Reduce, Reuse, Recycle: Green Information Retrieval Research

Harry Scells, Shengyao Zhuang, Guido Zuccon

h.scells@uq.edu.au

The University of Queensland, Australia

[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*

[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

ie
lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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

ie lab

The University Of Queensland Australia

Create Change

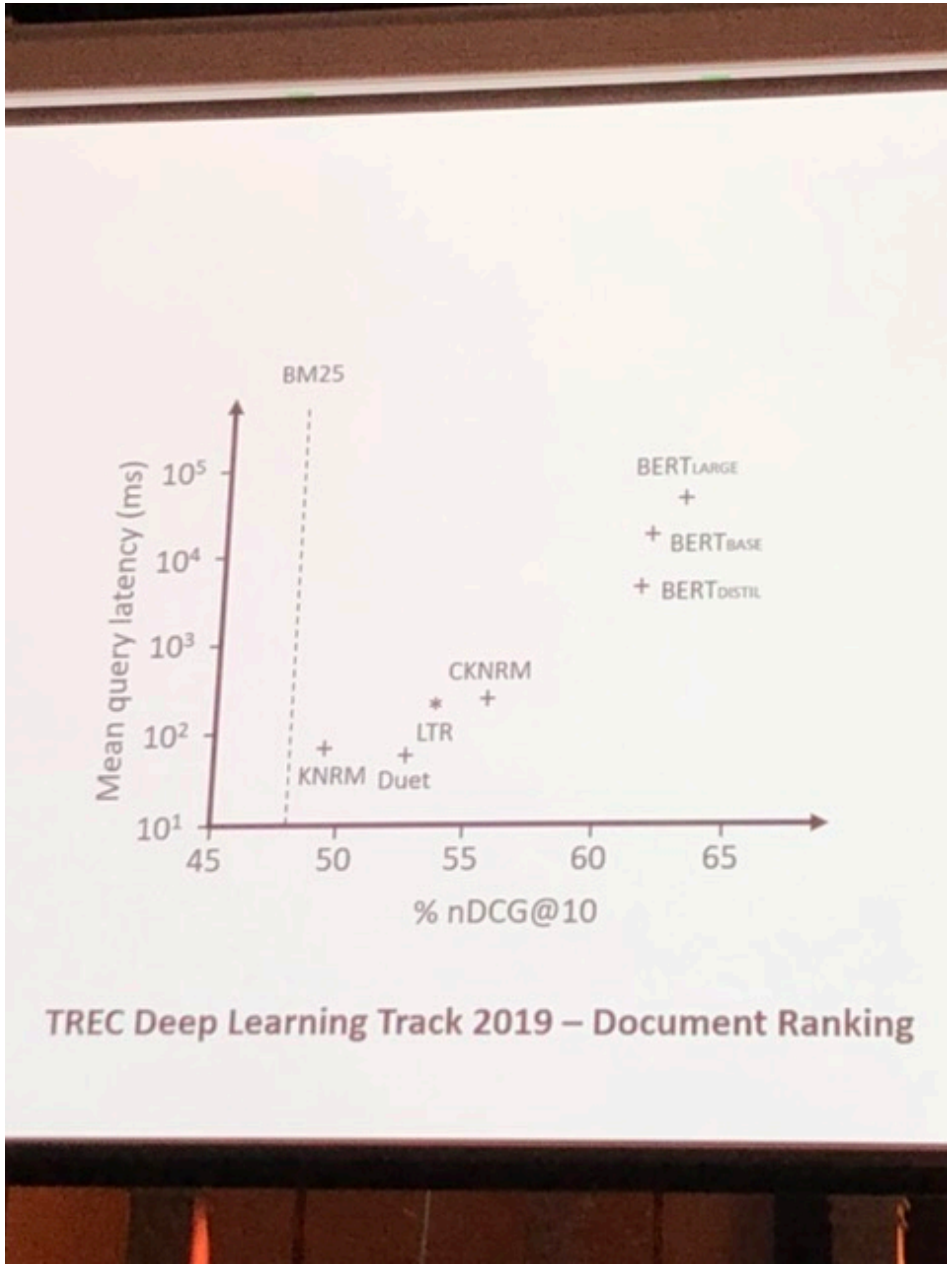
NLP

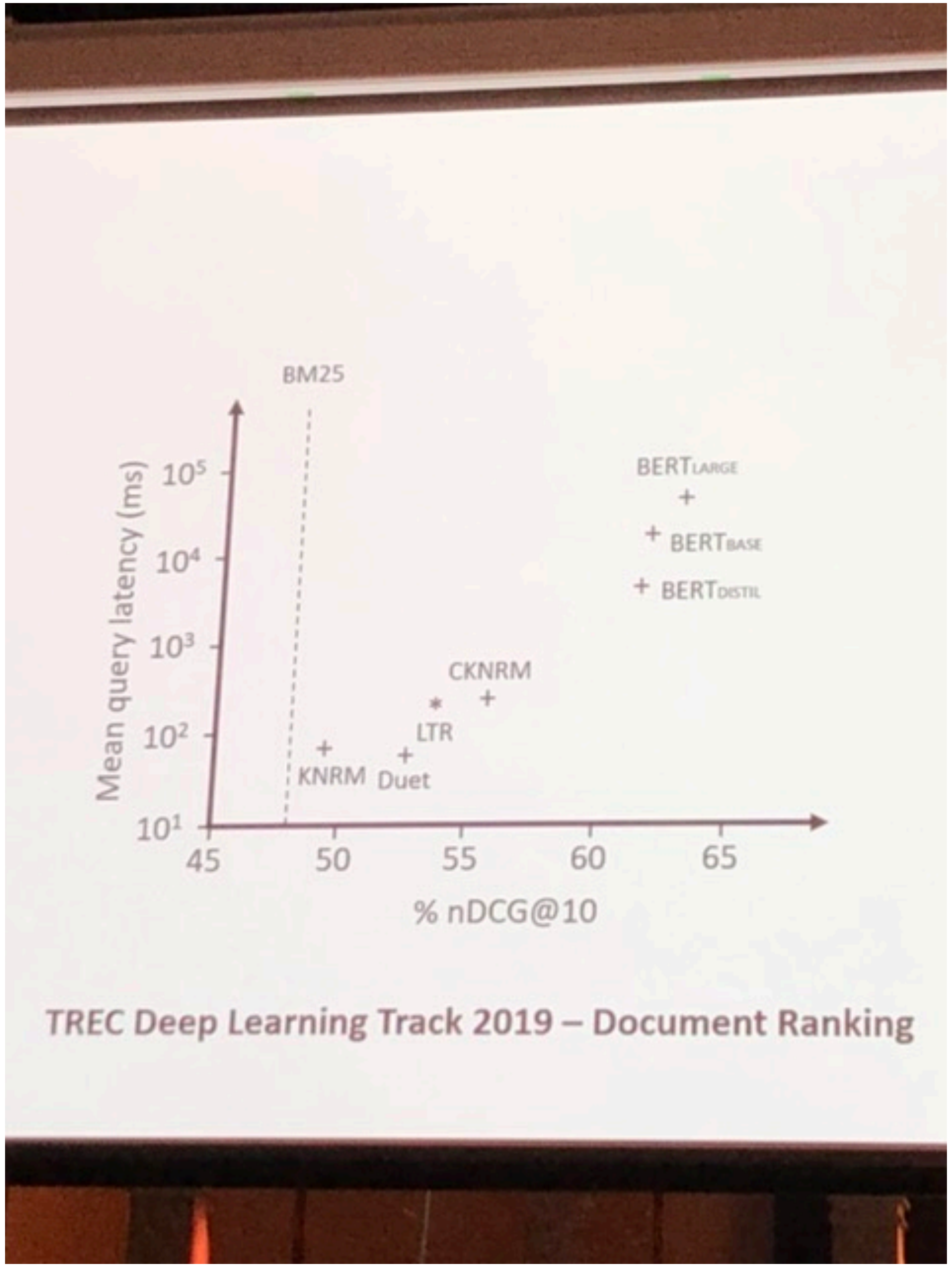
ML

What about IR research?

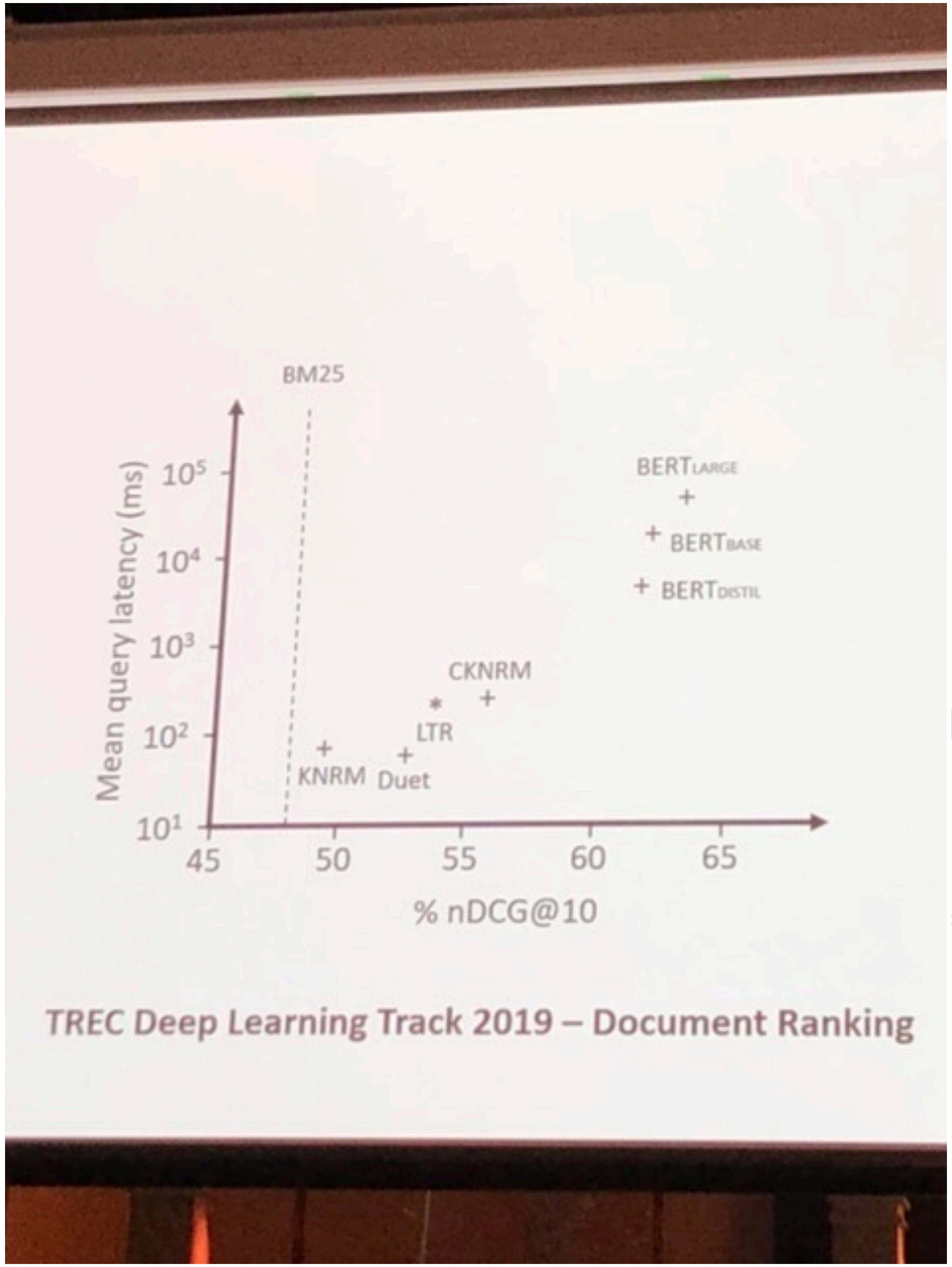
Isn't this just retrieval efficiency?

