# How to Choose a Good LLM

Team Members:

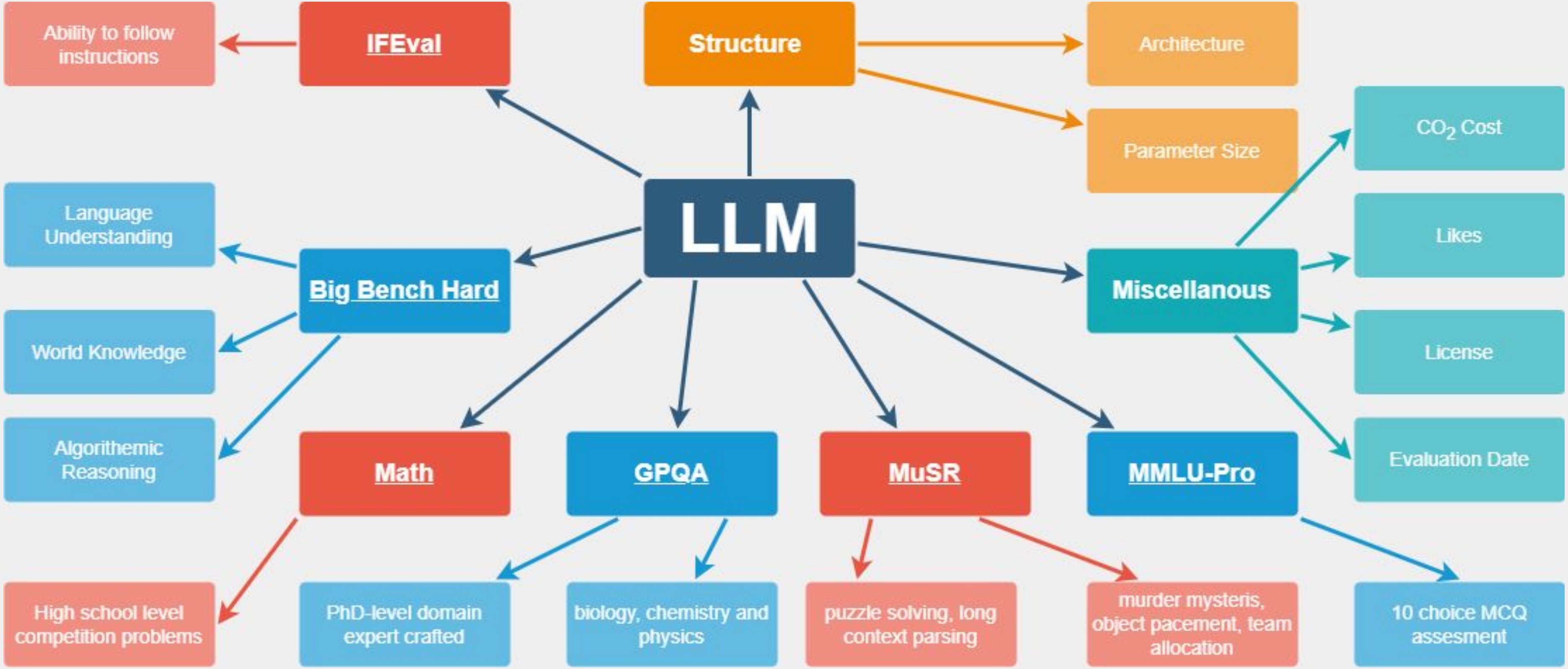
- Wang, Yifan
- Khanna, Sunreet
- Li, Jason
- Xu, Jiabao

# Motivation

- Growing importance of Large Language Models (LLMs)

- **Challenges in selecting the right LLM:**
  - Increasing number of models
  - Countless evaluation metrics
  - Varying needs

