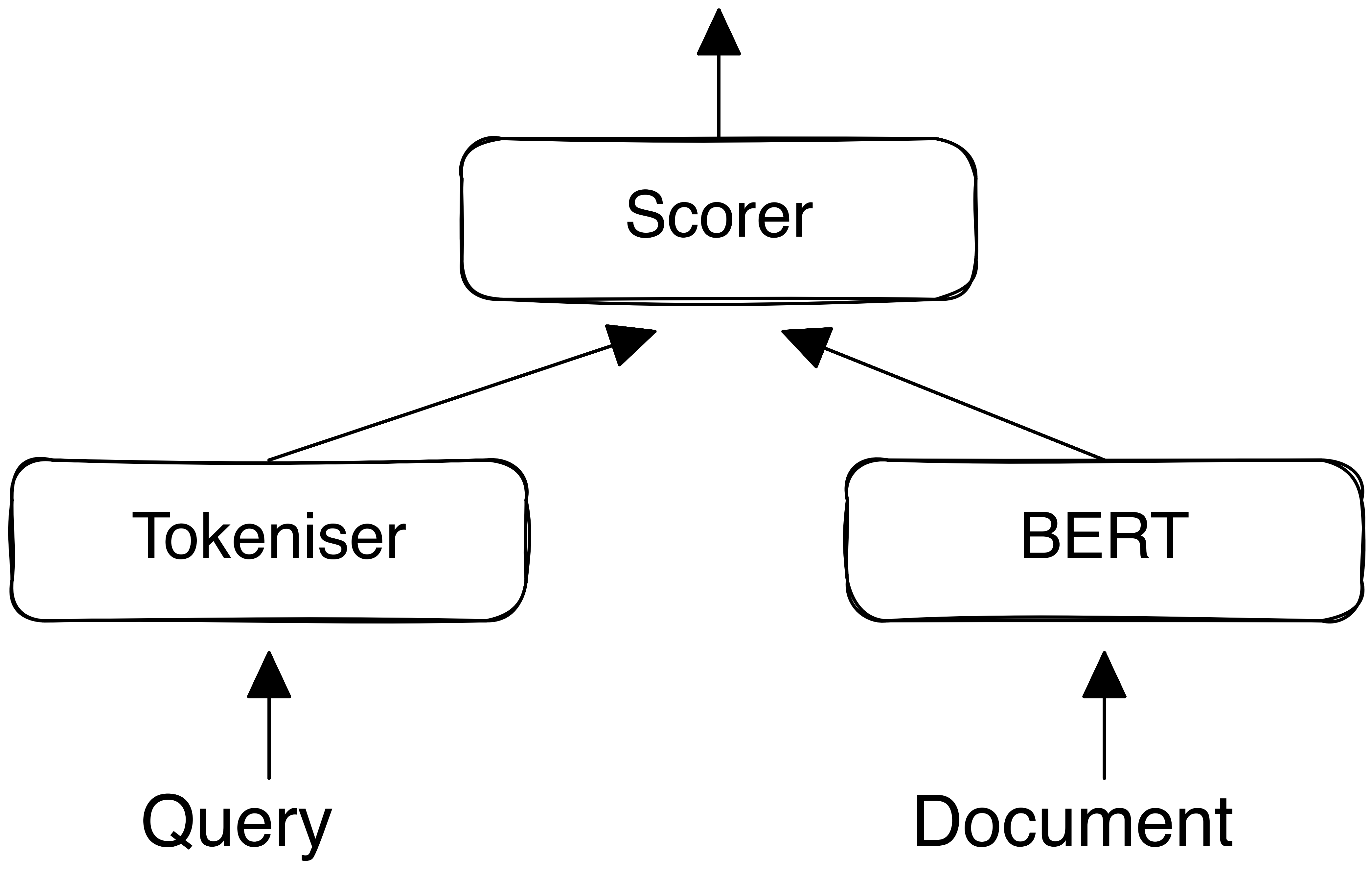
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Dense retriever (bi-encoder)

BERT (cross-encoder)

Sparse retrievers



Experimental Setup Overview

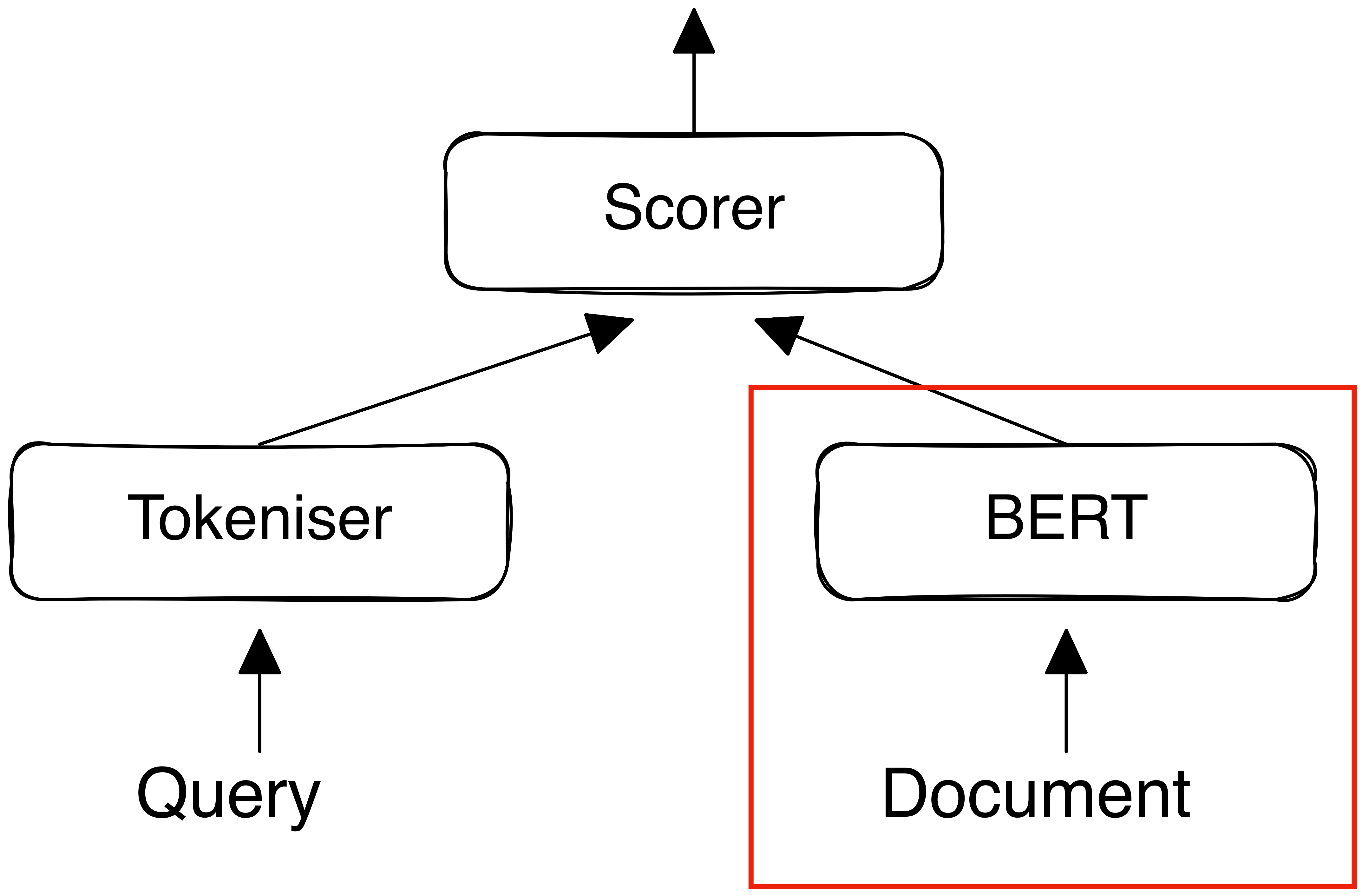
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Dense retriever (bi-encoder)

BERT (cross-encoder)

Sparse retrievers



Process documents offline

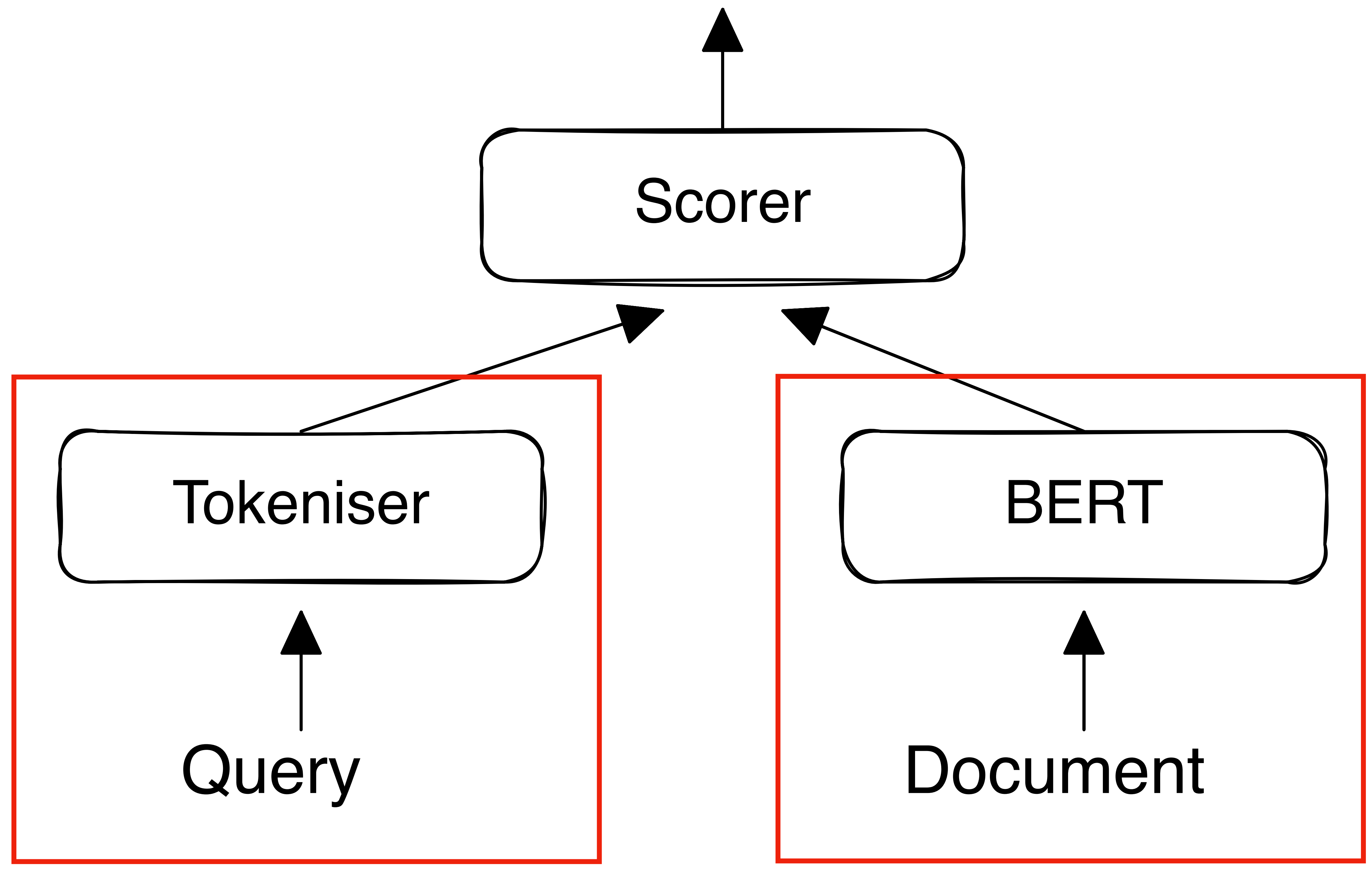
Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART

Non-neural

  - DPR **Dense retriever (bi-encoder)**
  - monoBERT **BERT (cross-encoder)**
  - uniCOIL
  - TILDEv2

Sparse retrievers



Fast inference time
(Can even be done on CPU)

Process documents offline

Experimental Setup Overview

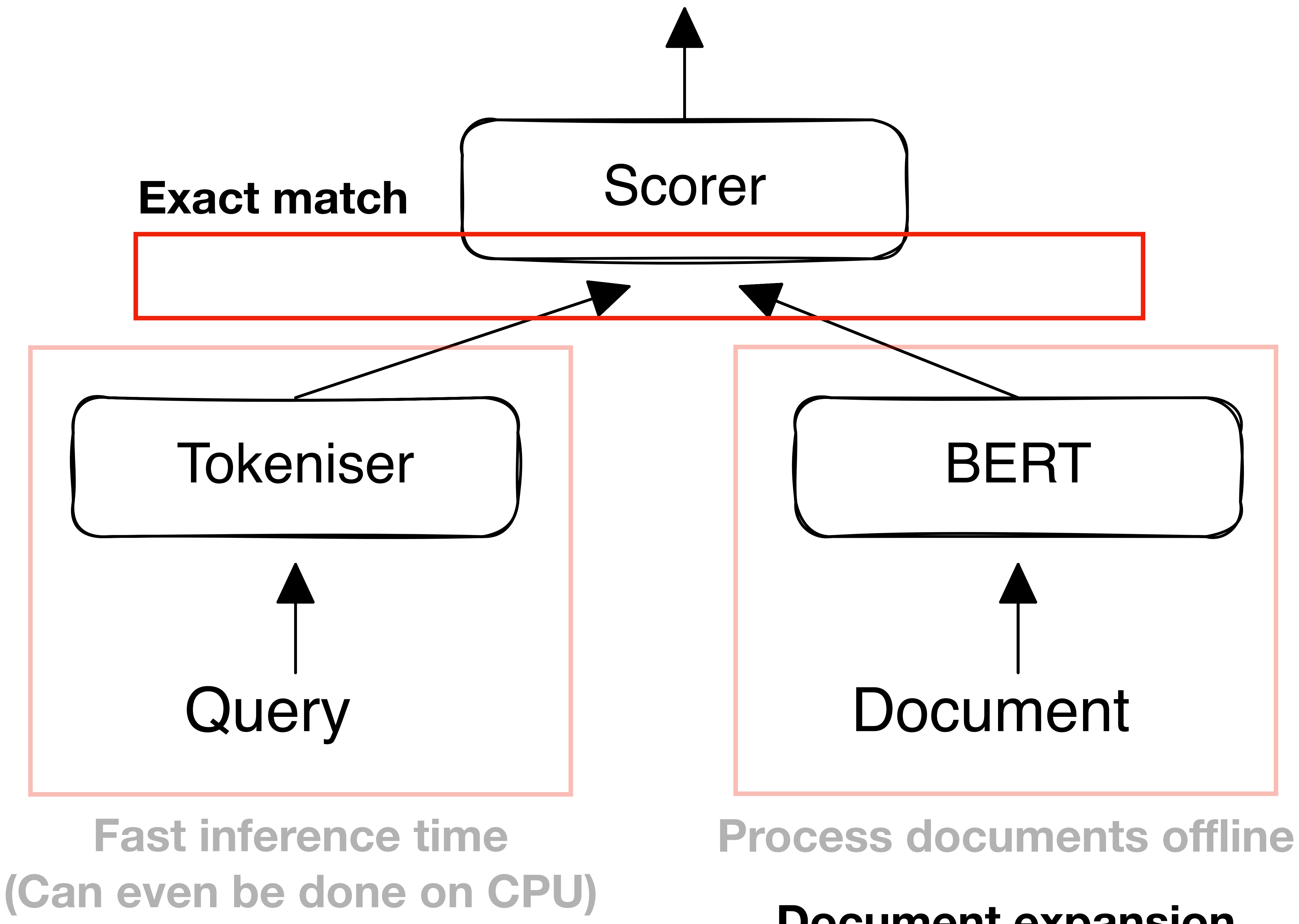
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