[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*


BM25
Mean query latency (ms)
10^5
10^4
10^3
10^2
10^1
BERTLARGE
BERTBASE
BERTDISTIL
CKNRM
LTR
KNRM
Duet
45
50
55
60
65
% nDCG@10
TREC Deep Learning Track 2019 – Document Ranking

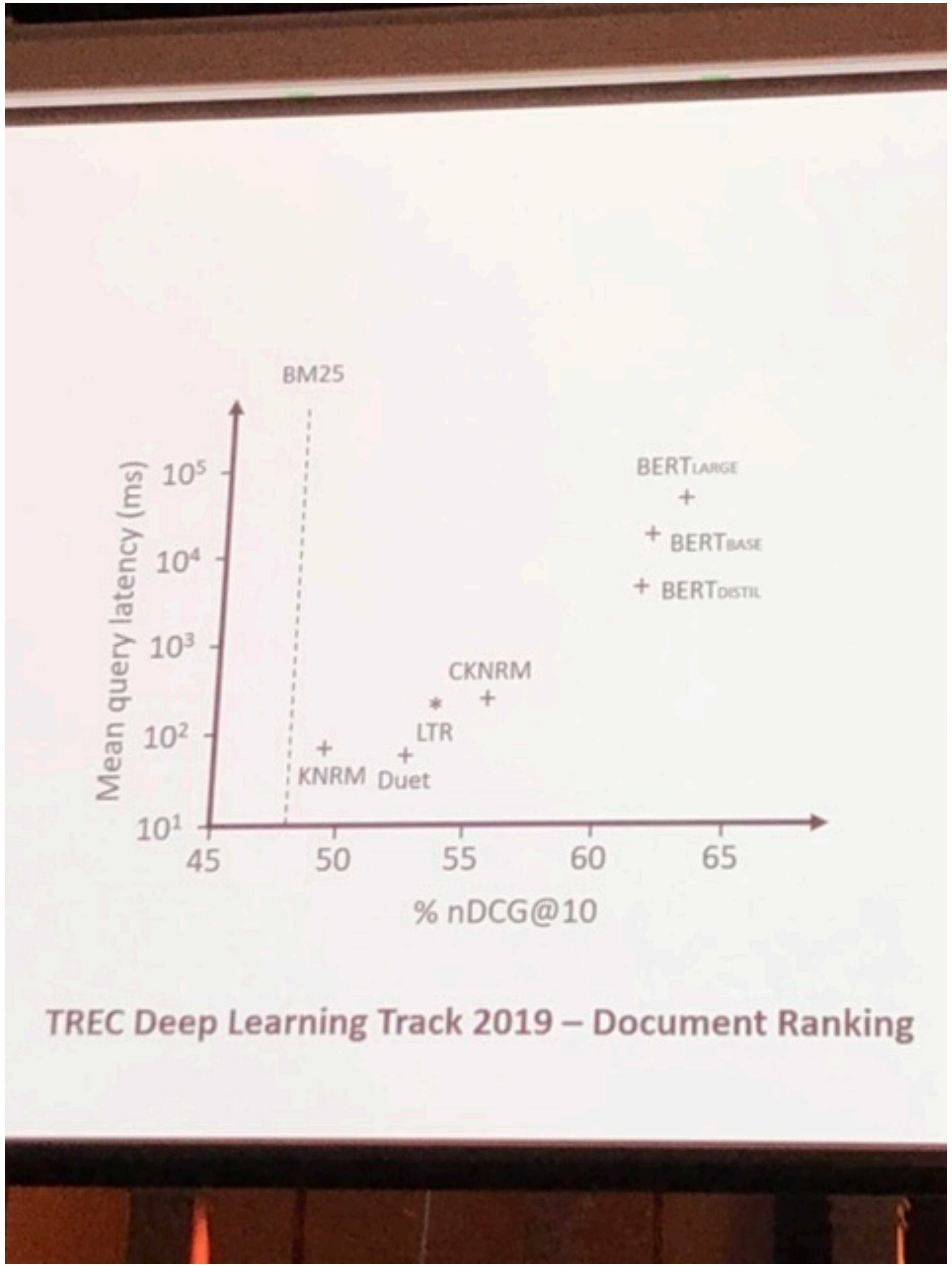

BM25
Mean query latency (ms)
10^5
10^4
10^3
10^2
10^1
BERTLARGE
BERTBASE
BERTDISTIL
CKNRM
LTR
KNRM
Duet
45
50
55
60
65
% nDCG@10
TREC Deep Learning Track 2019 – Document Ranking


RR@10
0.35
0.30
0.25
0.20
PISA
JASS-A
Time [ms]
PISA
JASS-A


BM25
Mean query latency (ms)
10⁵
10⁴
10³
10²
10¹
BERTLARGE
BERTBASE
BERTDISTIL
CKNRM
LTR
KNRM
Duet
45
50
55
60
65
% nDCG@10
TREC Deep Learning Track 2019 – Document Ranking

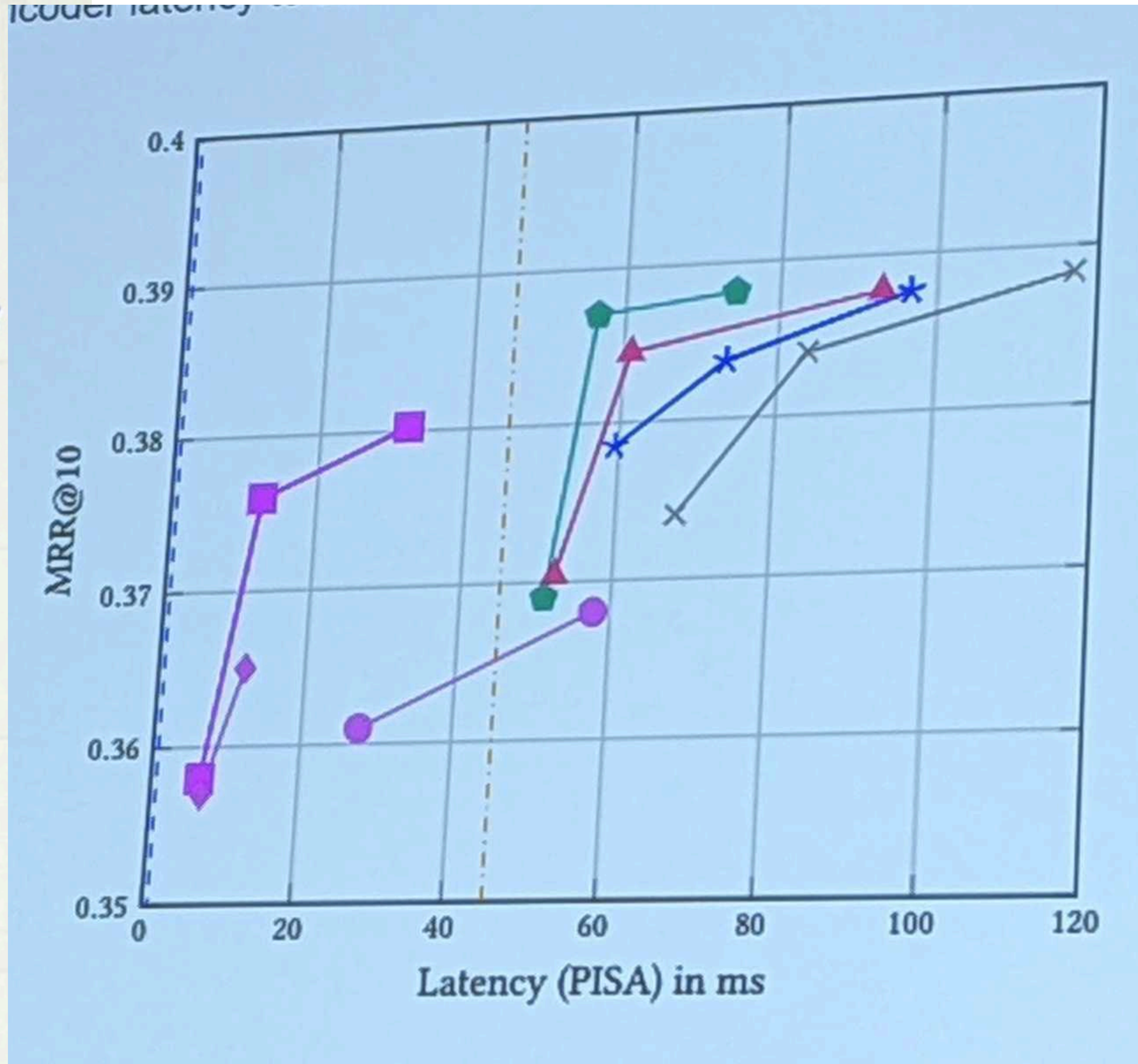

RR@10
PISA
JASS-A
Time [ms]
PISA
JASS-A


SelfDistil
EnsembleDistil
DistilMSE
SPLADE
MRR@10 (MS MARCO dev)
38
37
36
35
34
33
0
1
2
3
4
5
6
7
8
FLOPS


BM25
Mean query latency (ms)
10^1
10^2
10^3
10^4
10^5
BERT LARGE
BERT BASE
BERT DISTIL
CKNRM
LTR
KNRM
Duet
45
50
55
60
65
% nDCG@10
TREC Deep Learning Track 2019 – Document Ranking


RR@10
PISA
JASS-A
Time [ms]
PISA
JASS-A


SelfDistil
EnsembleDistil
DistilMSE
SPLADE
MRR@10 (MS MARCO dev)
33
34
35
36
37
38
0
1
2
3
4
5
6
7
FLOPS