# Dataset Introduction



LLM

IFEval

Ability to follow instructions

Structure

Architecture

Parameter Size

Big Bench Hard

Language Understanding

World Knowledge

Algorithemic Reasoning

Miscellanous

$CO_2$ Cost

Likes

License

Evaluation Date

Math

High school level competition problems

GPQA

PhD-level domain expert crafted

biology, chemistry and physics

MuSR

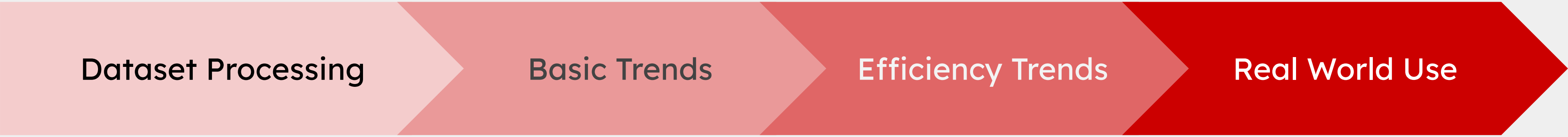
puzzle solving, long context parsing

murder mysteris, object pacement, team allocation

MMLU-Pro

10 choice MCQ assesment

# Methodology



Dataset Processing

Basic Trends

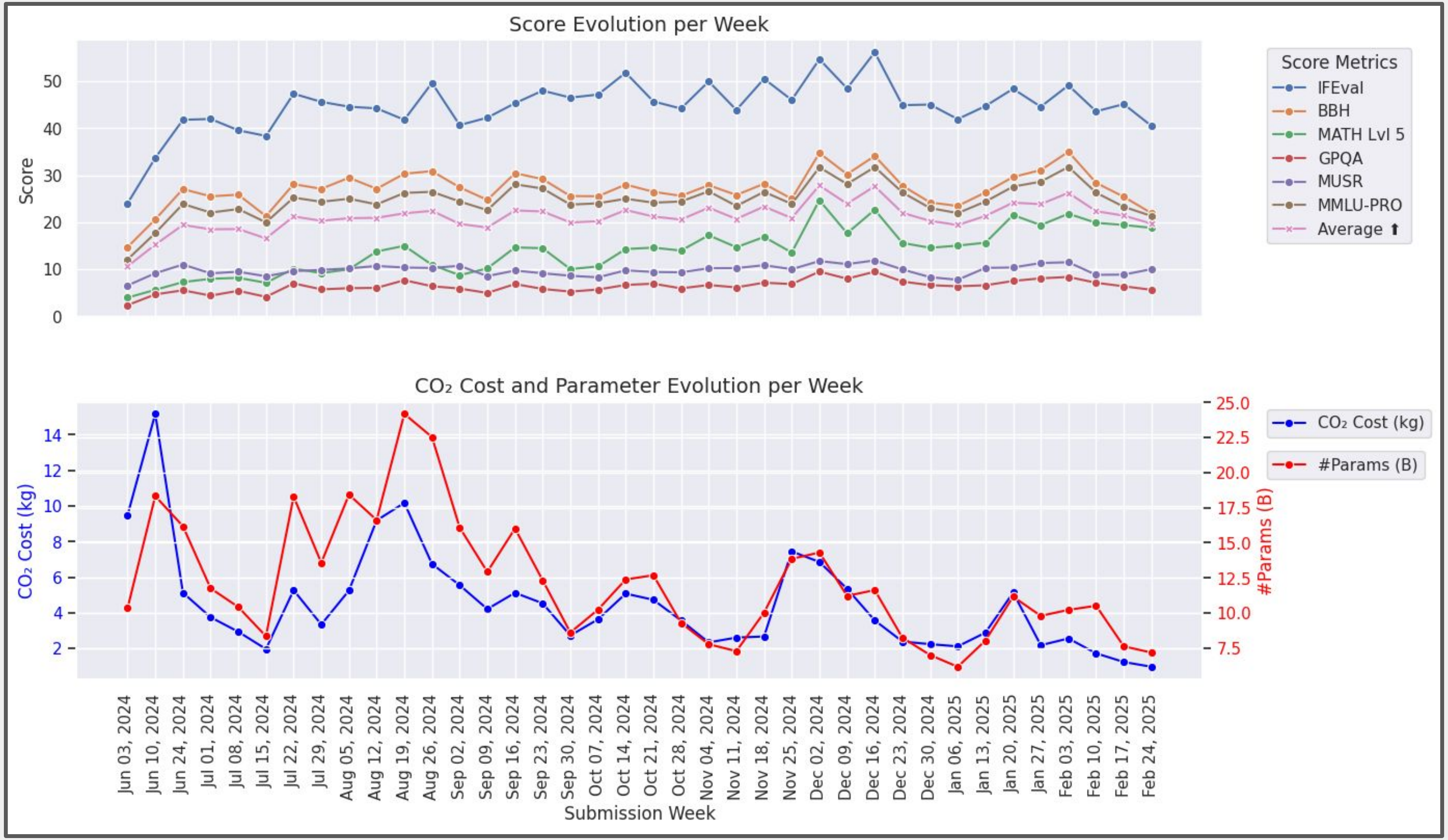
Efficiency Trends

Real World Use

- Correlations
- Importance of Architecture