Dense retriever (bi-encoder)

BERT (cross-encoder)

Sparse retrievers



Document expansion
TILDE/doc2query

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

“Neural”

- Collection:
  - MSMARCOv1
- Experiments:

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

"Neural"

- Collection:
  - MSMARCOv1
- Experiments:
  - How many emissions do these methods produce to obtain an experimental result?

Experimental Setup Overview

- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

"Neural"

- Collection:
  - MSMARCOv1
- Experiments:
  - How many emissions do these methods produce to obtain an experimental result?
  - What are the effectiveness-utilisation trade-offs of these methods?

Experimental Setup Overview

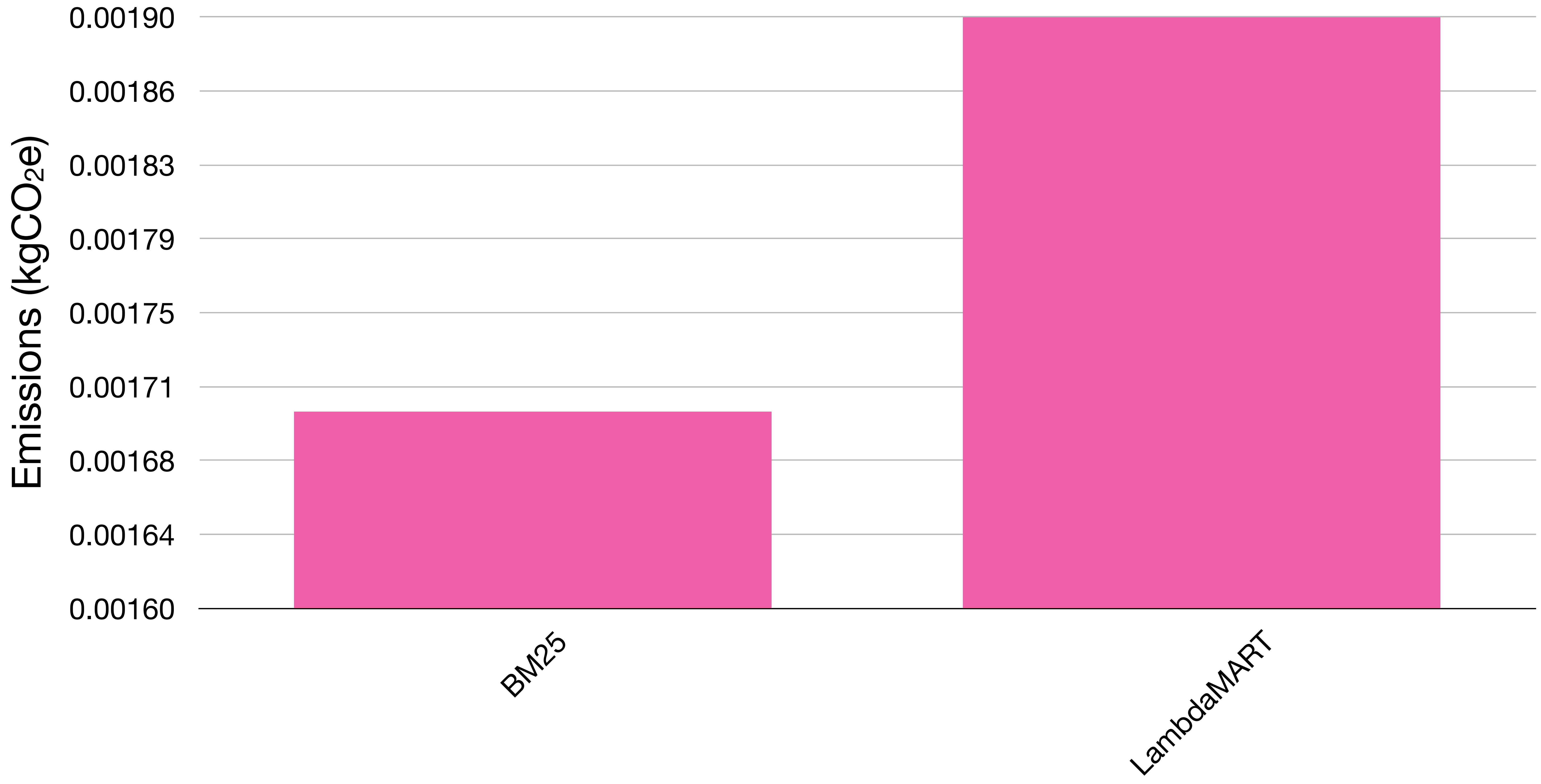
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - uniCOIL
  - TILDEv2

Non-neural

“Neural”

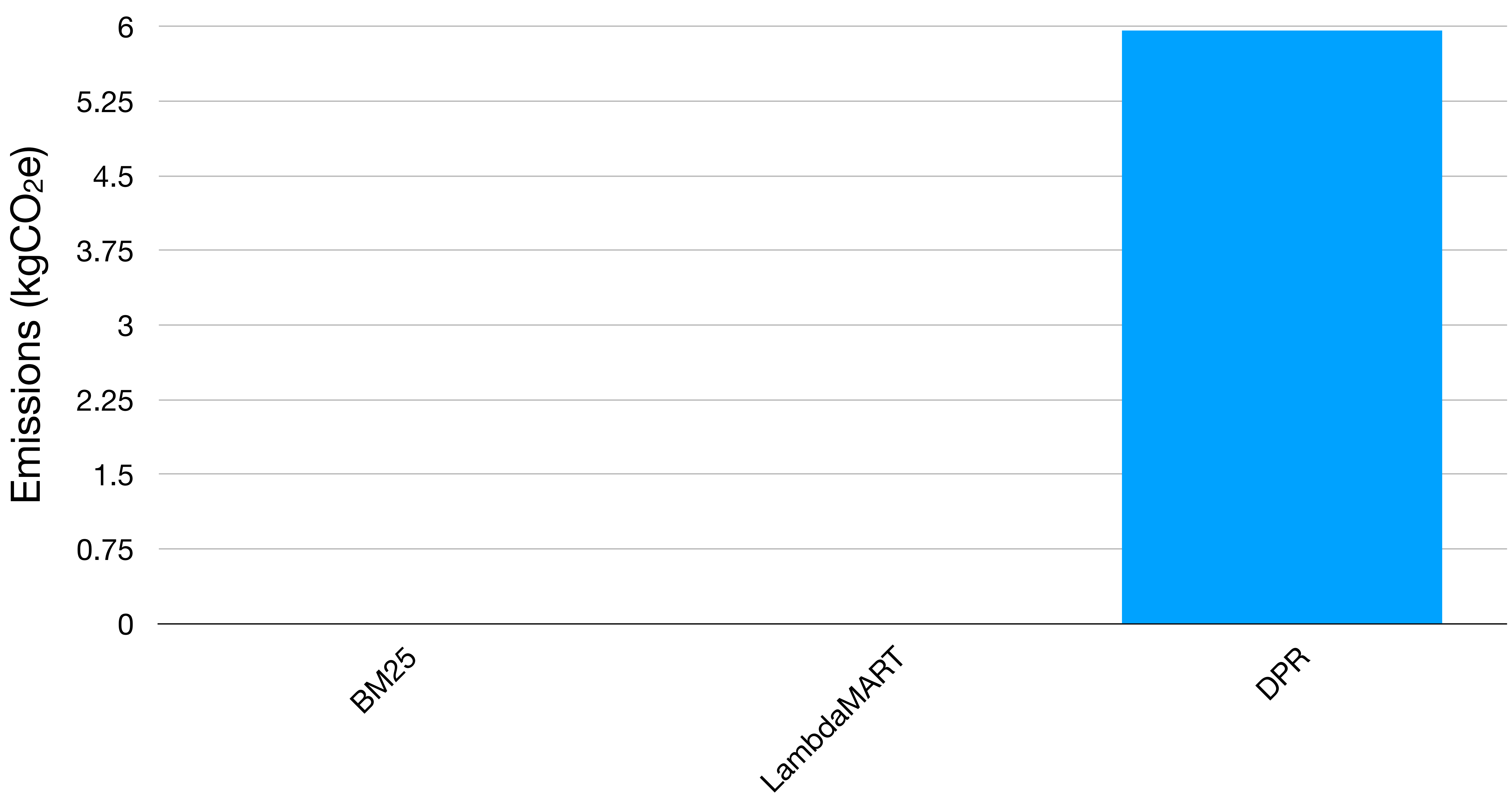
- Collection:
  - MSMARCOv1
- Experiments:
  - How many emissions do these methods produce to obtain an experimental result?
  - What are the effectiveness-utilisation trade-offs of these methods?

How many emissions do these methods produce to obtain an experimental result?



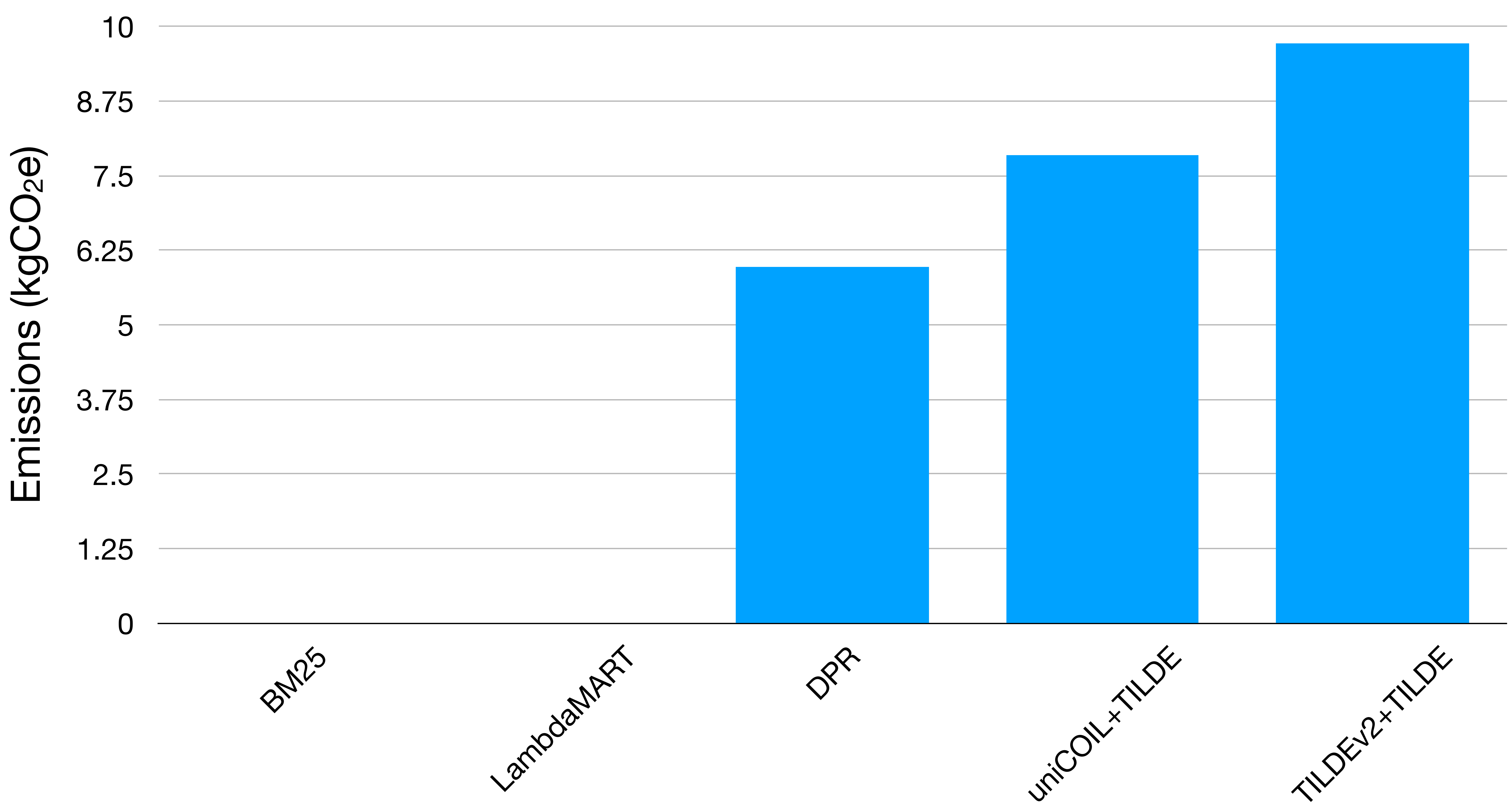
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?