MRR@10
0.35
0.36
0.37
0.38
0.39
0.4
0
20
40
60
80
100
120
Latency (PISA) in ms

Developing Energy Efficient Filtering Systems

Leif Azzopardi, Wim Vanderbauwhede, Mahmoud Moadeli
Dept. of Comp. Sci., University of Glasgow
Glasgow, United Kingdom
{leif, wim, mahmoudm}@dcs.gla.ac.uk

ABSTRACT

Processing large volumes of information generally requires massive amounts of computational power, which consumes a significant amount of energy. An emerging challenge is the development of "environmentally friendly" systems that are not only efficient in terms of time, but also energy efficient. In this poster, we outline our initial efforts at developing greener filtering systems by employing Field Programmable Gate Arrays (FPGA) to perform the core information processing task. FPGAs enable code to be executed in parallel at a chip level, while consuming only a fraction of the power of a standard (von Neuman style) processor. On a number of test collections, we demonstrate that the FPGA filtering system performs 10-20 times faster than the Itanium based implementation, resulting in considerable energy savings.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software: Performance evaluation

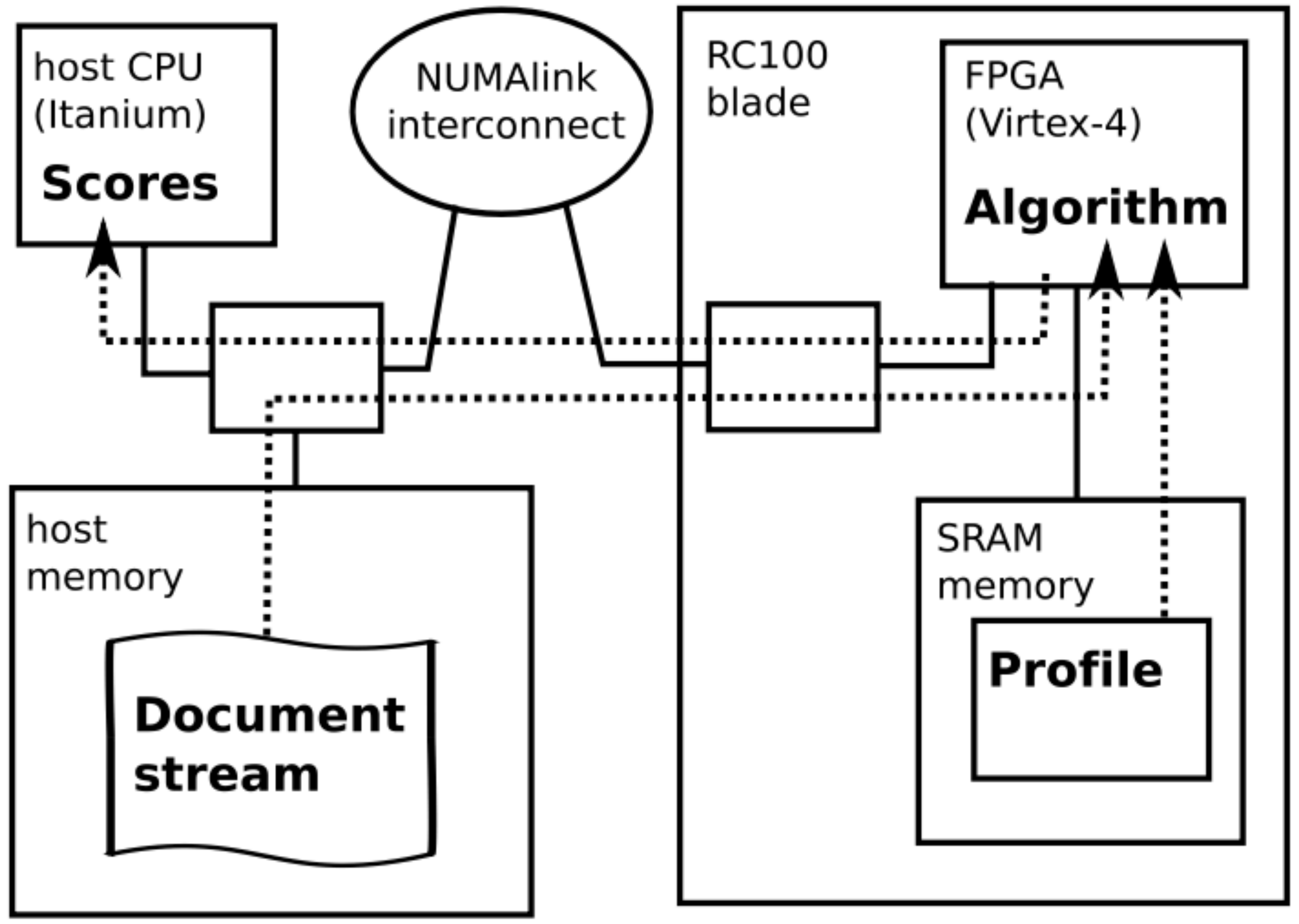
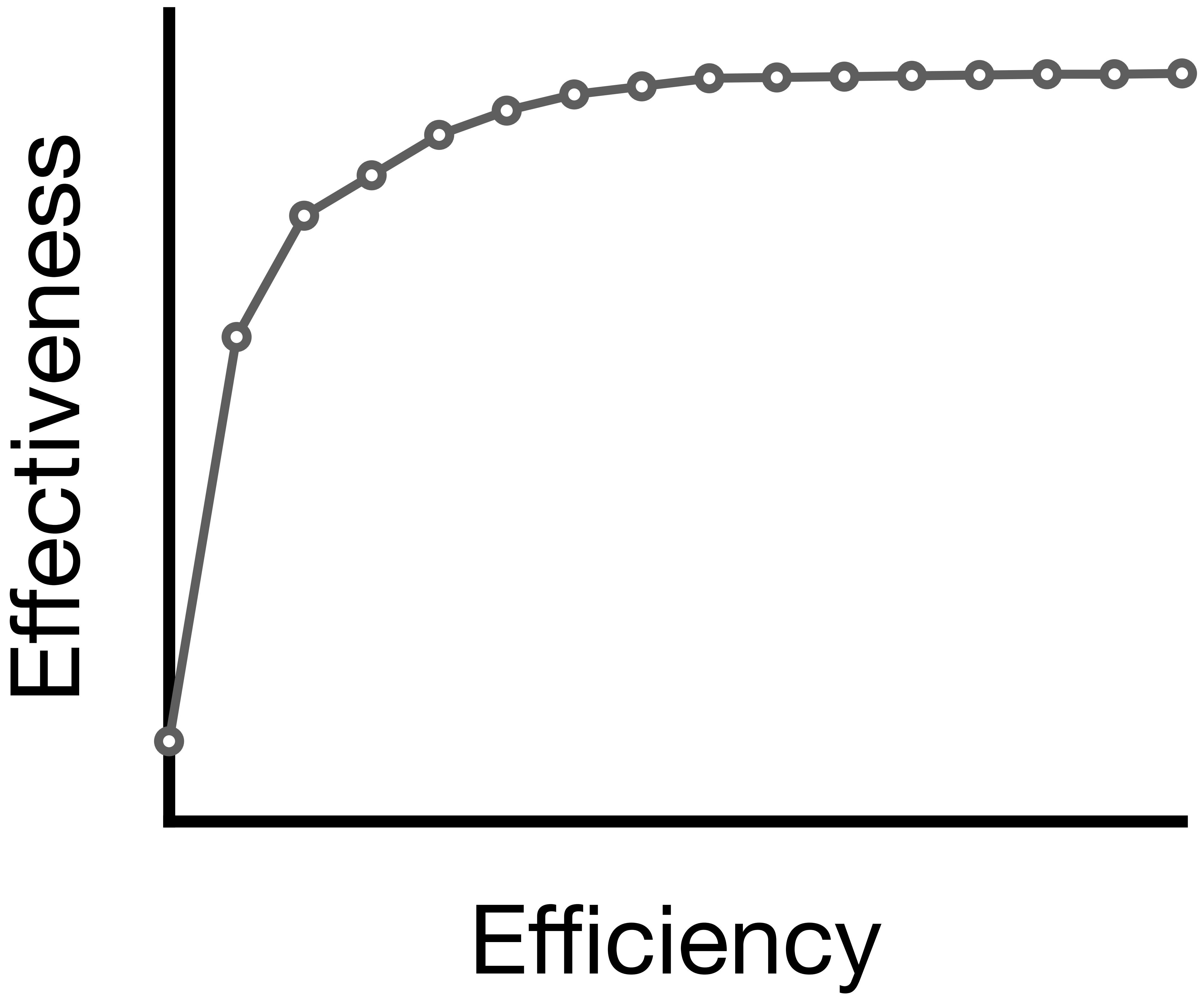