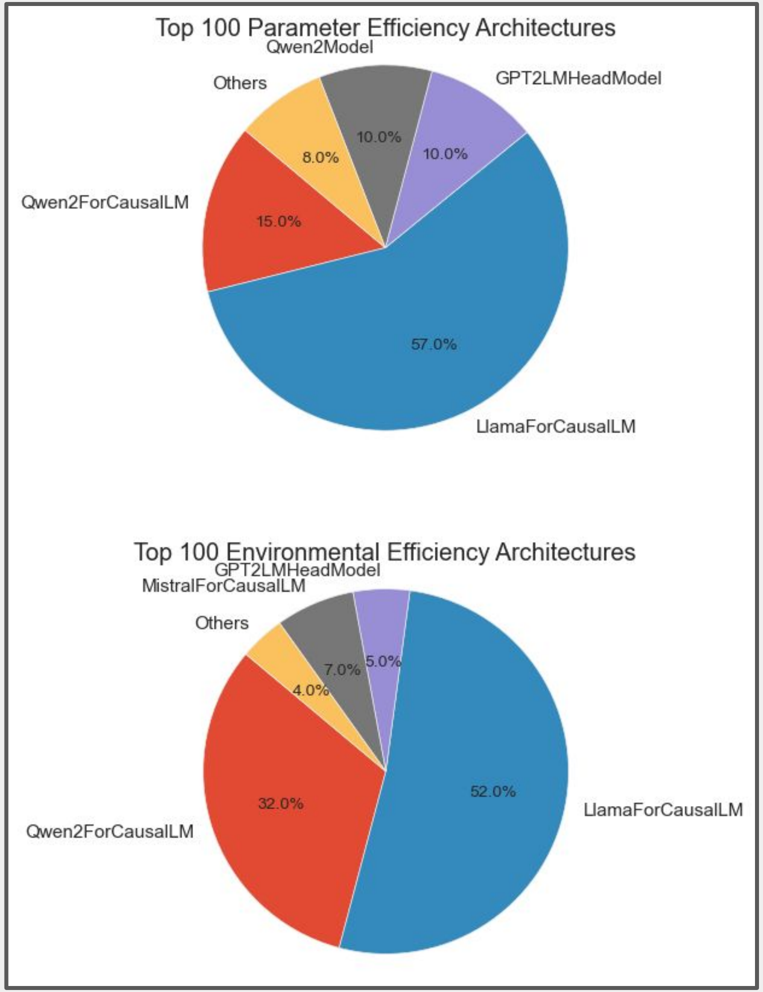
- Industry requirement
- Ranking models

# Same Performance but Smaller size, Lower Cost

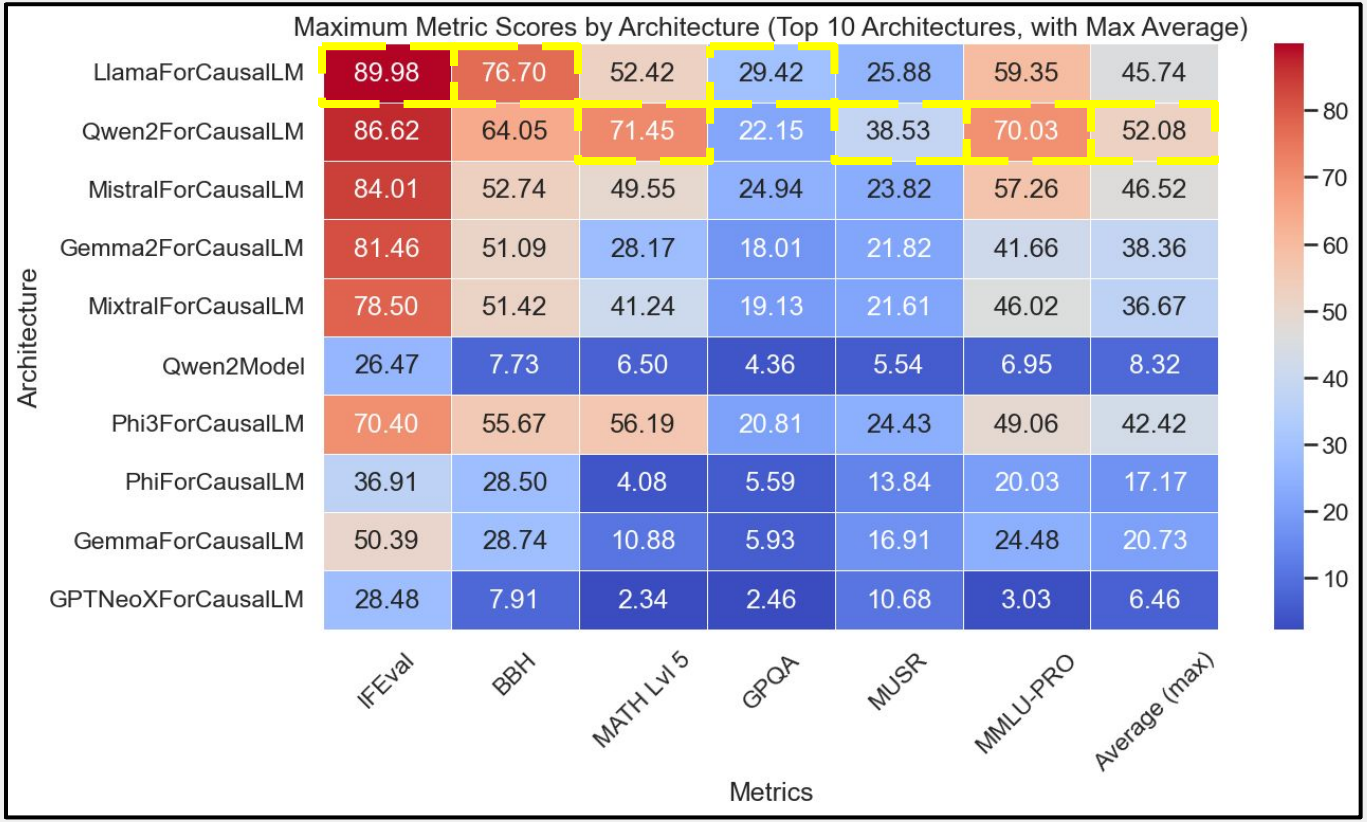


# Efficiency Matters

- Parameter Efficiency
  - Metrics Average / #Params (B)
- Environmental Efficiency
  - Metrics Average / $CO_2$ cost (kg)
- **Architecture** is important