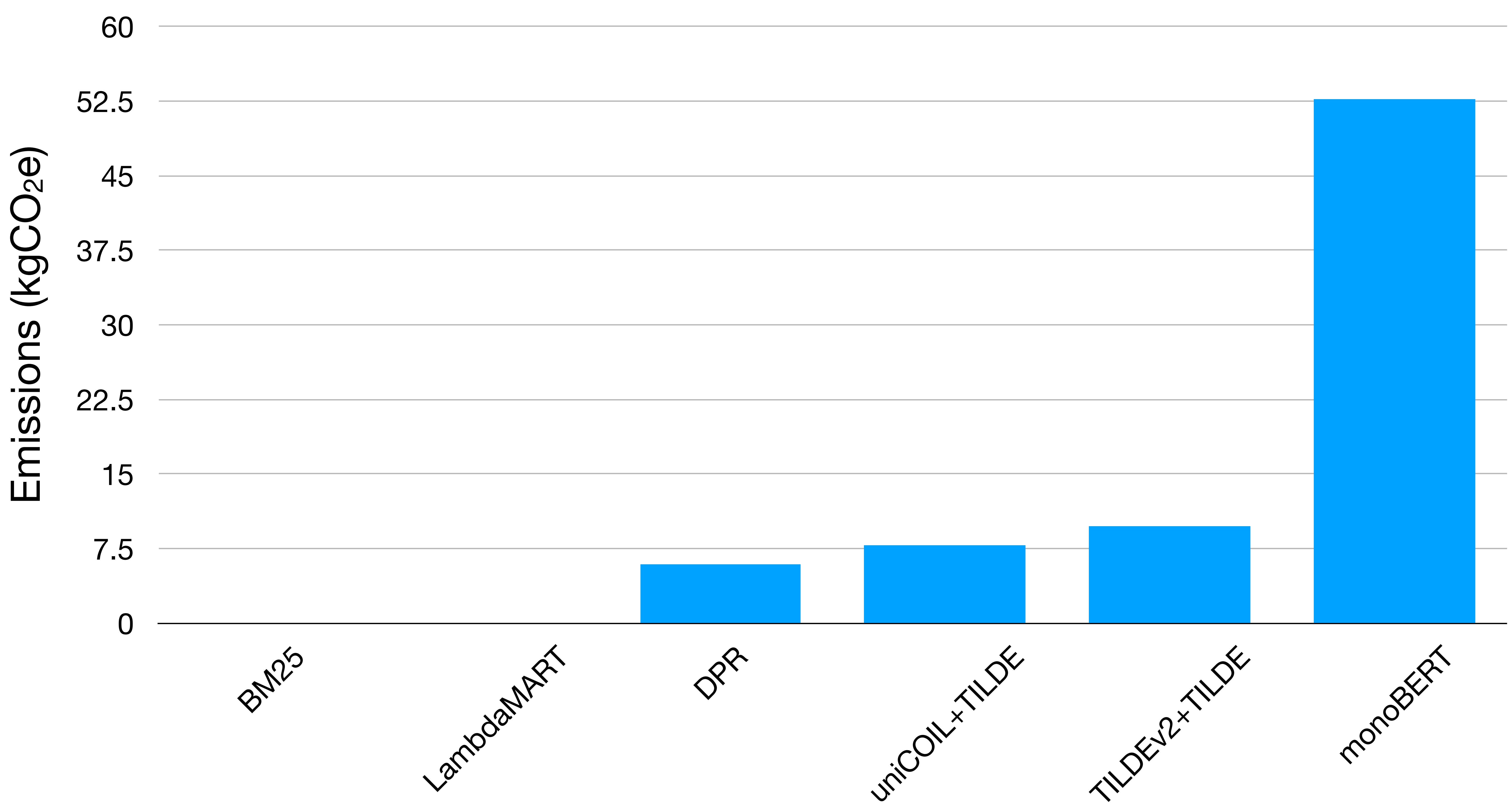
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?



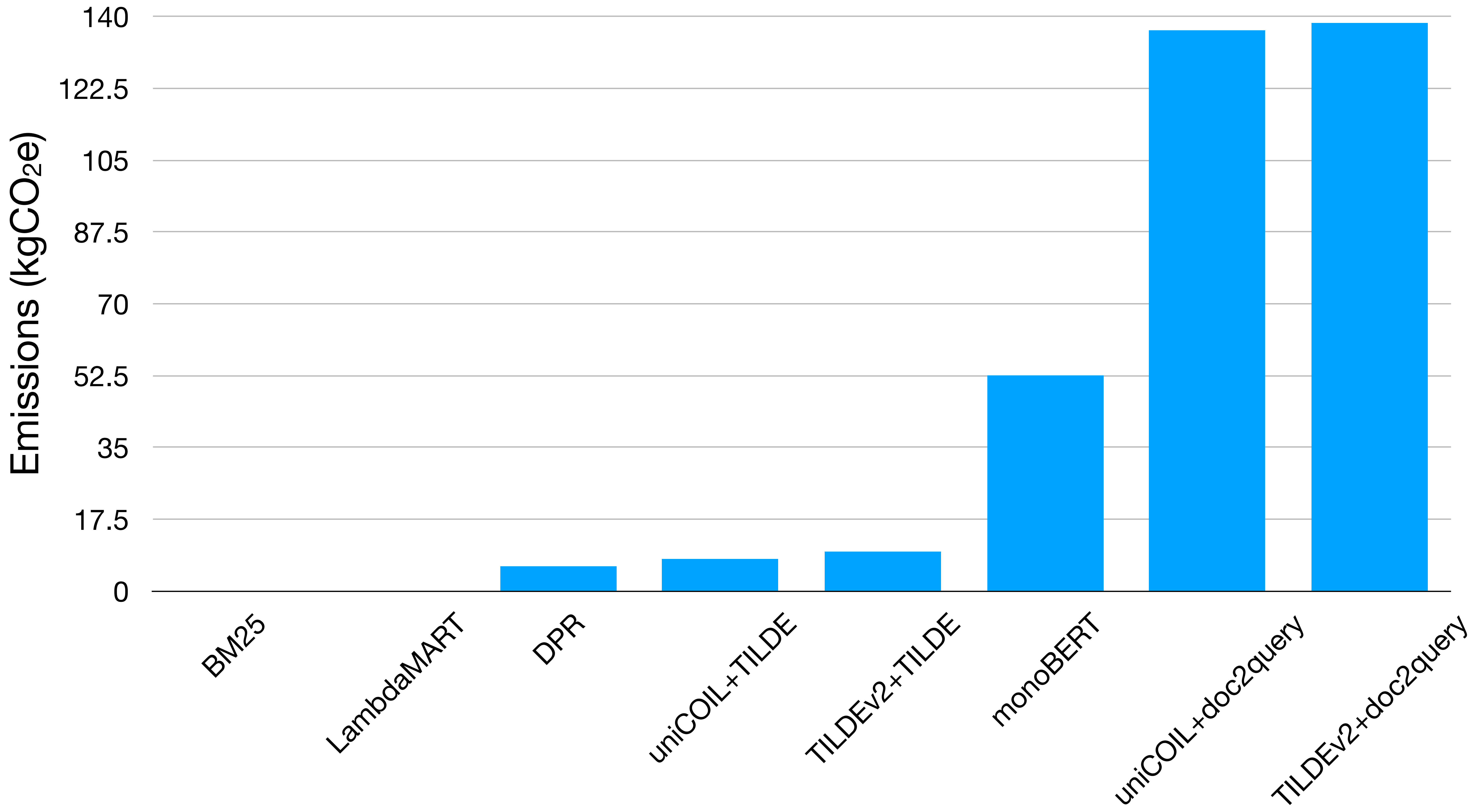
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?



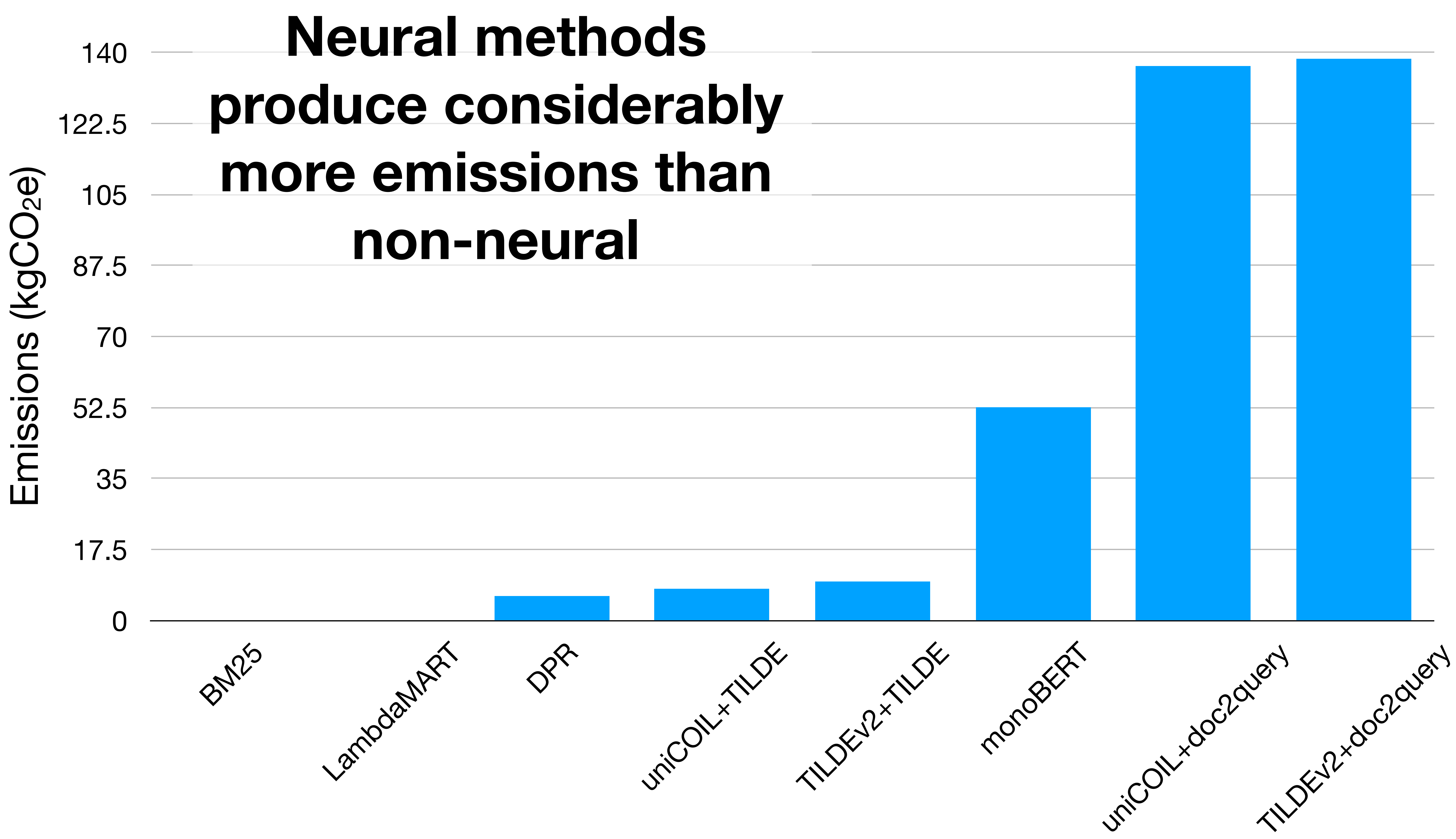
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?



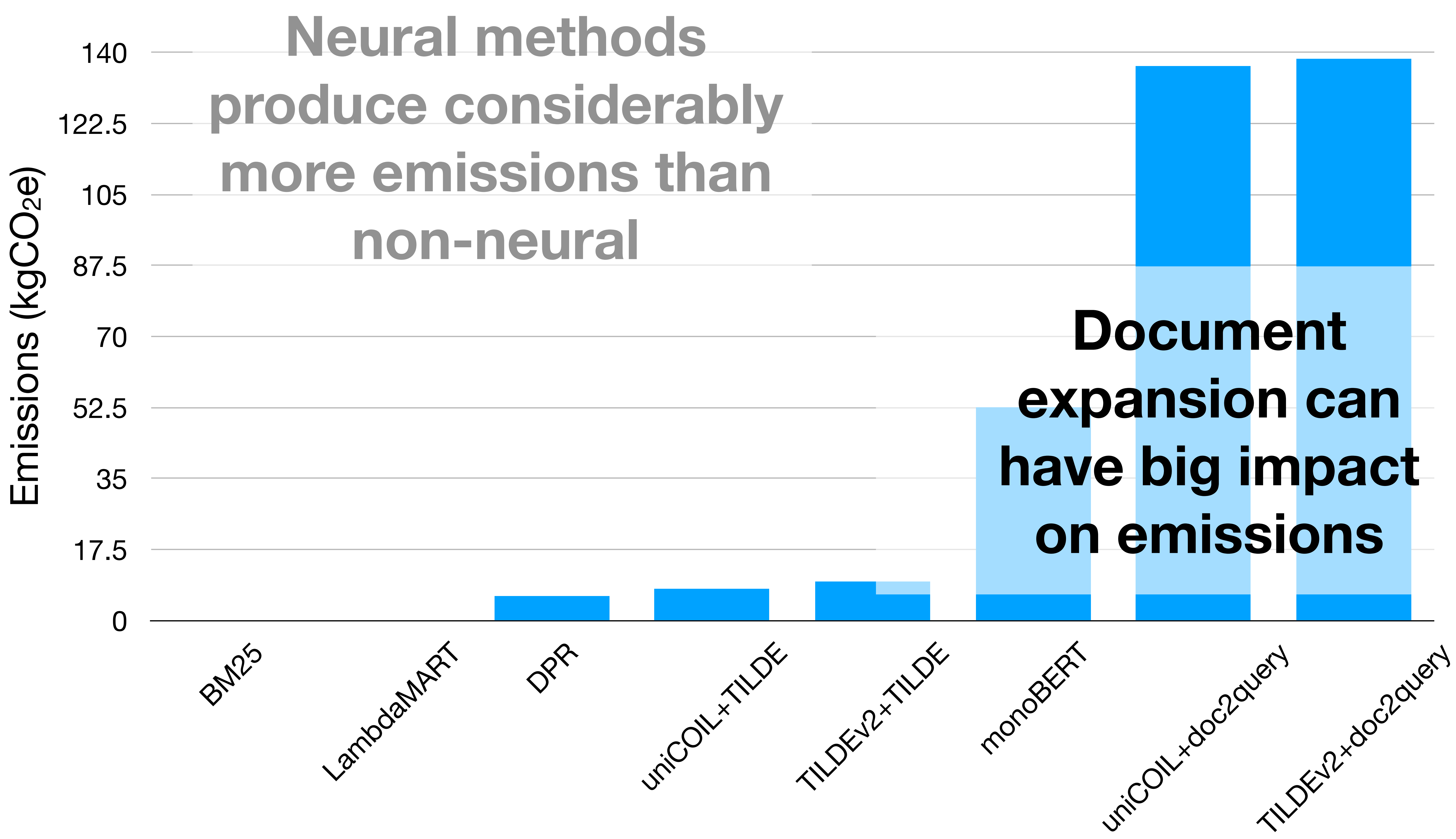
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?