Figure 1: Schematic of FPGA-accelerated filtering

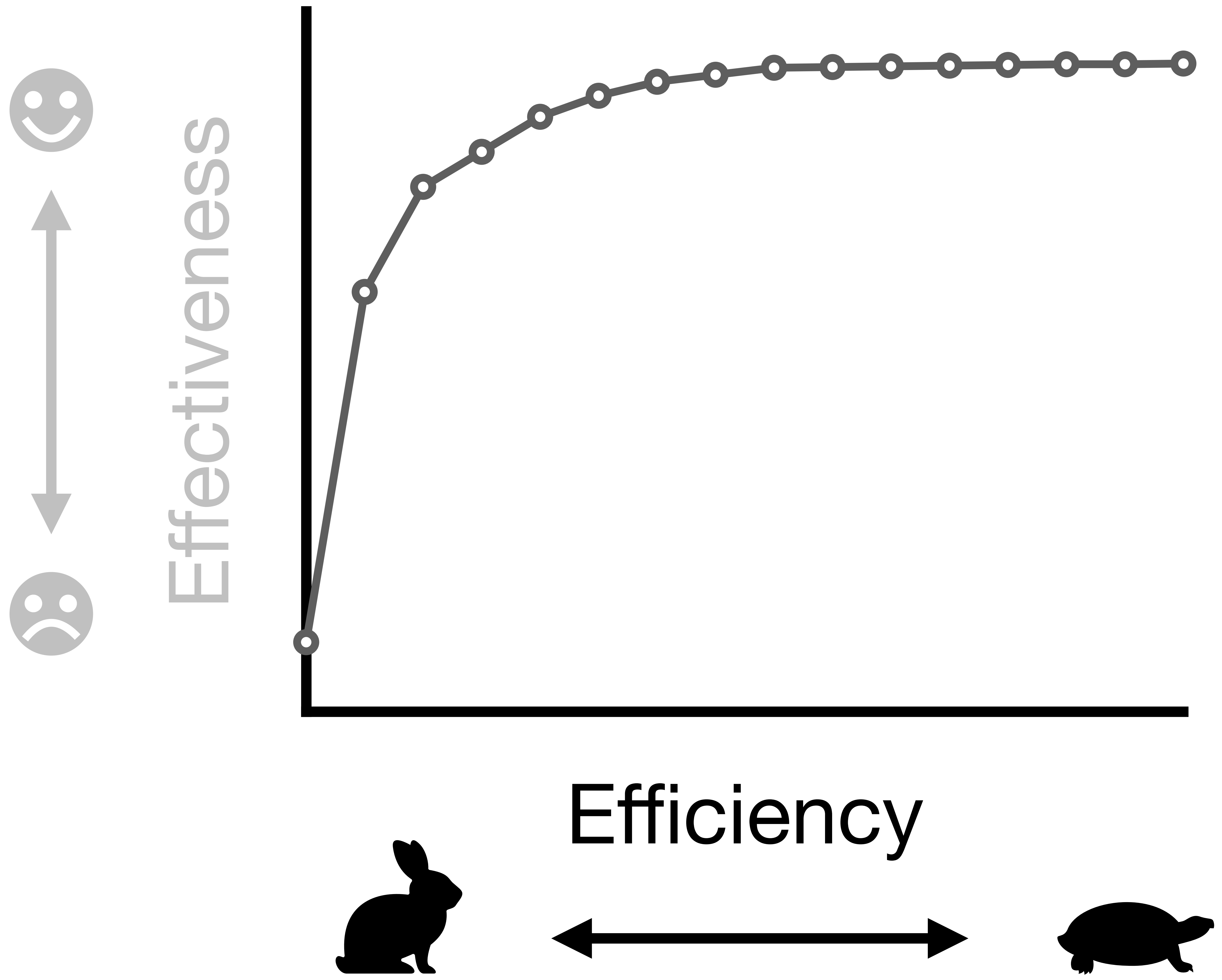
2. SYSTEM ARCHITECTURE

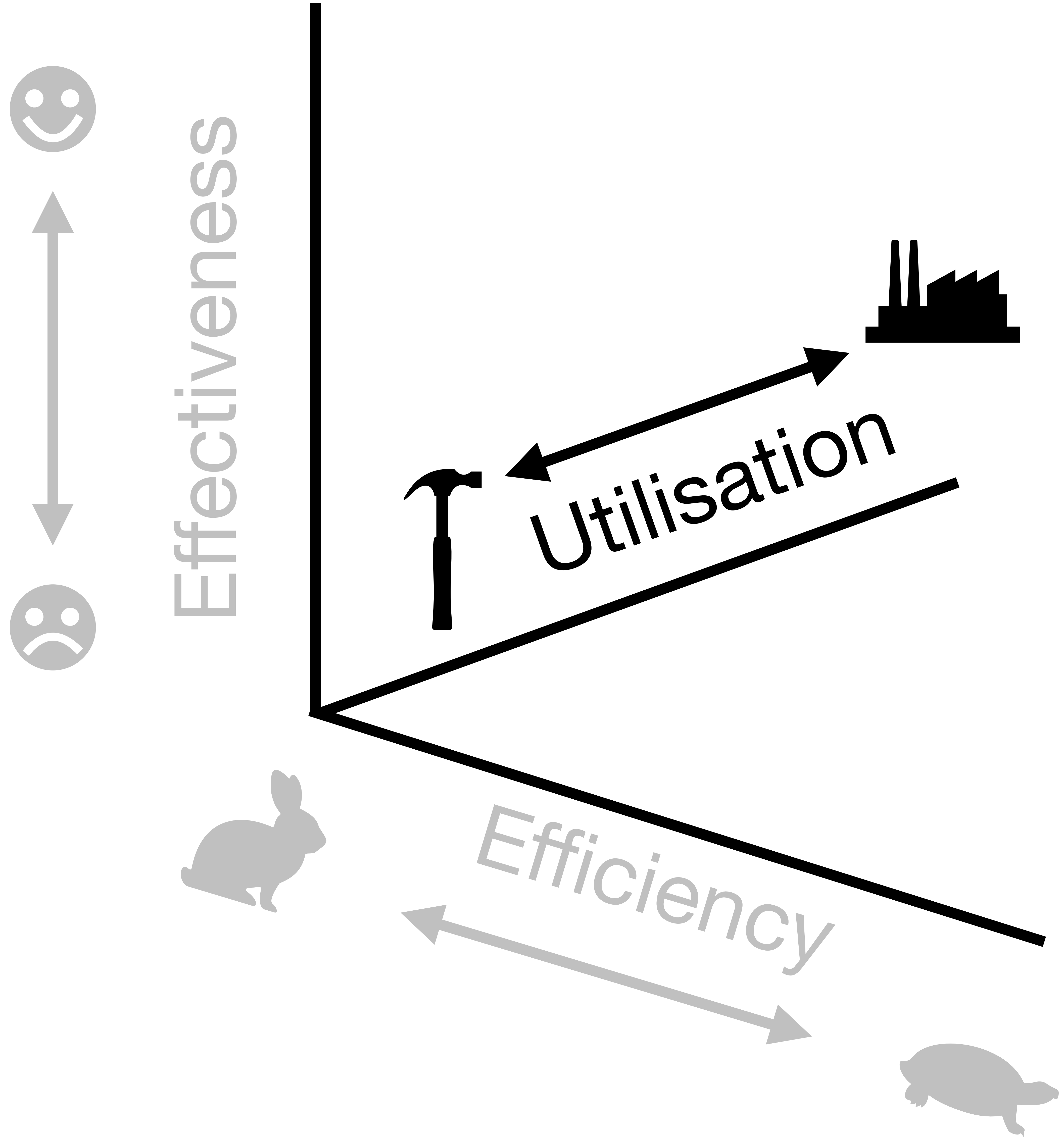
An FPGA is a reconfigurable semiconductor device which


Effectiveness
Efficiency

ie
lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE


Effectiveness
Efficiency



ie
lab

THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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:

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

Measuring emissions

- First, measure power consumption:

PUE

watts

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

ie
lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

Measuring emissions

- First, measure power consumption:

Running Time

PUE

watts

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

Measuring emissions

- First, measure power consumption:

PUE

Running Time

CPU, RAM, GPU power draw

watts

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

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:

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:

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

ie lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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:

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

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

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

ie lab

The University of Queensland Australia

Create Change

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

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

Power consumption of experiments

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

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:

$$\underbrace{\mathbf{kgCO_2e}}_{\text{emissions}} = \overbrace{\theta}^{\text{avg. } CO_2e \text{ (kg) per kWh where experiments took place}} \cdot \underbrace{p_t}_{\text{Power consumption of experiments}}$$

- Emissions of my search engine:

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

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:

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

Power consumption of a single query

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

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

Power consumption of experiments

- Emissions of my search engine:

No. queries issued per unit time

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

Power consumption of a single query

ie lab

The University of Queensland Australia

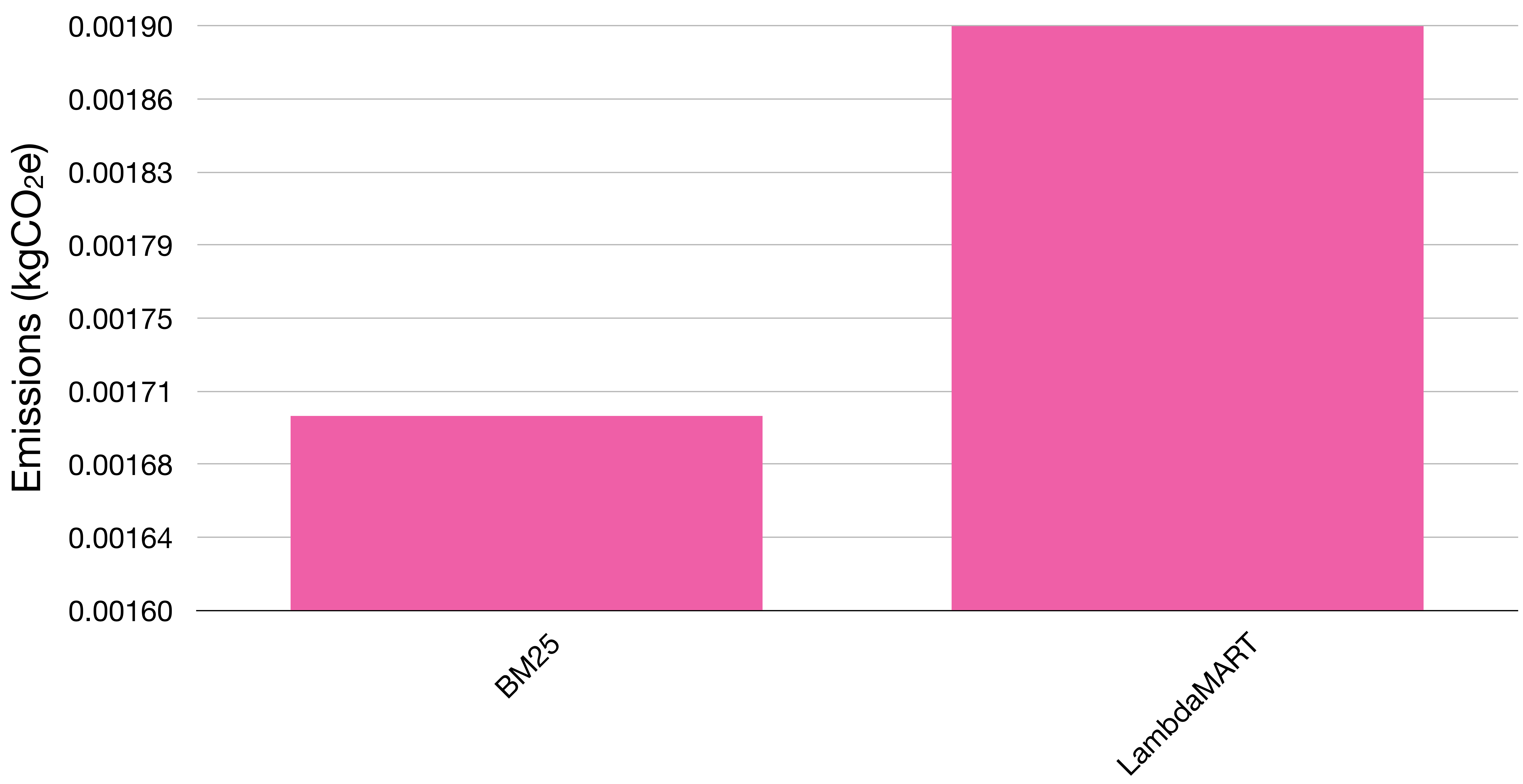
Create Change

Measuring energy & emissions of your model

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

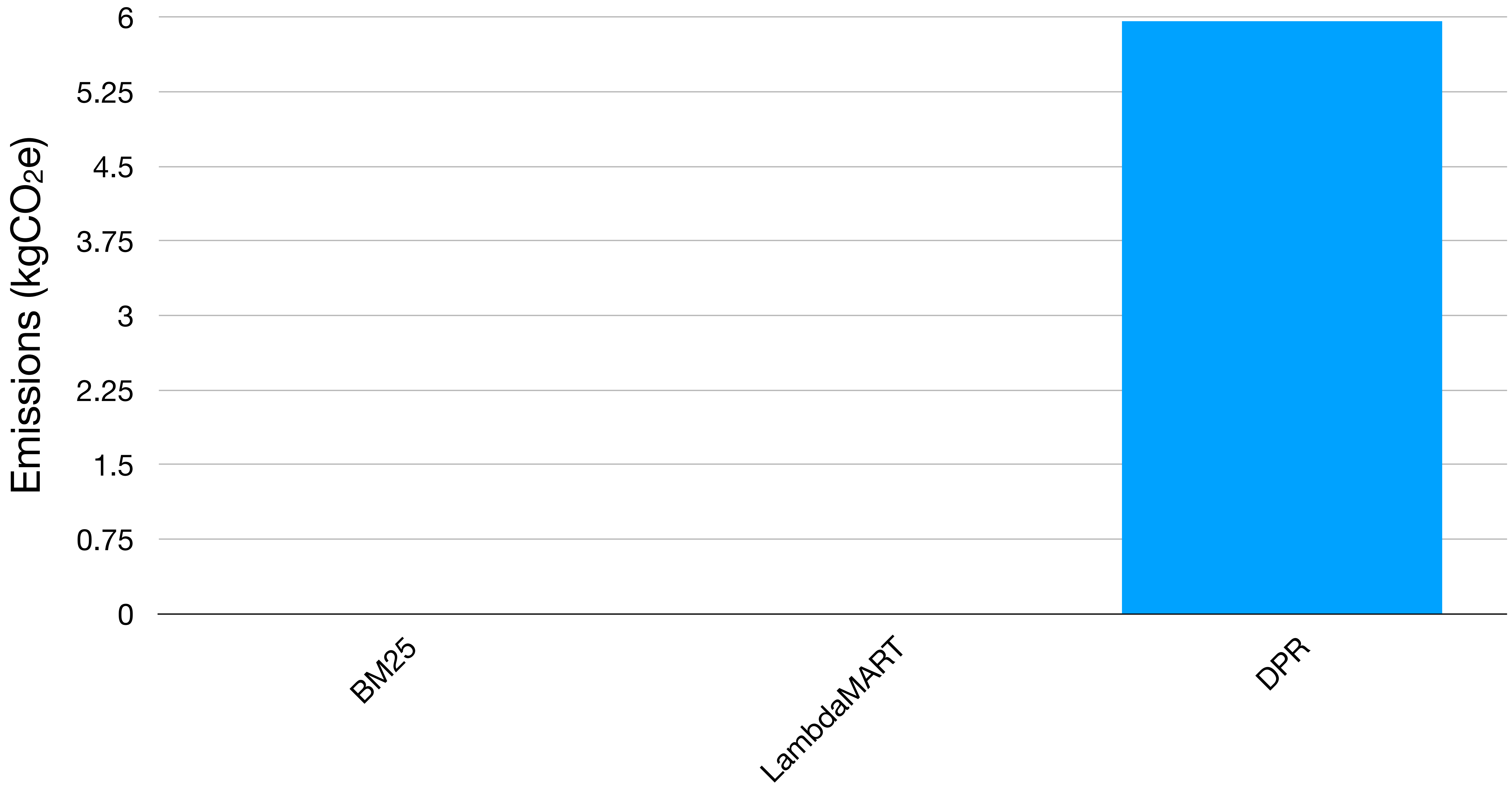
```
from codecarbon import EmissionsTracker

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

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



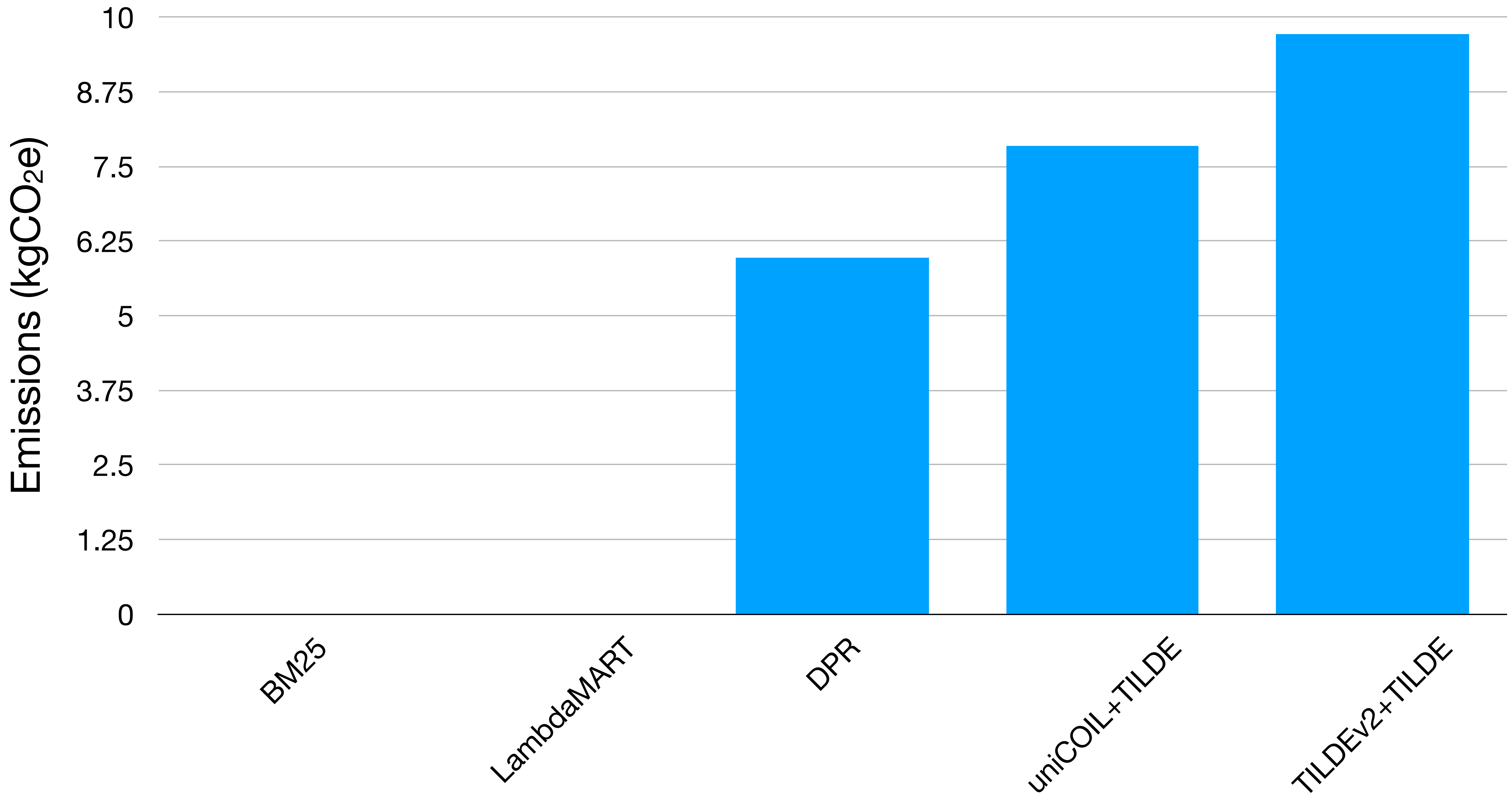
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

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?

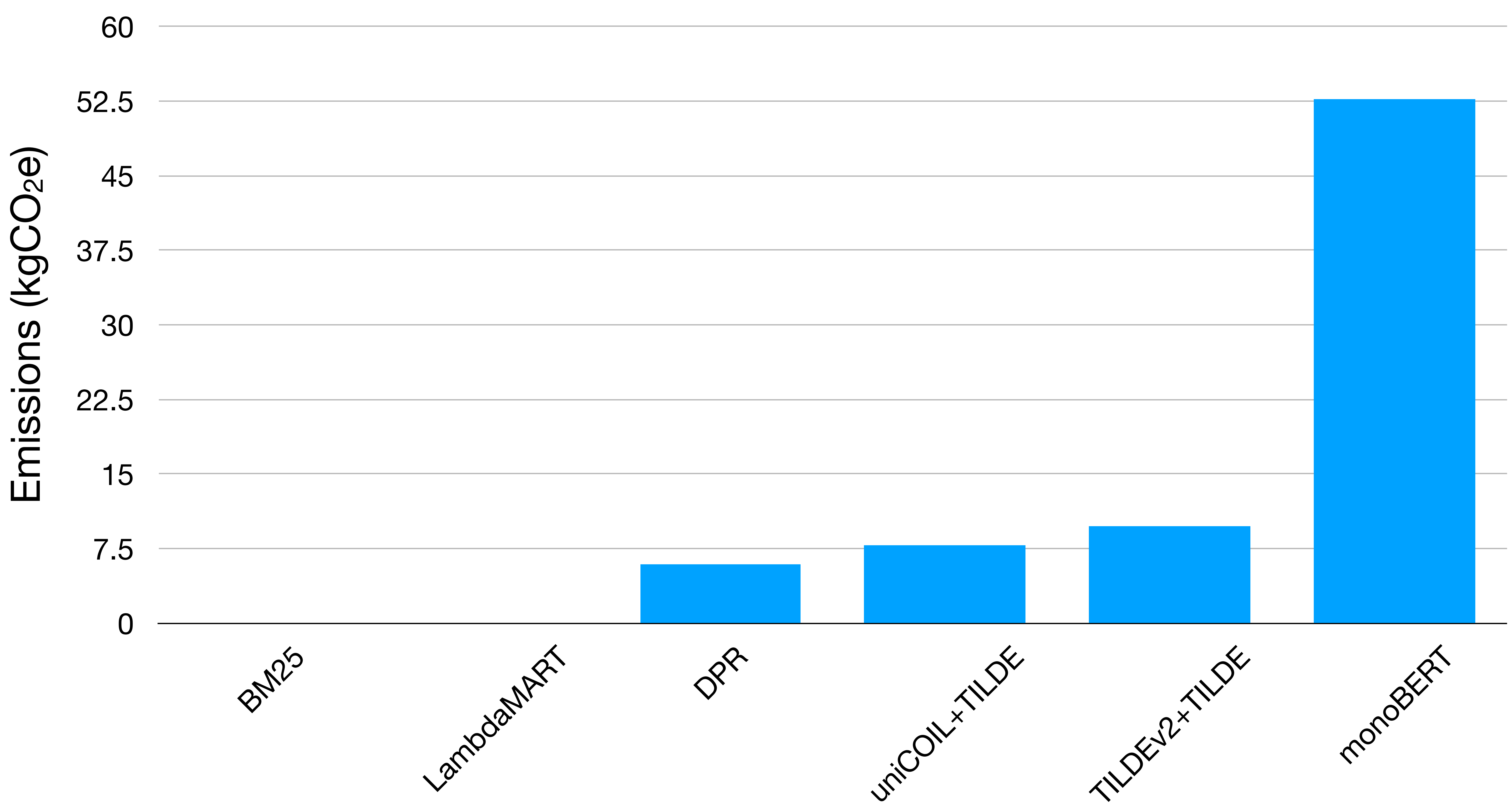


- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

ie lab

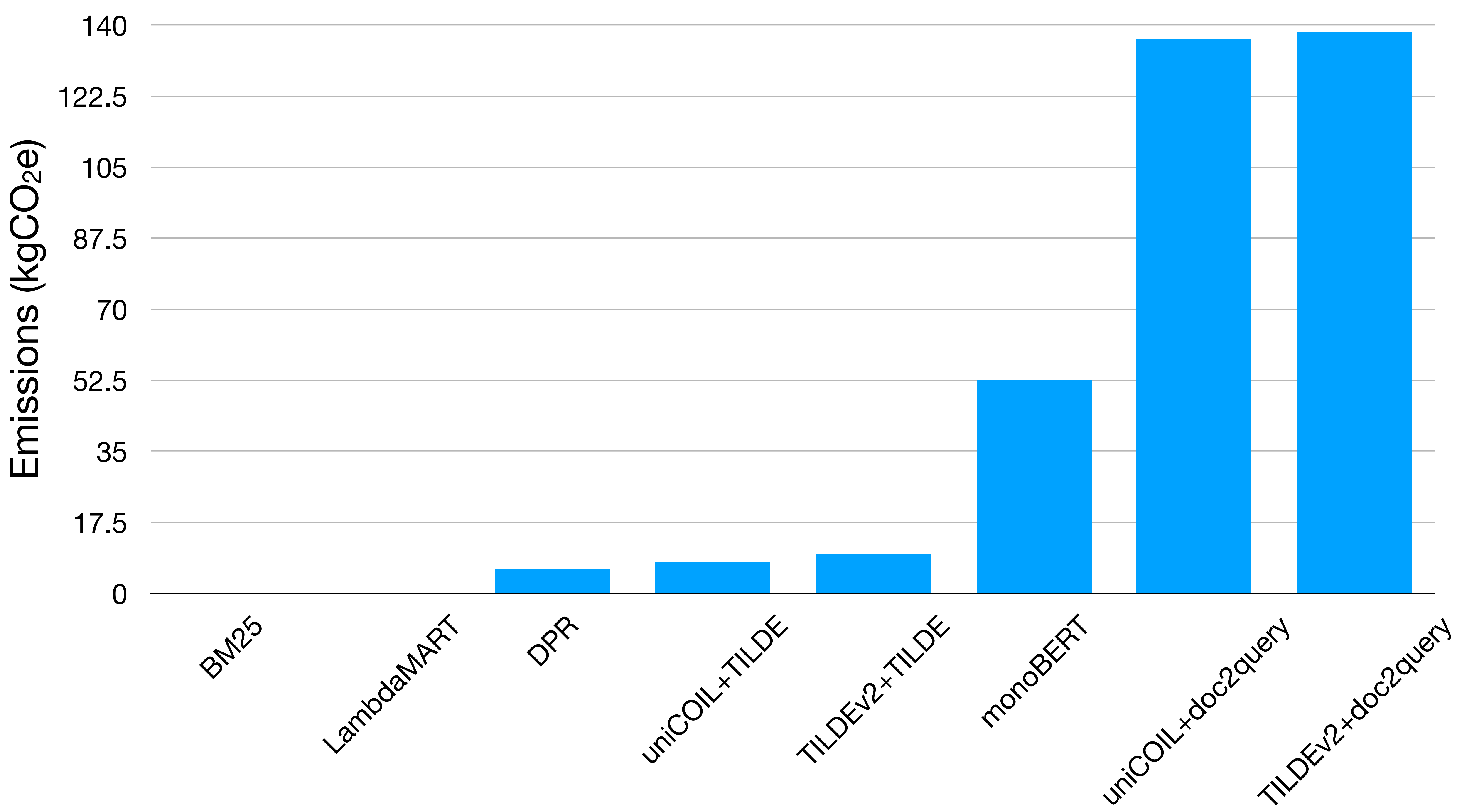
The University Of Queensland Australia
Create Change

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?