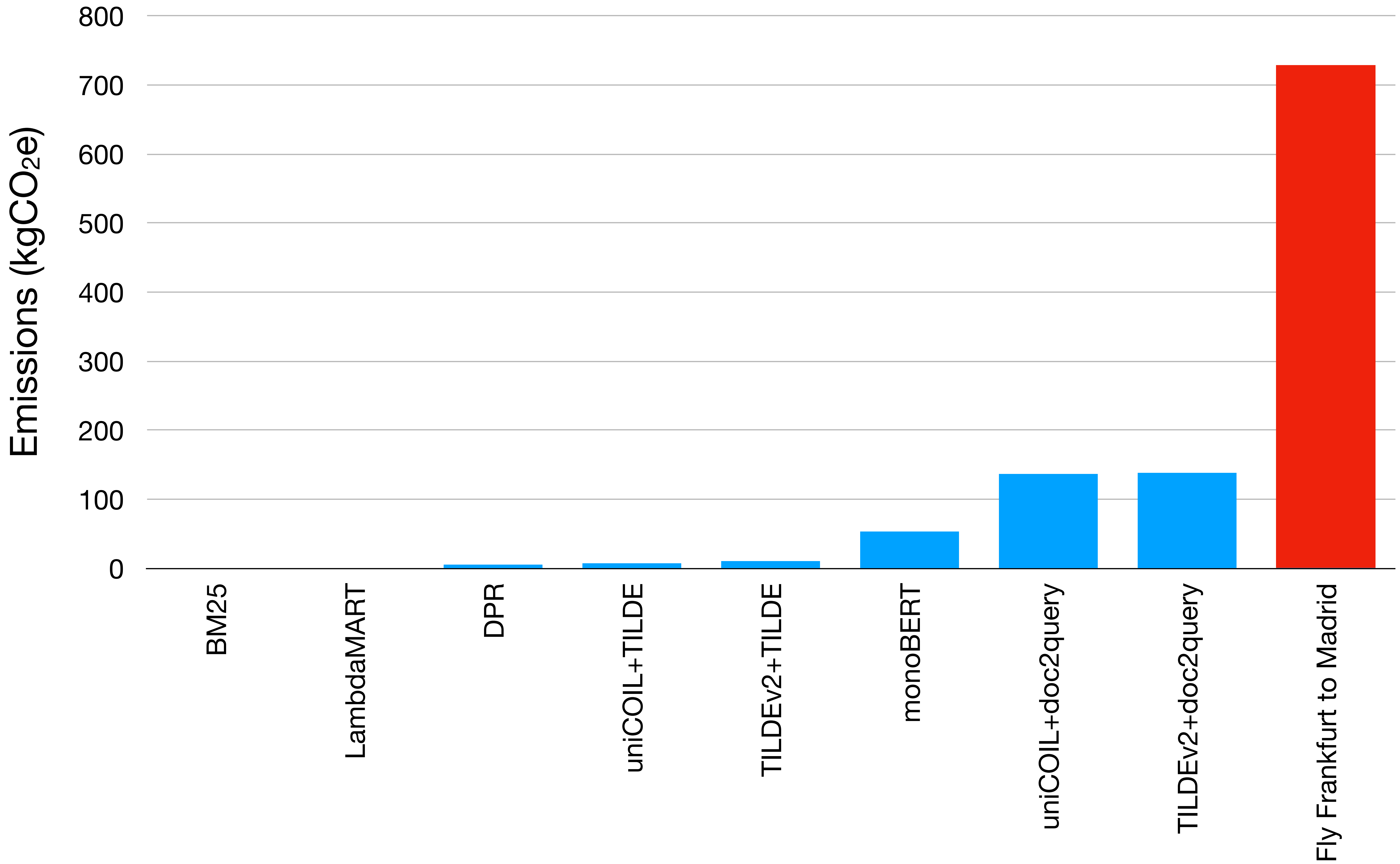
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

How many emissions do these methods produce to obtain an experimental result?

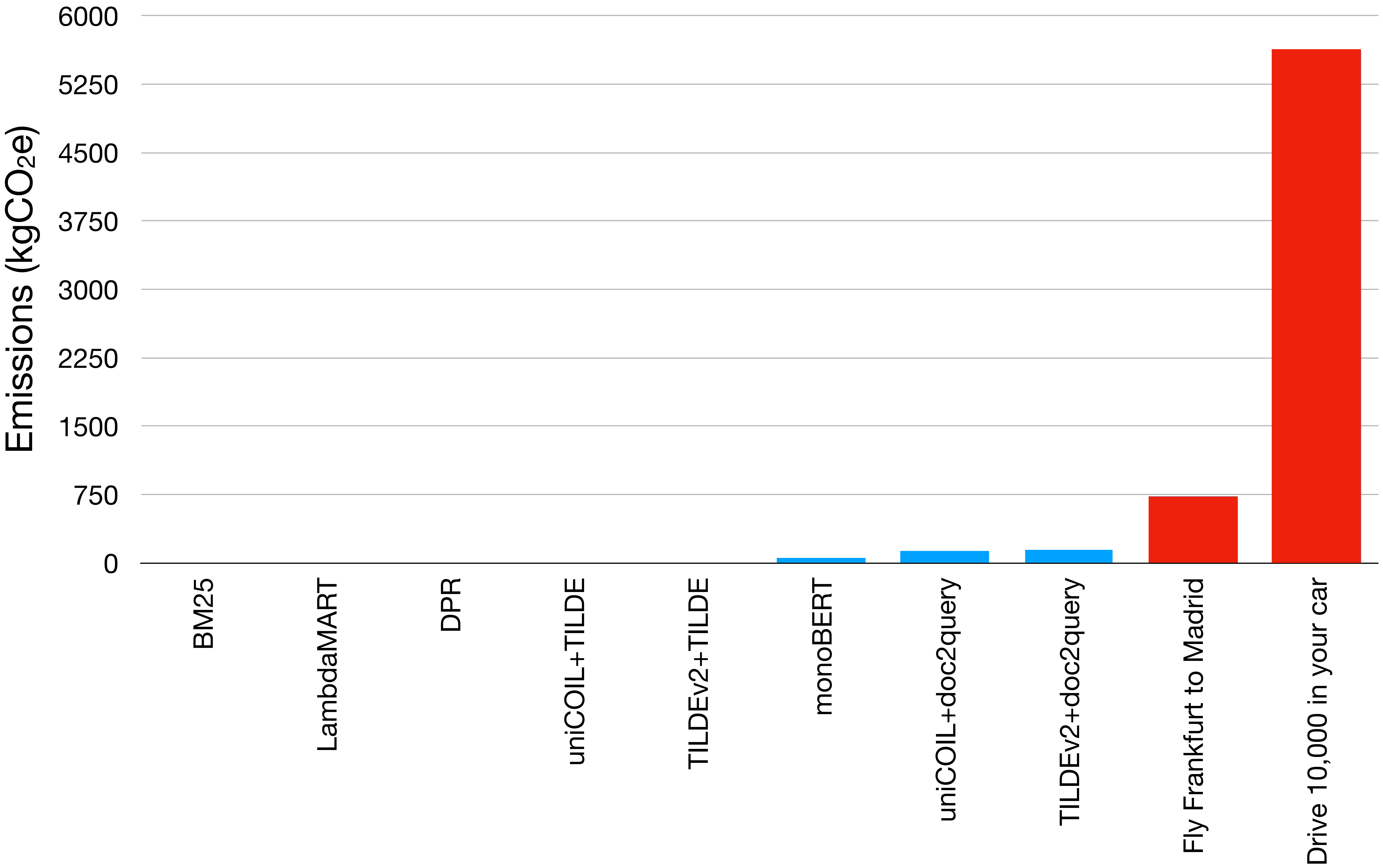


- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

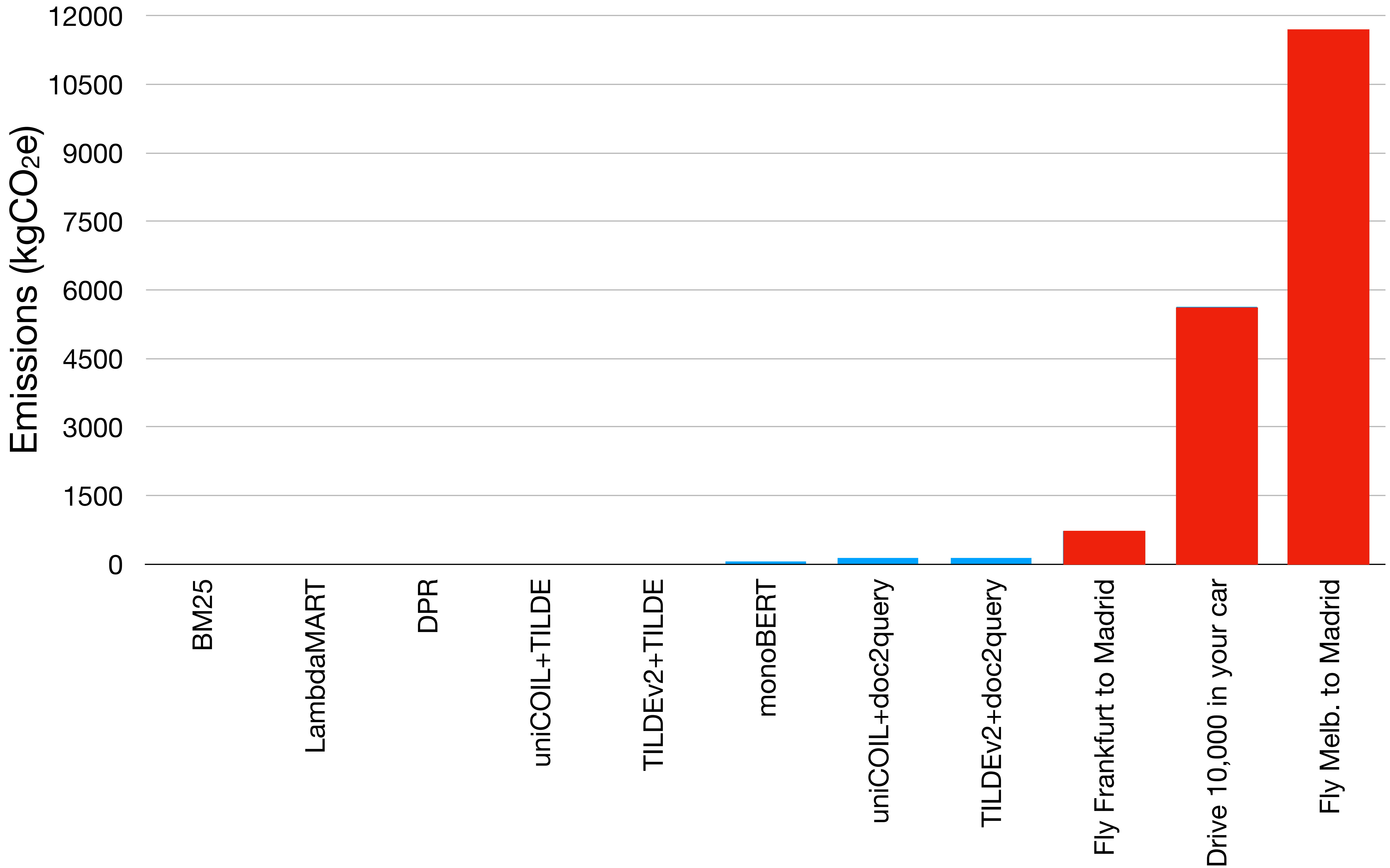
How many emissions do these methods produce to obtain an experimental result?



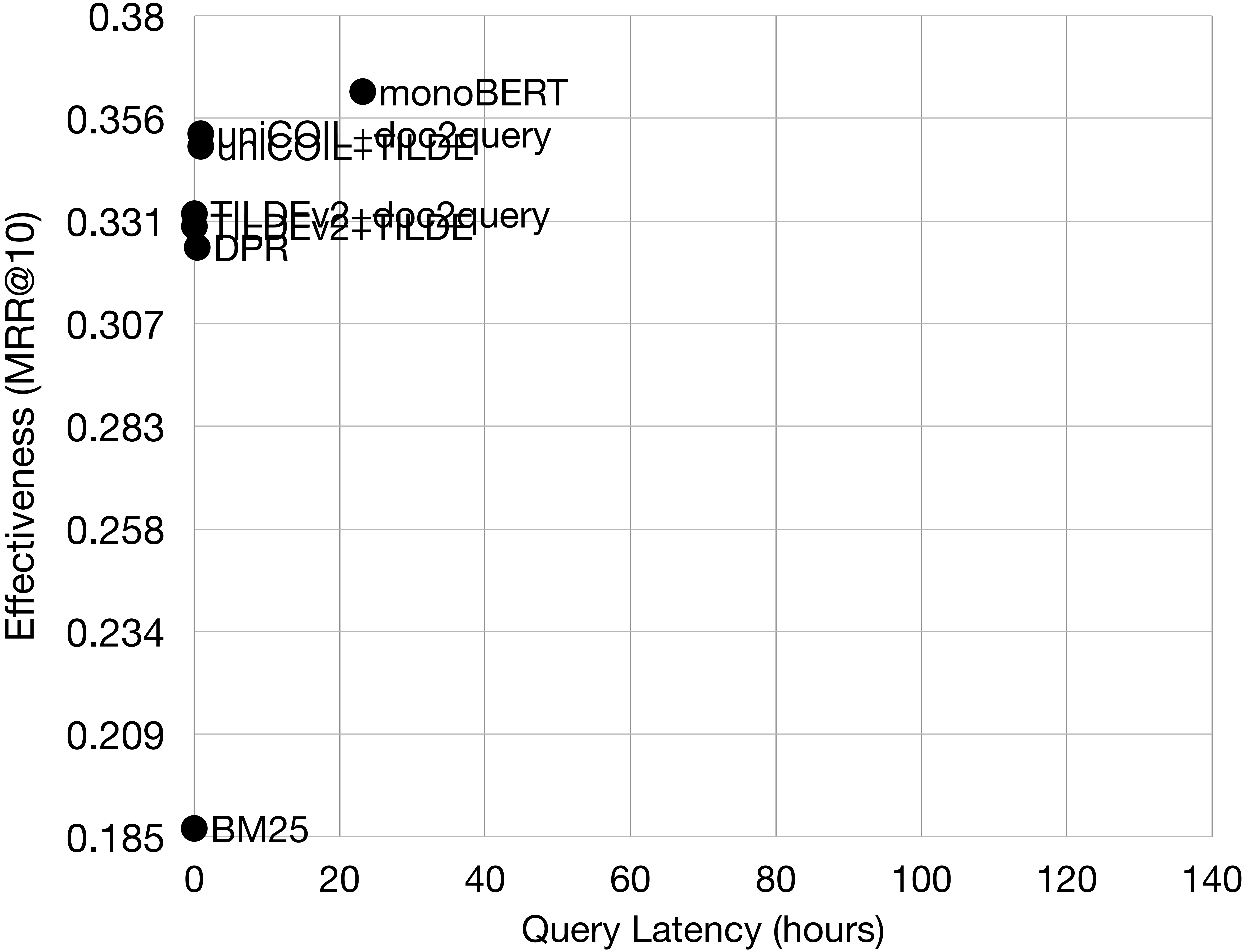
How many emissions do these methods produce to obtain an experimental result?



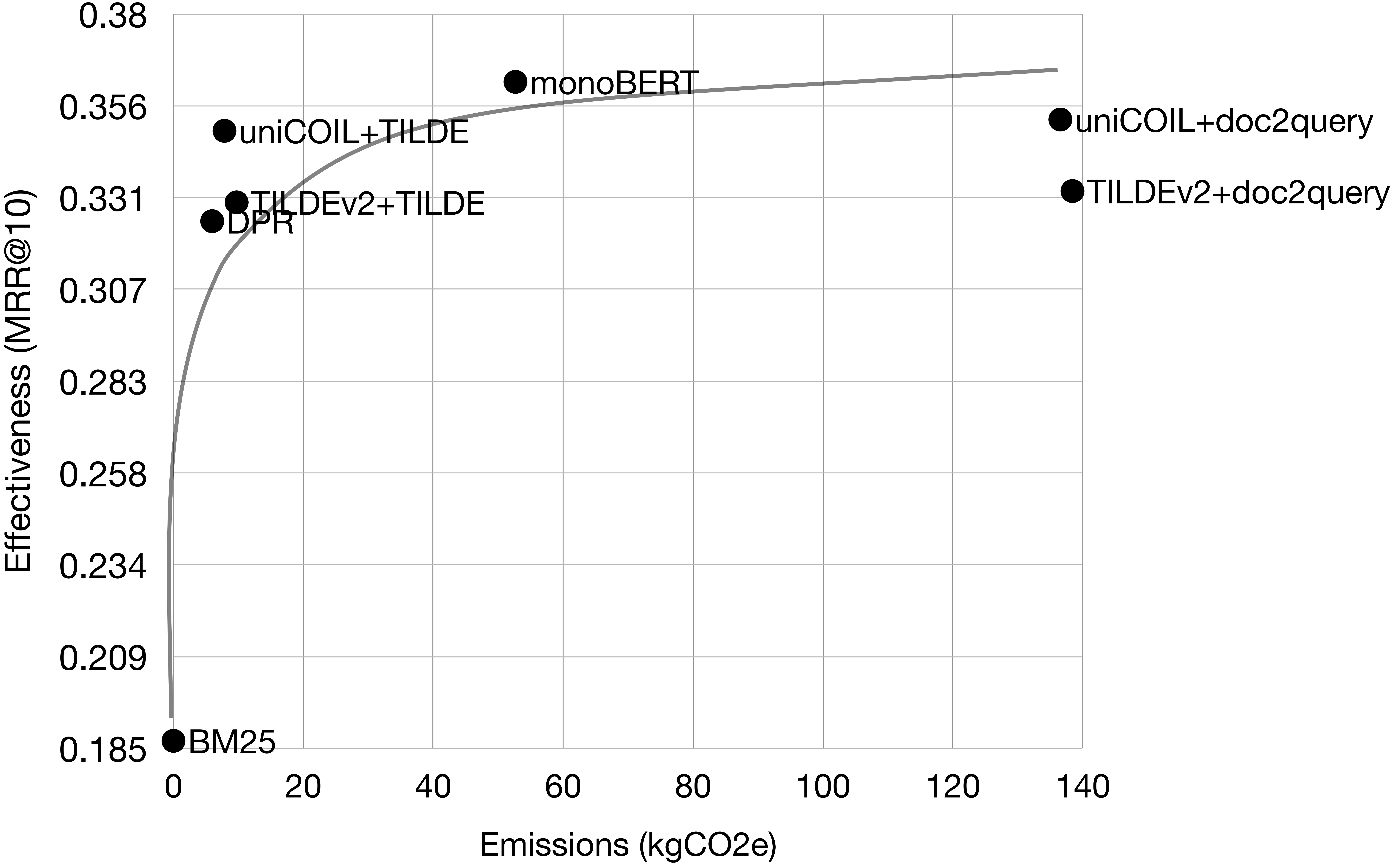
How many emissions do these methods produce to obtain an experimental result?



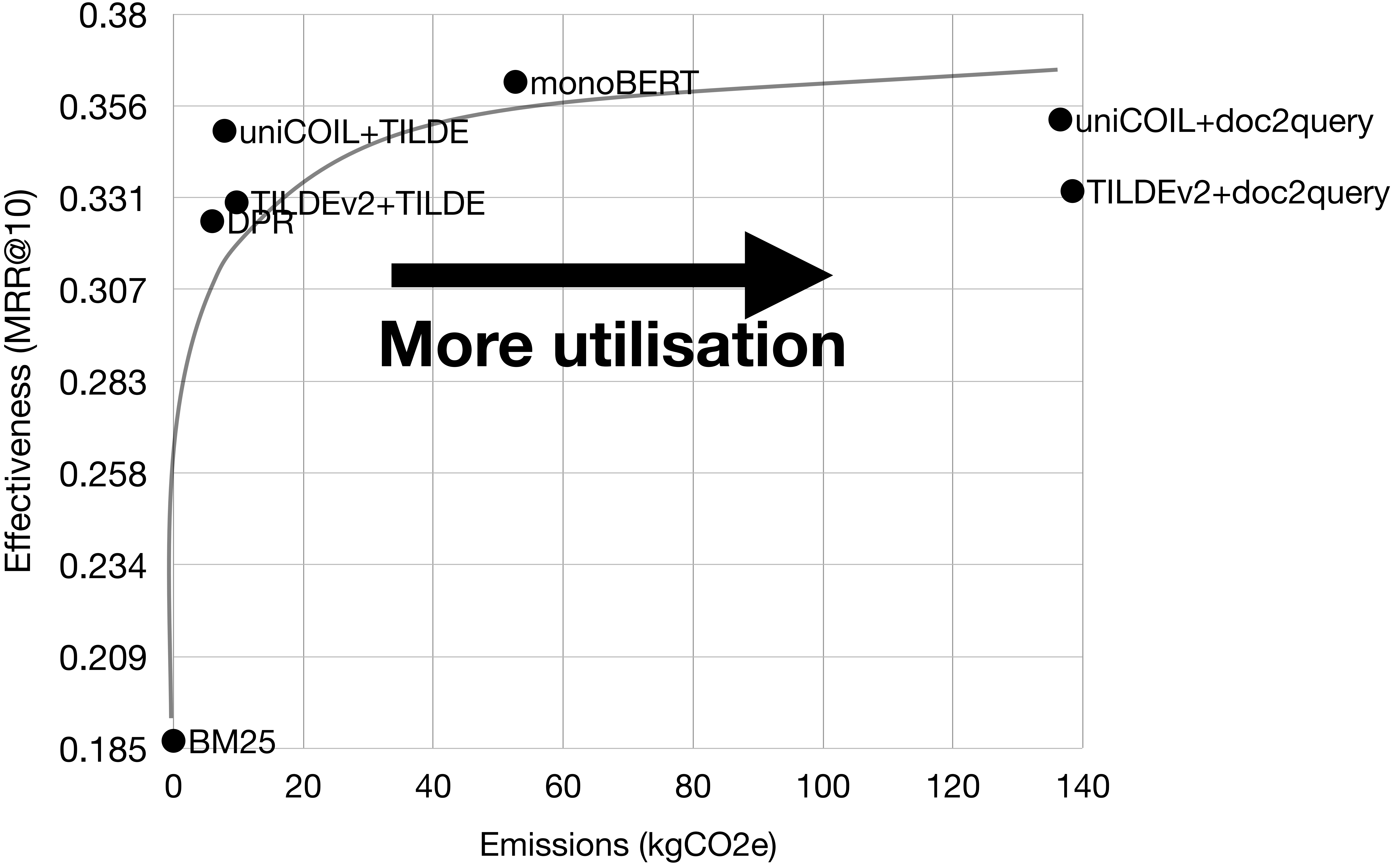
What are the effectiveness-utilisation trade-offs of these methods?



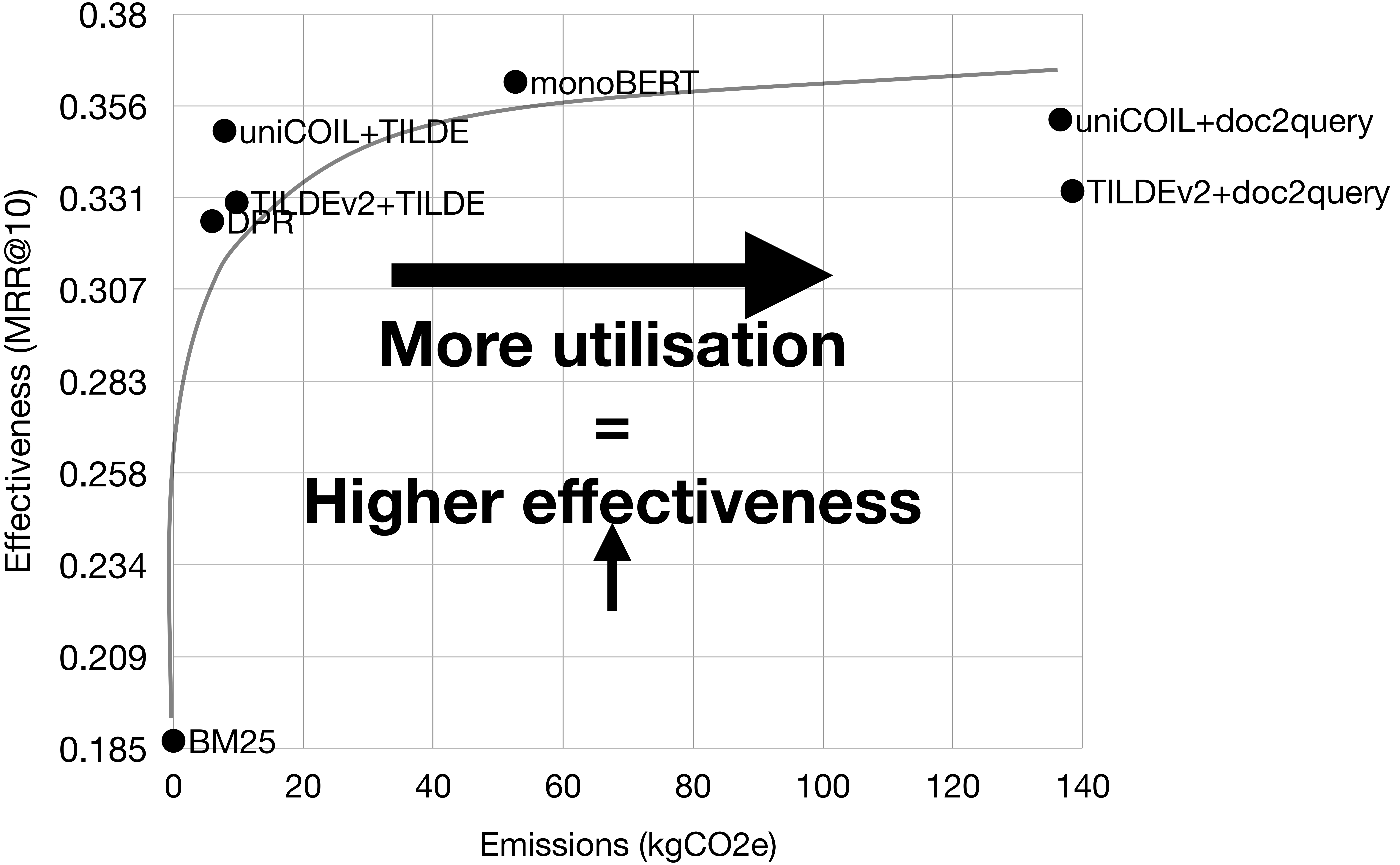
What are the effectiveness-utilisation trade-offs of these methods?