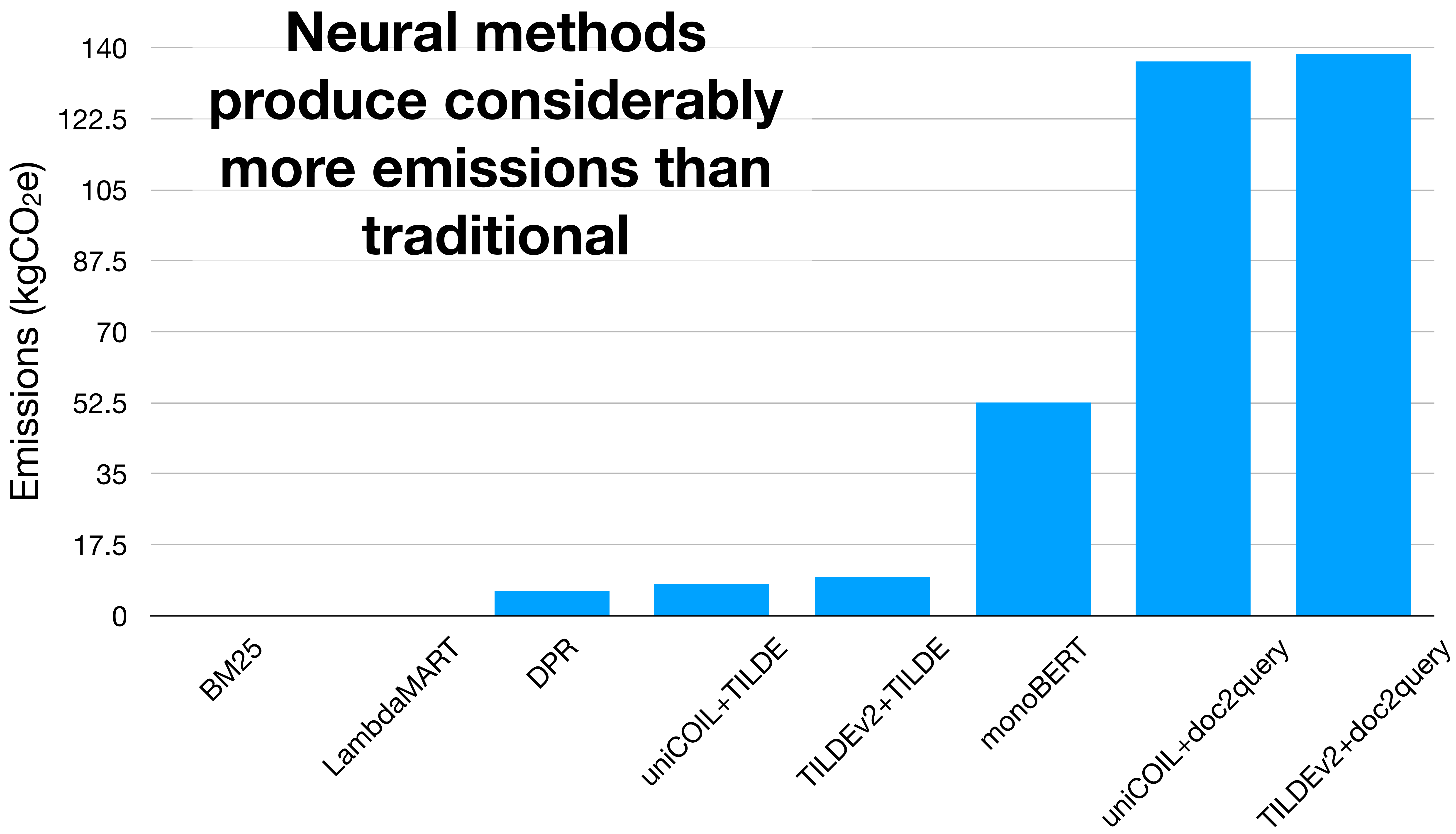
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

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



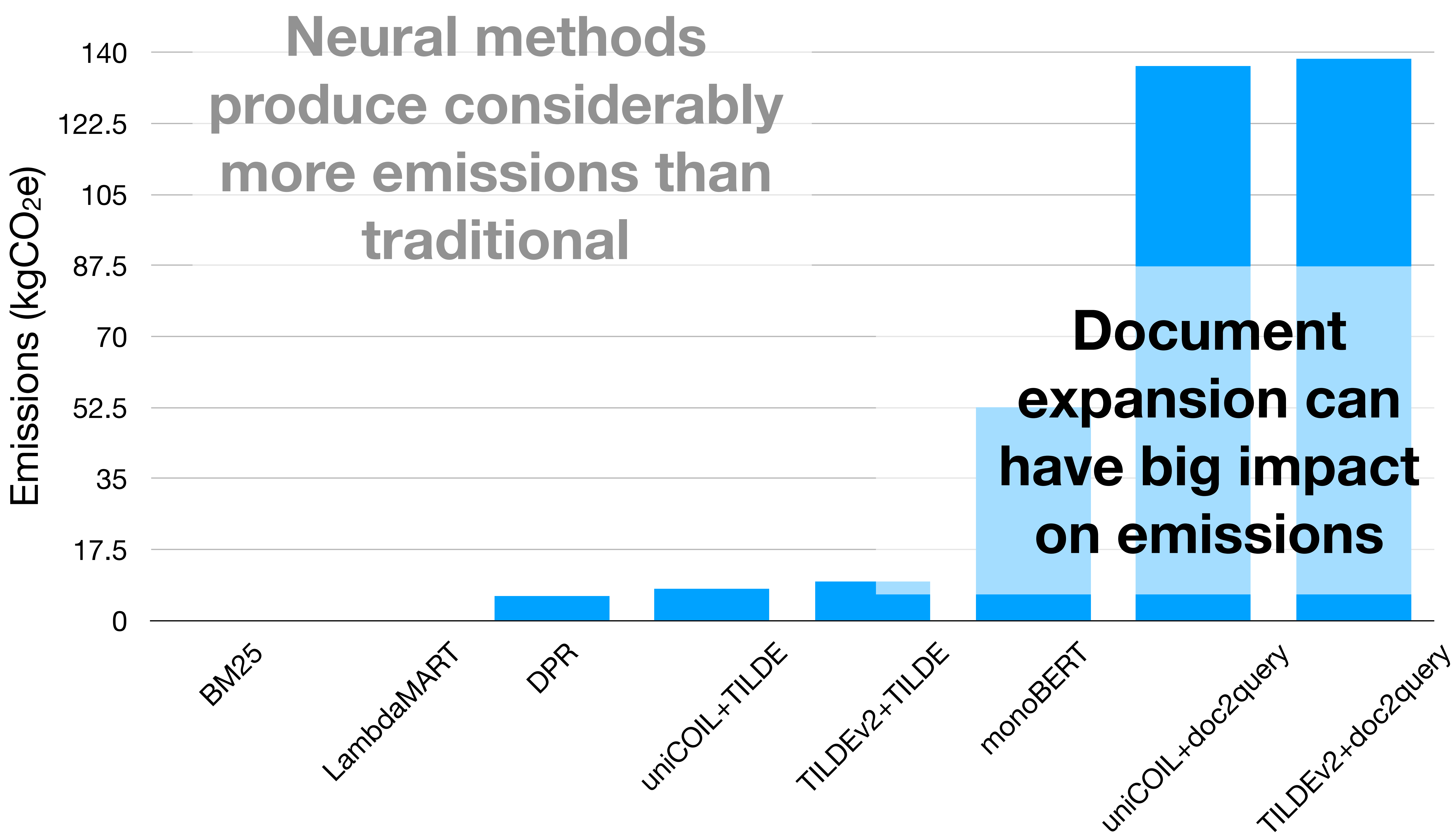
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

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



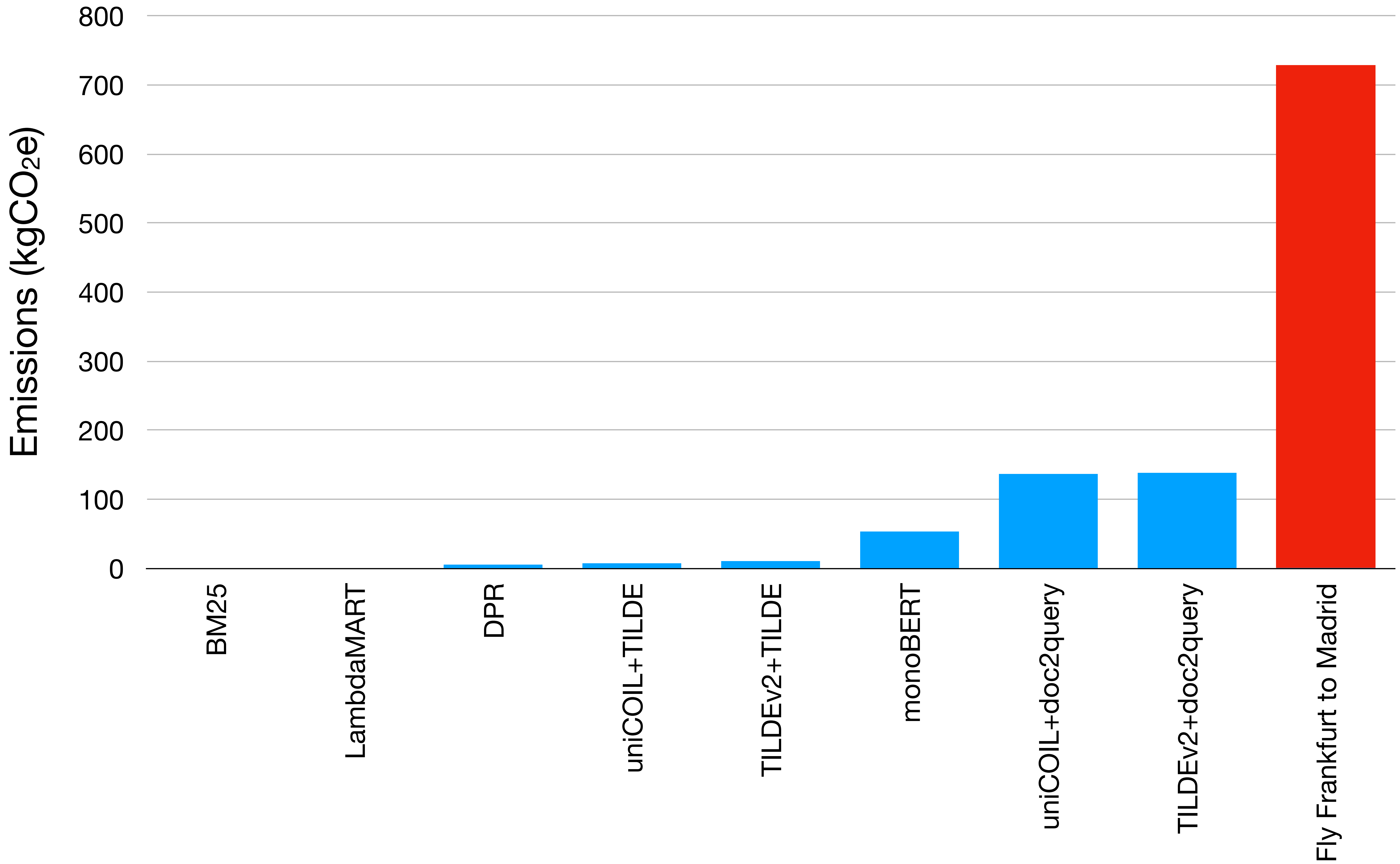
- Methods:
  - BM25
  - LambdaMART
  - DPR
  - monoBERT
  - TILDEv2
  - uniCOIL

ie lab

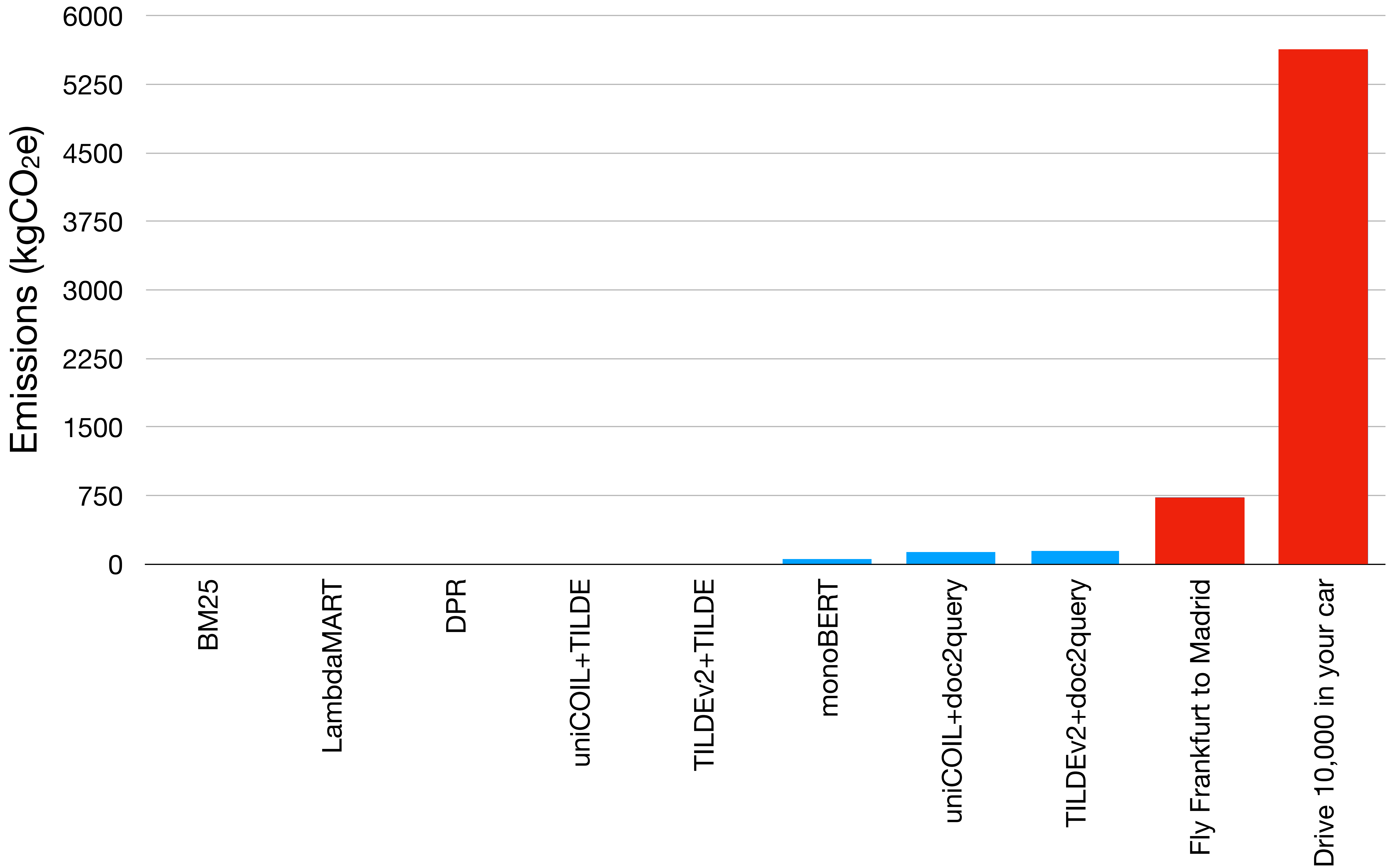
THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

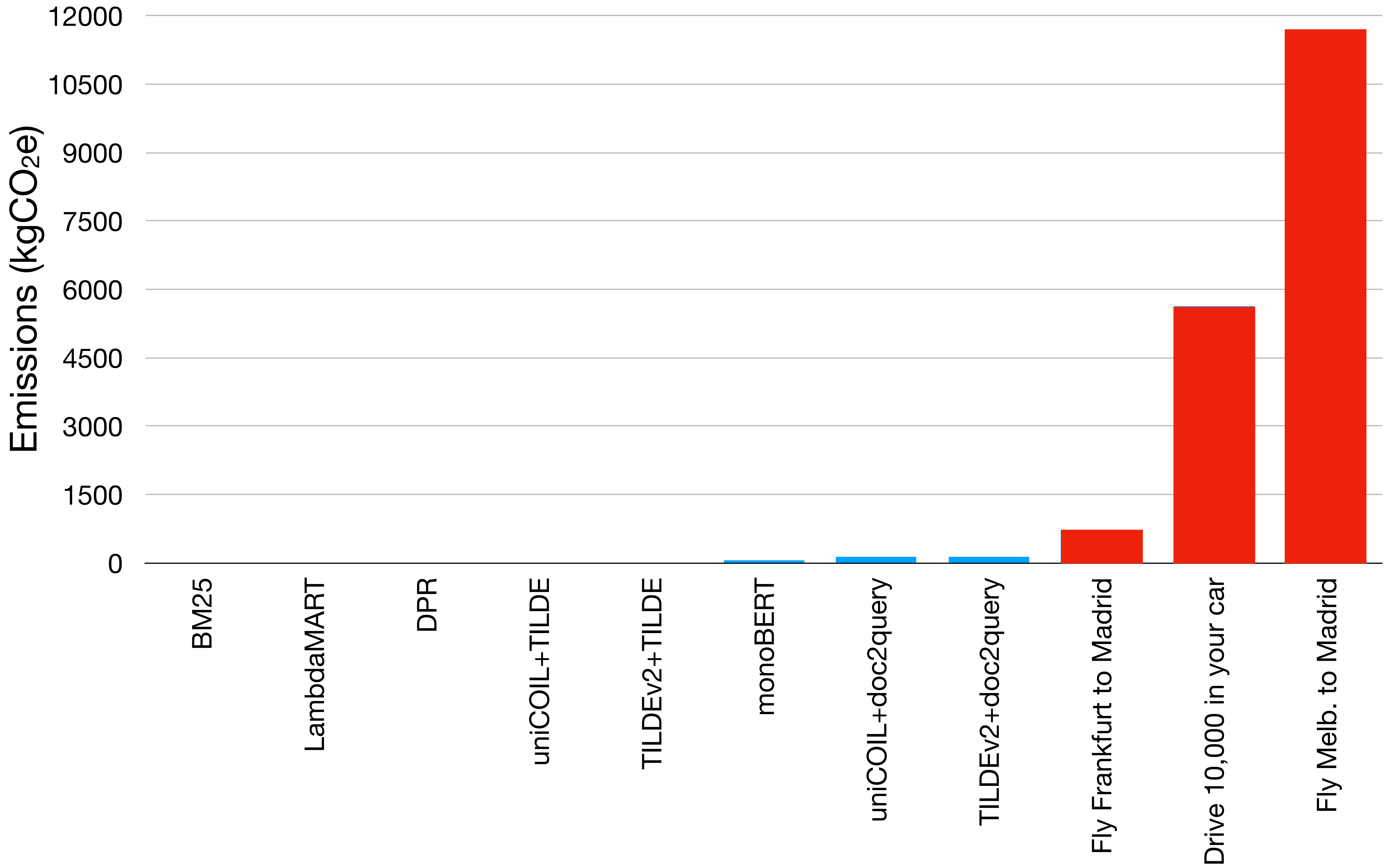
How many emissions do these methods produce to obtain an experimental result?



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



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

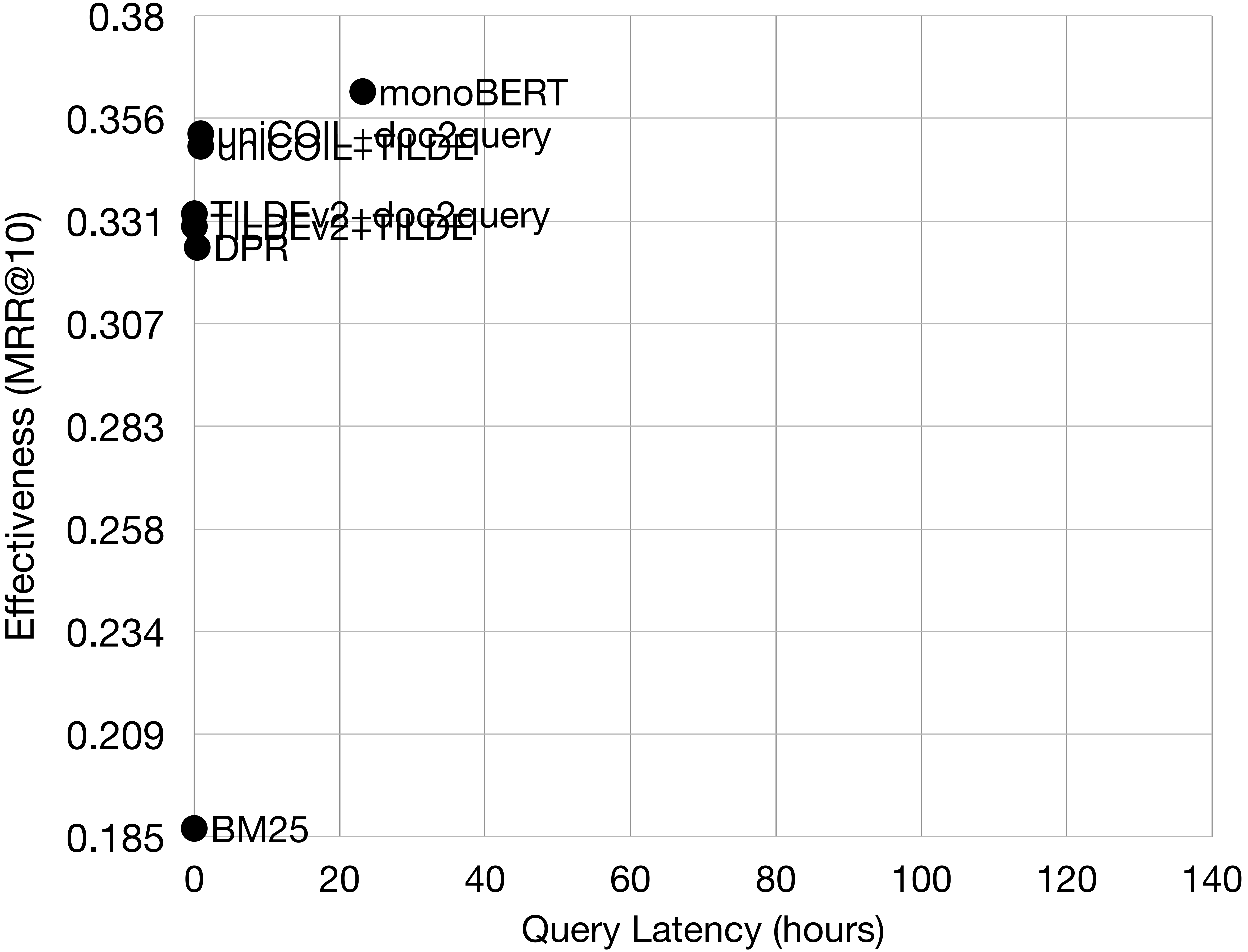


ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

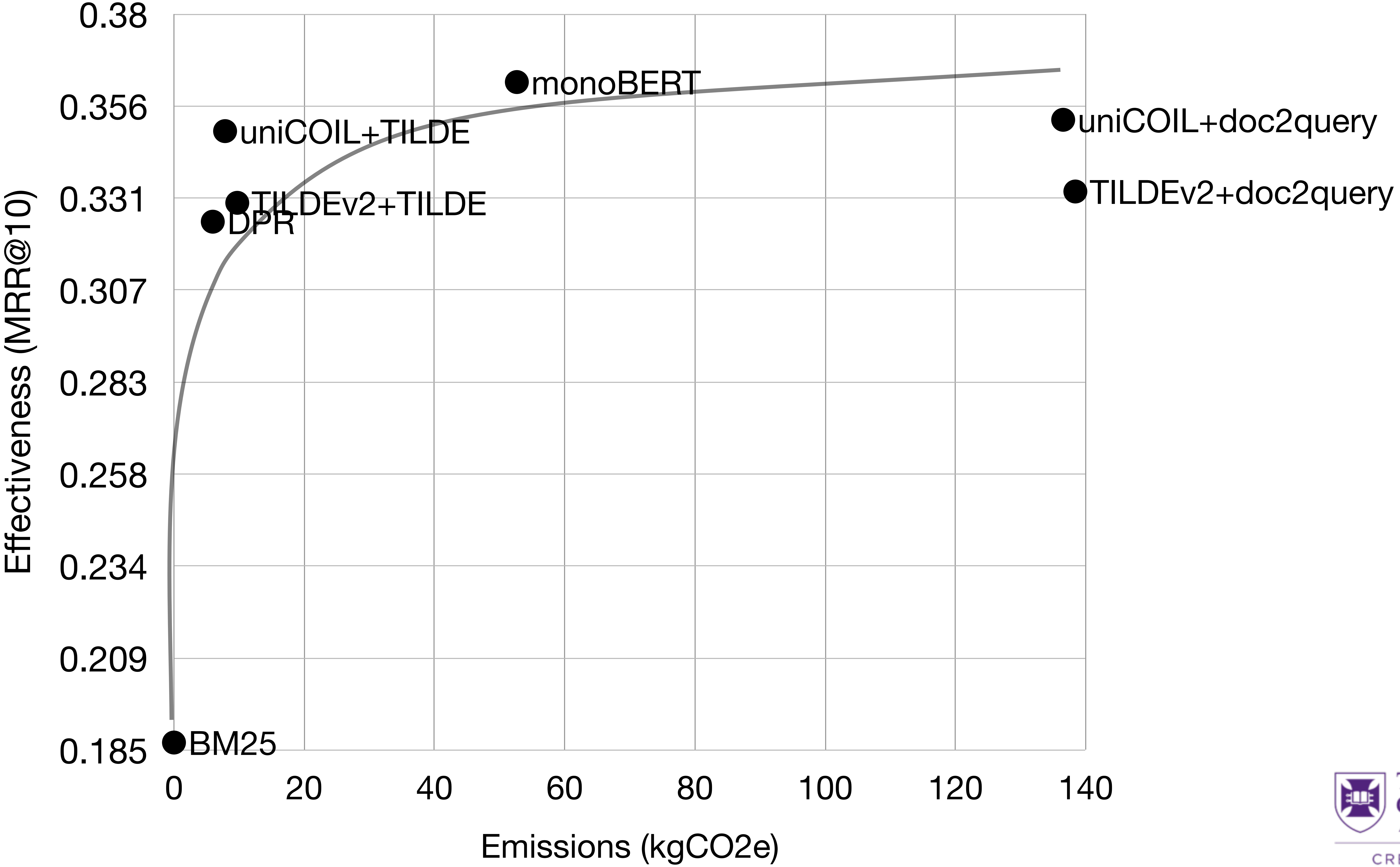
What are the effectiveness-utilisation trade-offs of these methods?



ie lab

The University of Queensland
Australia
Create Change

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



ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE


REDUCE
REUSE
RECYCLE

a framework for IR practitioners to remain mindful of the potential costs of IR research

Reduce

vs

Reduce

vs

expend fewer resources

ie
lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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
- 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

ie lab

The University of Queensland
Australia
Create Change

Recycle

ie lab

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

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 TILDE

ie lab

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

Outlook

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

Outlook

- **Larger neural methods** = power-hungry hardware = utilisation of more power
    - but: increase 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 entire indexing & searching architecture into single model

Outlook

- **Larger neural methods** = power-hungry hardware = utilisation of more power
  - but: increase 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 entire indexing & searching architecture into single model
- IR community at a **turning point**
  - Bigger/more complex models
  - Bigger collections

Outlook

- **Larger neural methods** = power-hungry hardware = utilisation of more power
  - but: increase 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 entire indexing & searching architecture into single model
- IR community at a **turning point**
  - Bigger/more complex models
  - Bigger collections
- Let's be mindful of the **cost** of IR research
  - Power usage → $$$
  - Emissions → CO_2e

ie lab

The University Of Queensland Australia
CREATE CHANGE