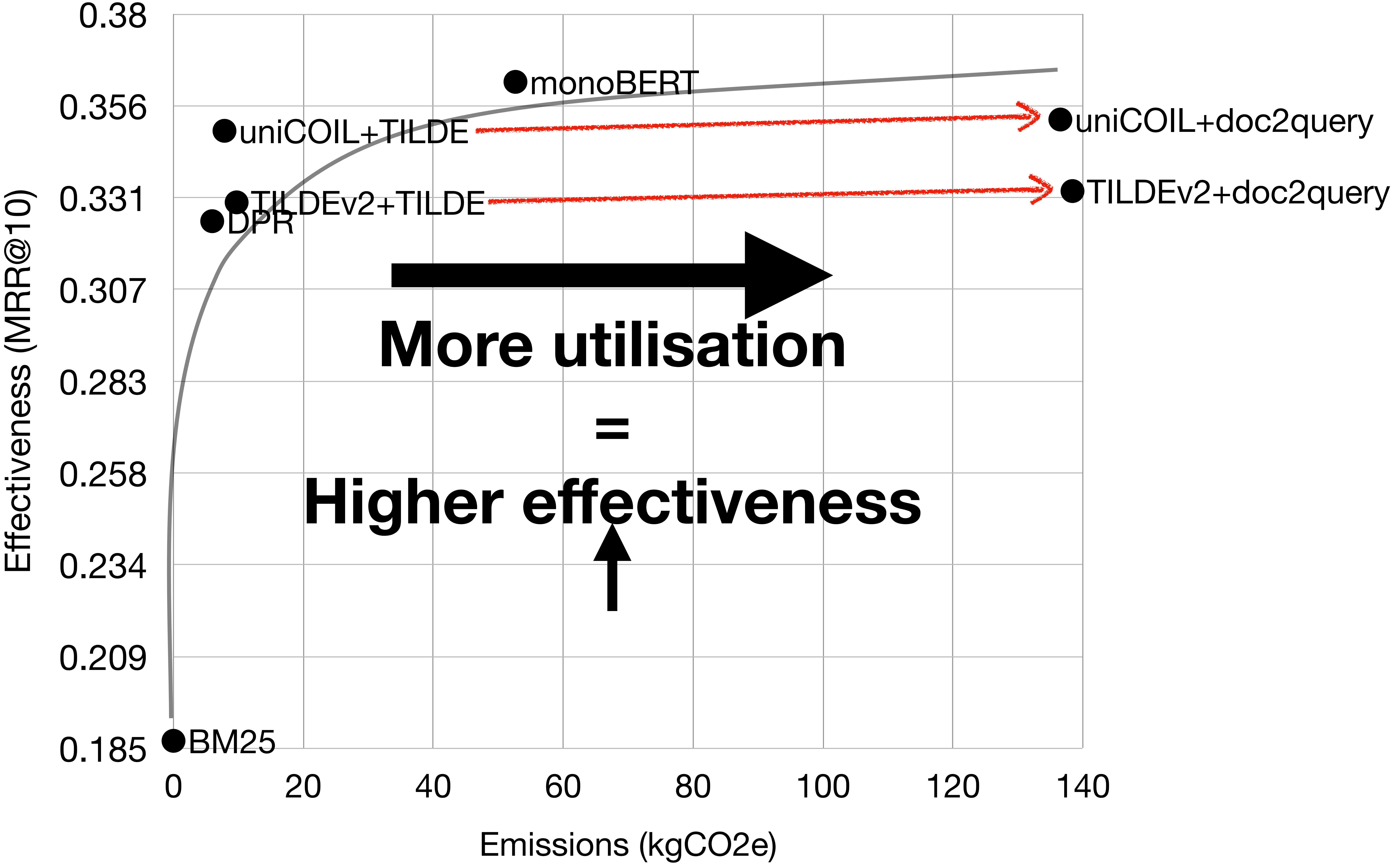
What are the effectiveness-utilisation trade-offs of these methods?



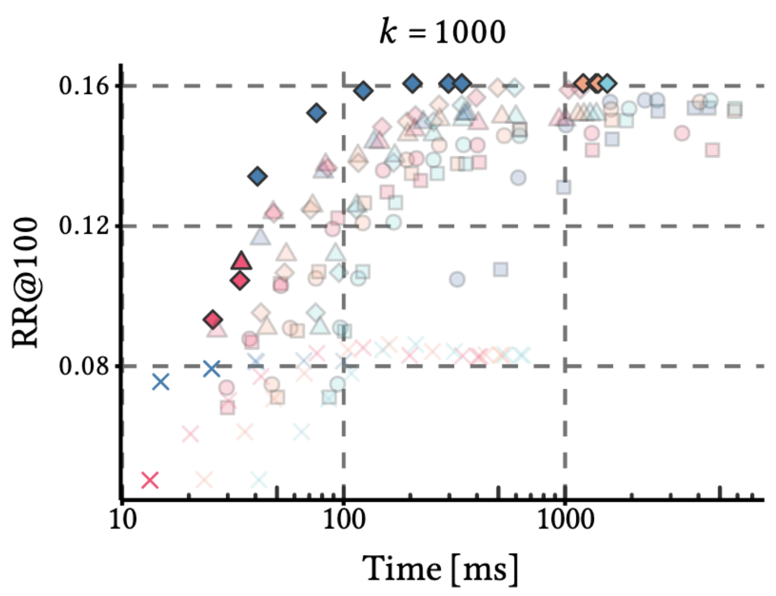
What are the effectiveness-utilisation trade-offs of these methods?

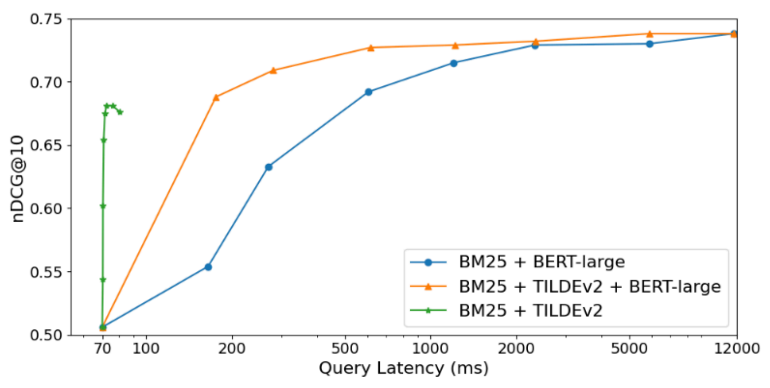


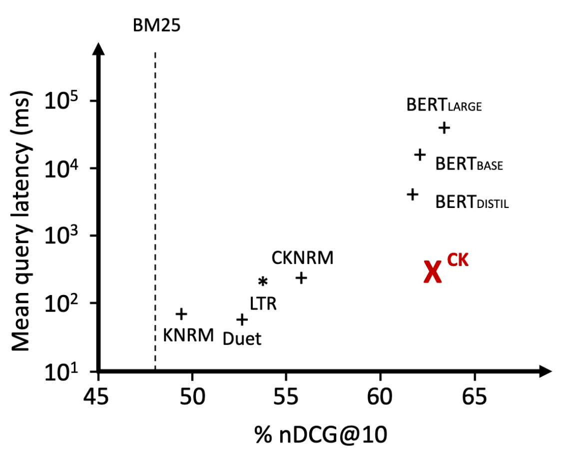
What are the effectiveness-utilisation trade-offs of these methods?

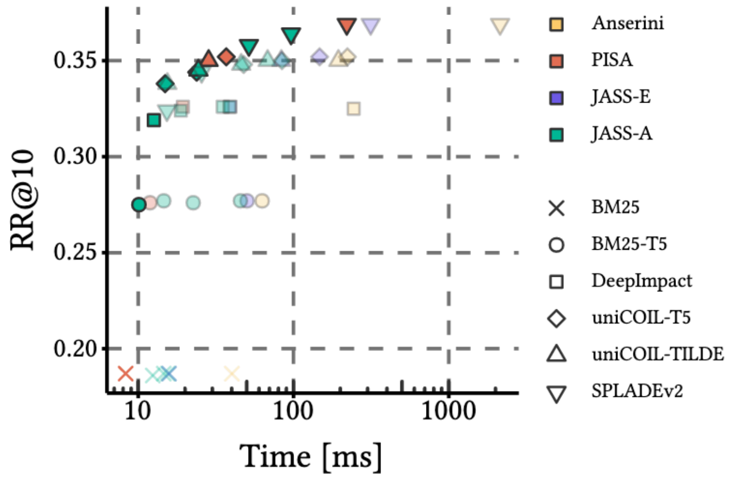


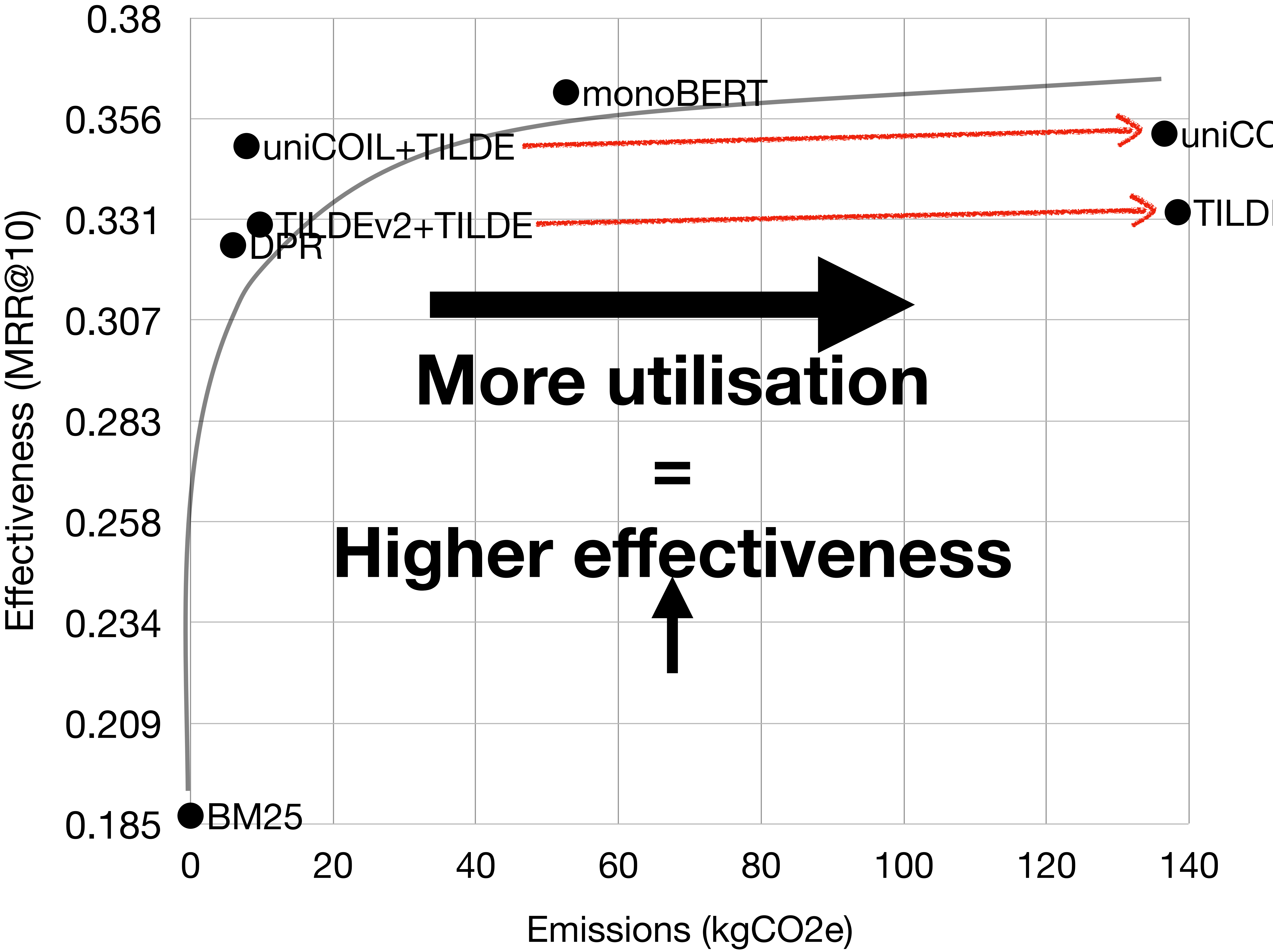
What are the effectiveness-utilisation trade-offs of these methods?

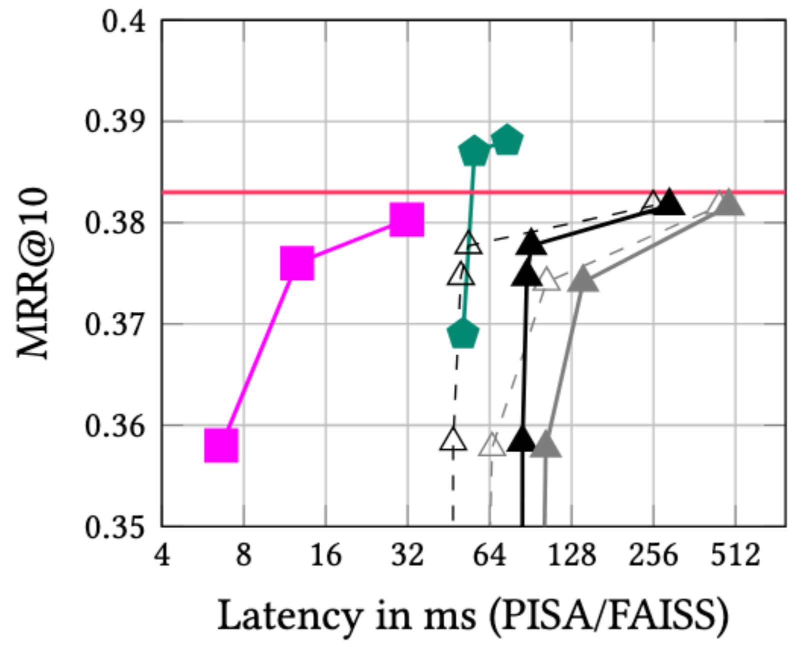


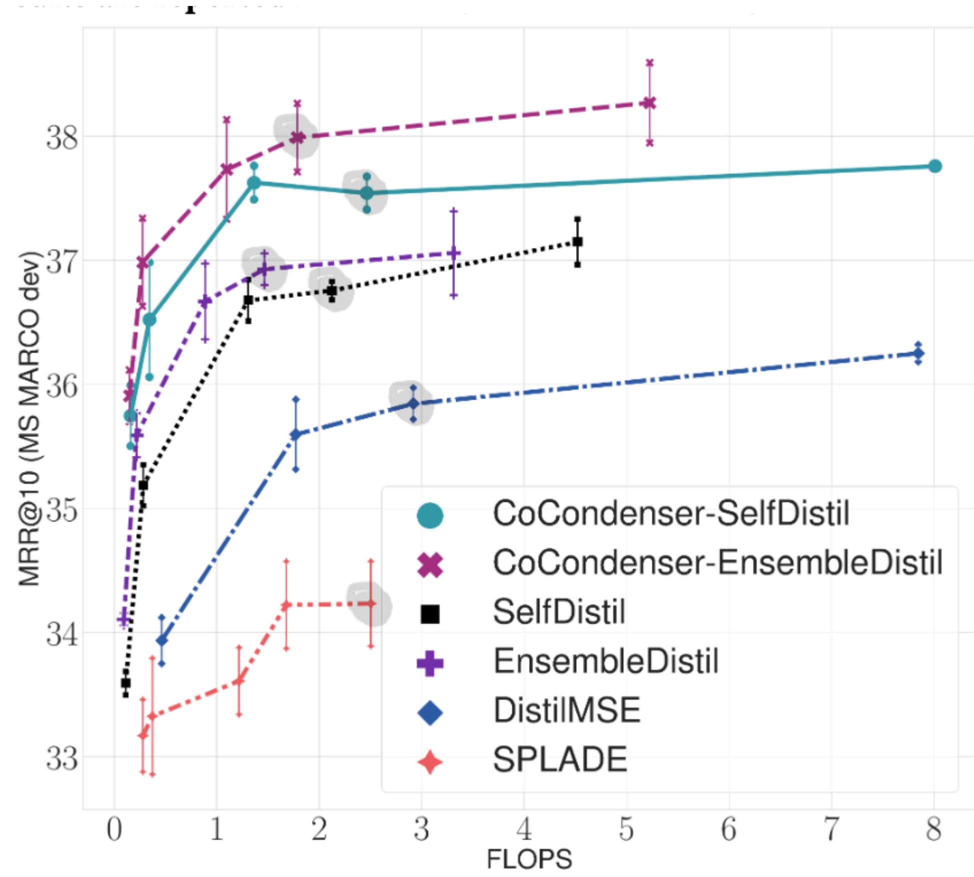










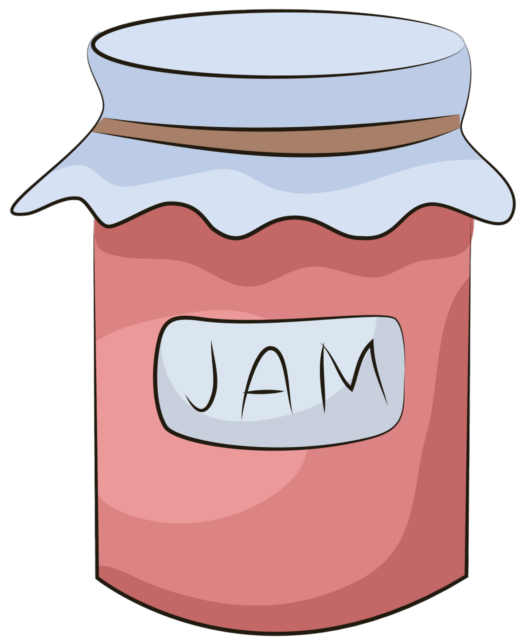


PART II

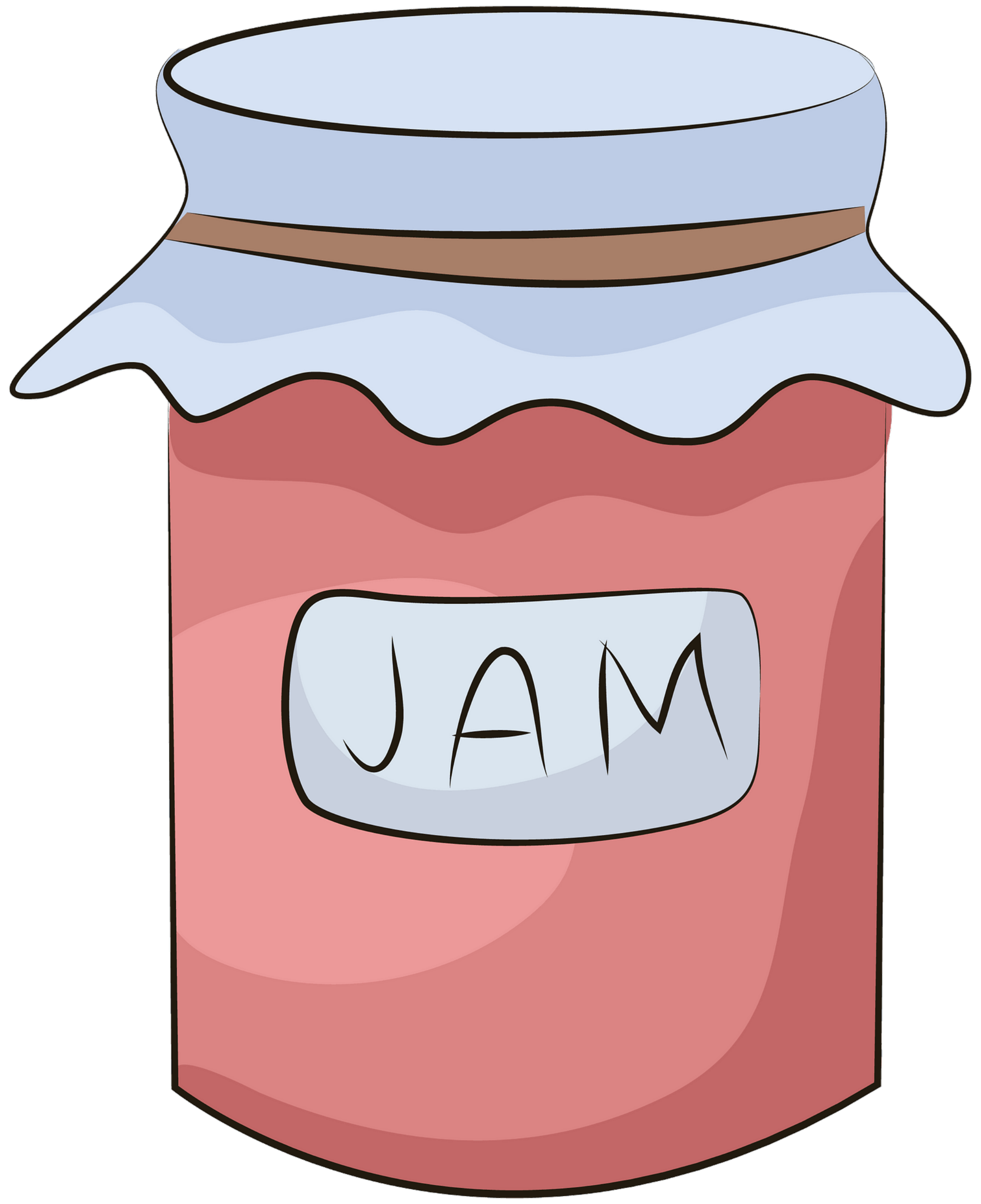
Green IR in Practice

A framework for practitioners to remain mindful of potential costs of IR research

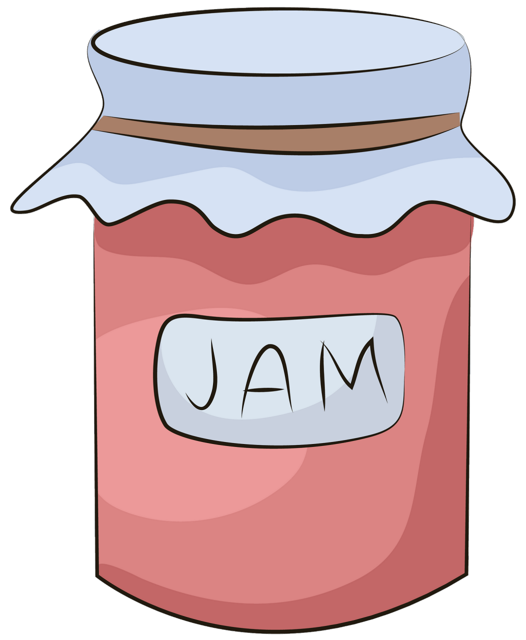
Reduce



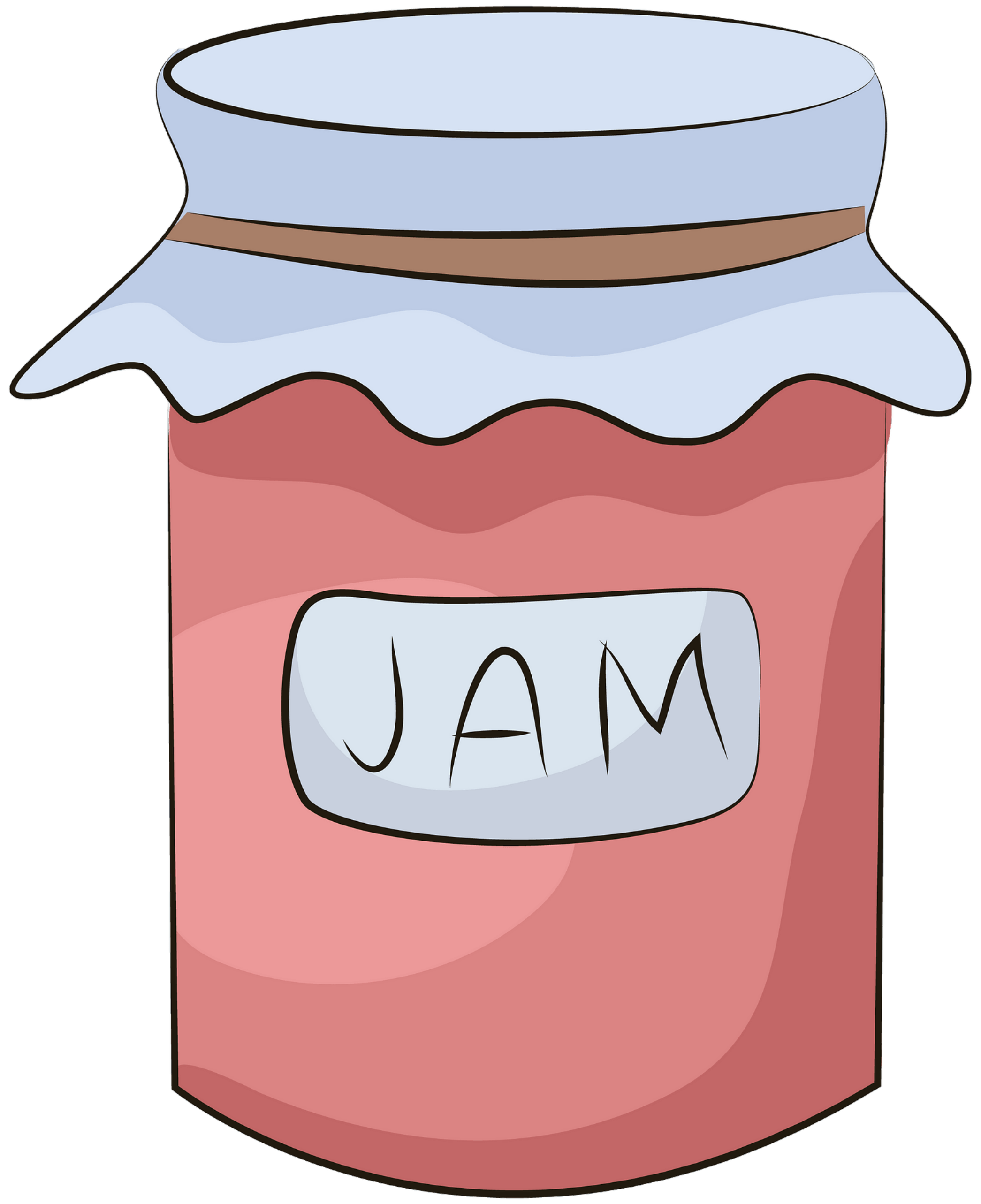
Vs



Reduce



Vs

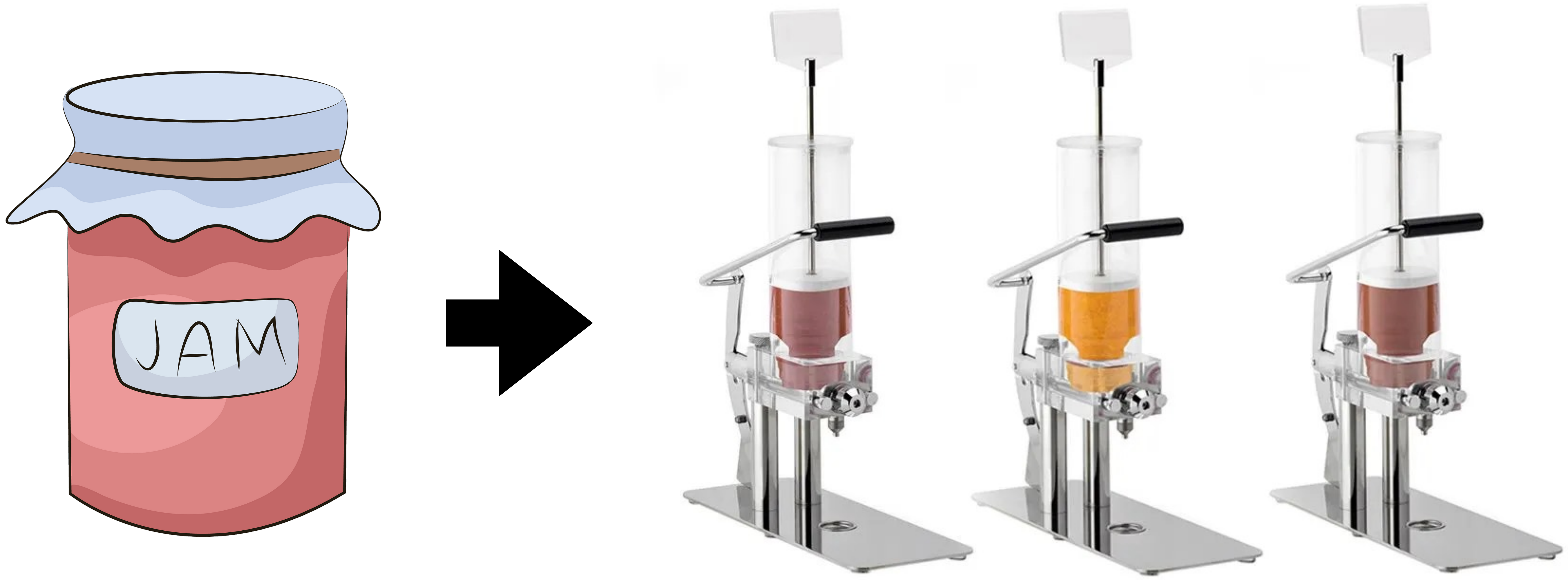


Expend fewer resources

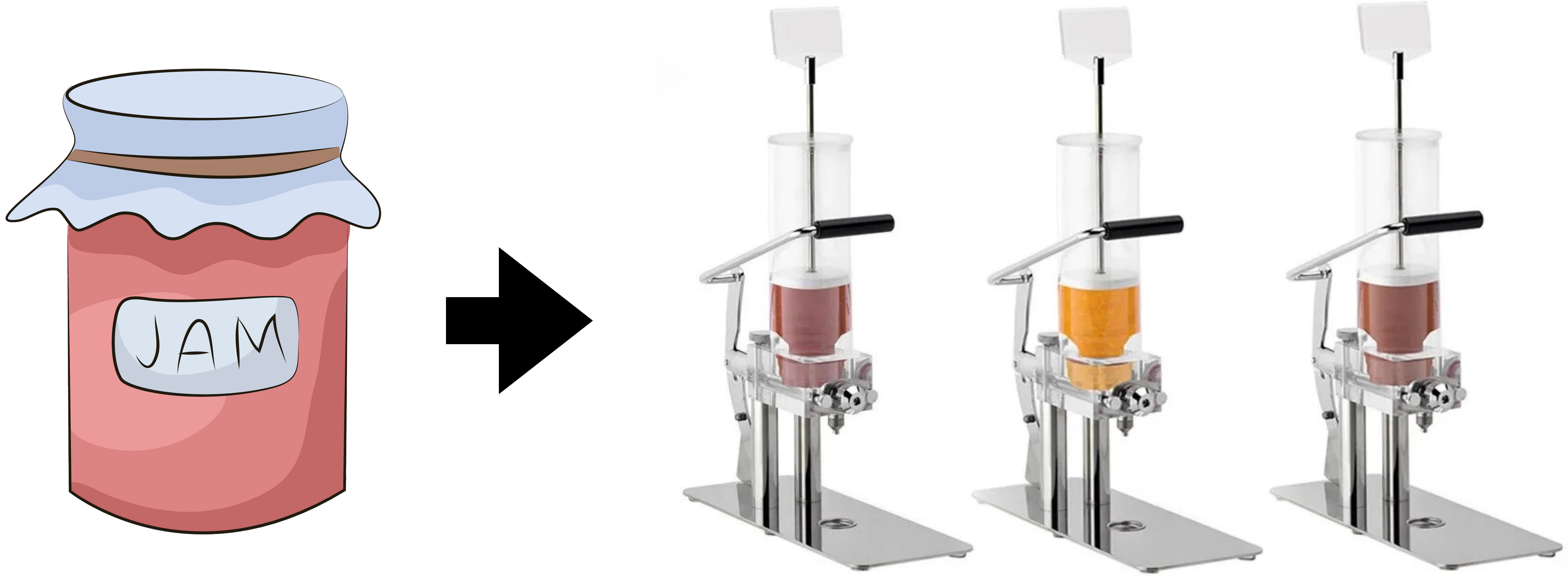
Reduce

- 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?*
    - Random hyper-parameter search
    - CPU-based inference

Reuse



Reuse

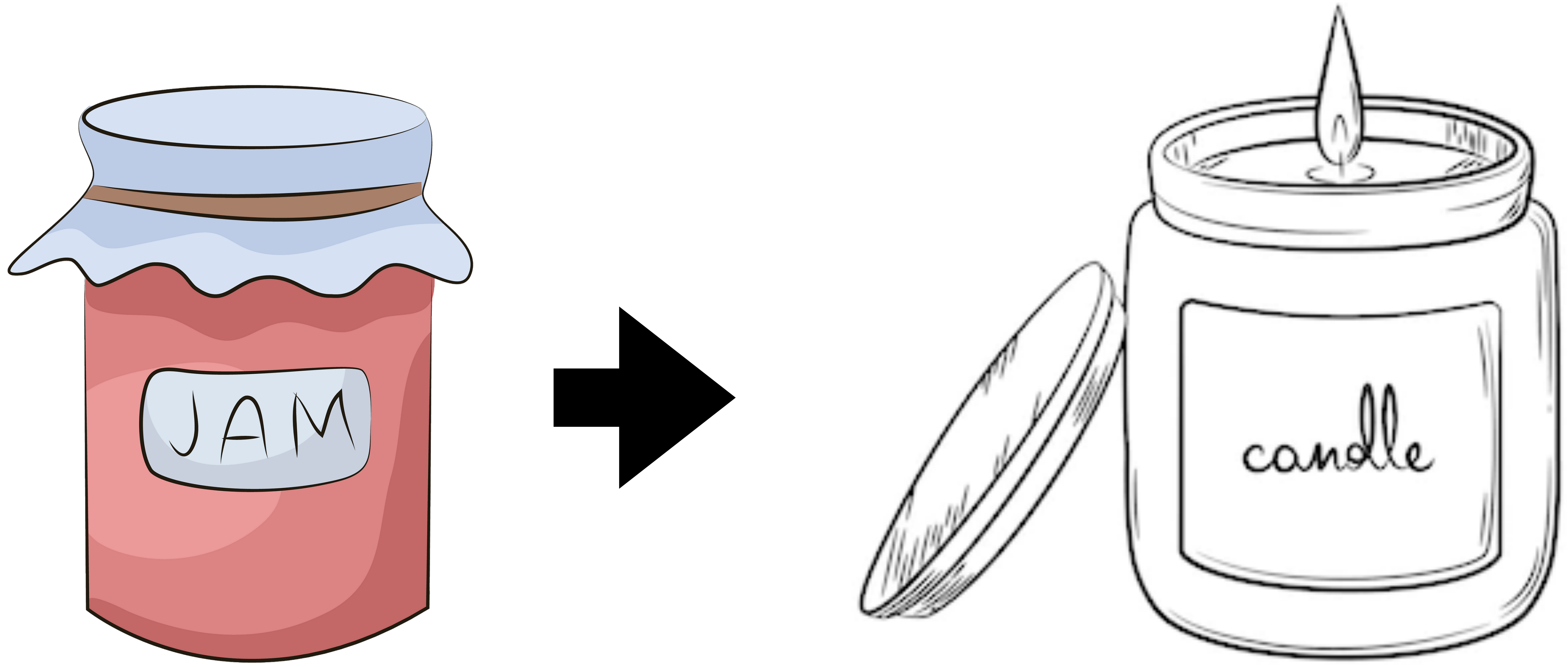


Repurpose resources intended for one task to the same task

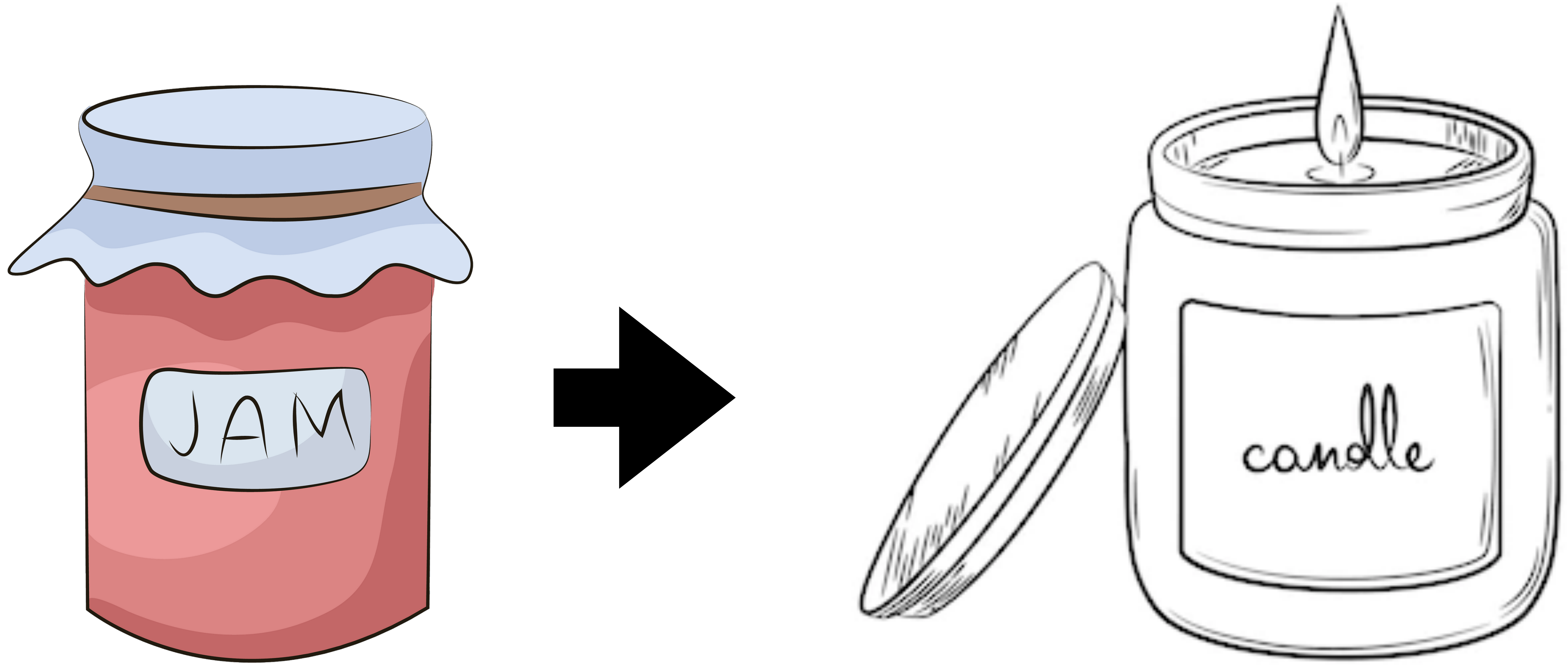
Reuse

- Reuse existing software artefacts such as data, code, or models
- Reuse: 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, code, or other digital artefacts meant for one task to the same task?*
    - Reuse large collections
    - Pre-indexing common collections

Recycle



Recycle



Repurpose resources intended for one task to a different task

Recycle

- Recycle existing software artefacts such as data, code, or models
- Recycle: the action of 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, code, or other digital artefacts meant for one task to a different task?*
    - Neural query expansion
    - Passage expansion with models like TILDE

reduce, reuse, recycle

- Reduce: Expend fewer resources
- Reuse: Repurpose resources intended for one task to the same task
- Recycle: Repurpose resources intended for one task to a different task

PART III

Summary

Efficiency is not just query latency

- There is a trend of “query efficient” neural models which move the heavy computation offline
- This computation still costs: time, hardware, energy, emissions
- It is not just a “once off” cost

Efficiency is not just latency, energy

- Data efficiency
- Learning with little data
- Frugal models, federated learning, few-shot, zero-shot, prompt learning

Summary

- **Larger neural models** = power-hungry hardware = utilisation of more power
  - However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

Summary

- **Larger neural models** = power-hungry hardware = utilisation of more power
  - However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

- Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… **pre-train for IR**
  - More power and more emissions
  - DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model

Summary

- **Larger neural models** = power-hungry hardware = utilisation of more power
  - However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML

- Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… **pre-train for IR**
  - More power and more emissions
  - DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model

- IR community at a **turning point**
  - Bigger/more complex models
  - Bigger collections of documents, queries

Summary

- **Larger neural models** = power-hungry hardware = utilisation of more power
  - However: increased model size for higher effectiveness may not apply to IR, as it does to NLP and ML
- Likely trend in neural IR: go beyond PLMs designed for NLP but are specialised for IR… **pre-train for IR**
  - More power and more emissions
  - DSI: end-to-end transformers that encapsulate the entire indexing and searching architecture into a single model
- IR community at a **turning point**
  - Bigger/more complex models
  - Bigger collections of documents, queries
- There is a cost to IR (+NLP, ML) research:
  - Power usage: $$$
  - Emissions: CO_2e