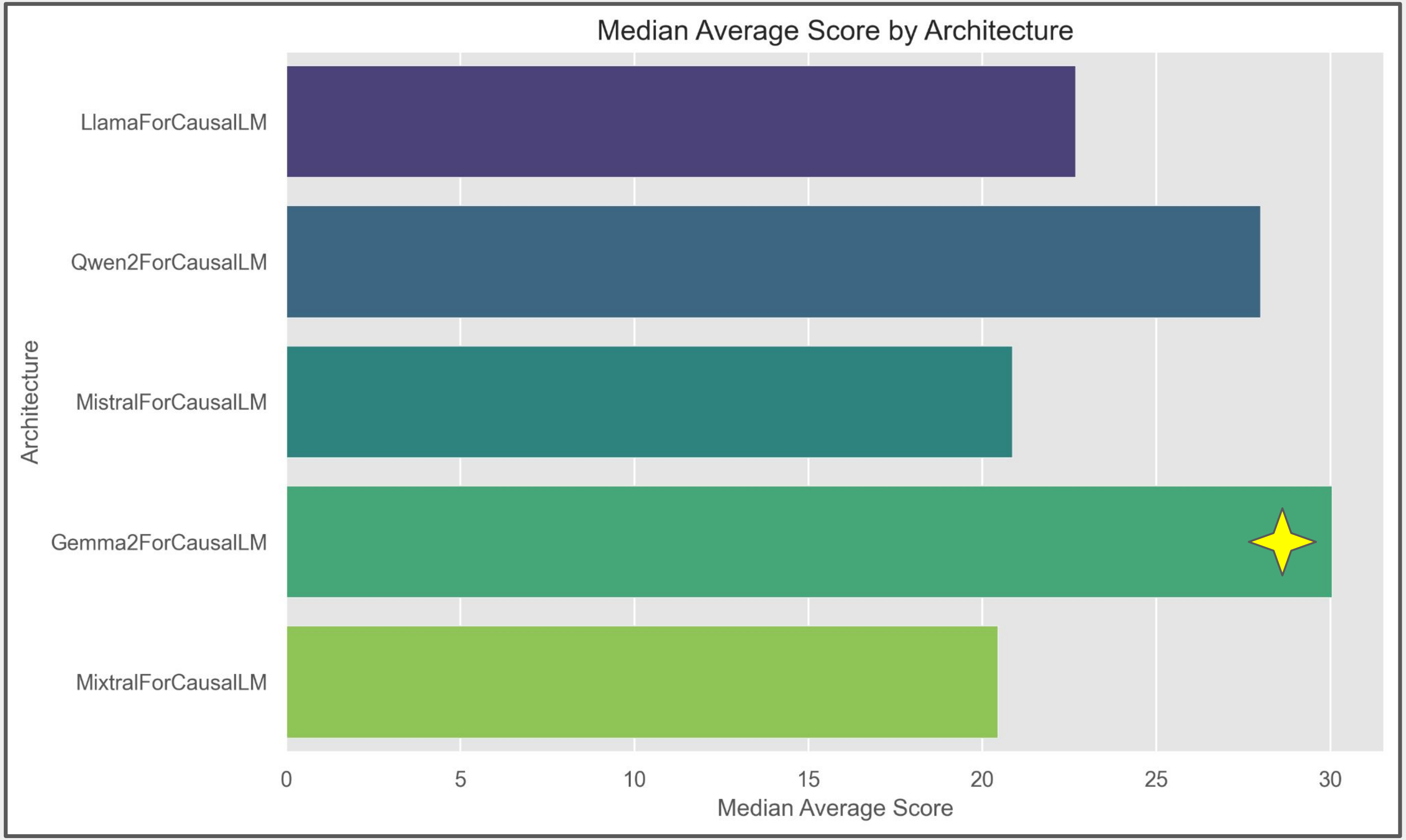
# Architecture- Max Performance



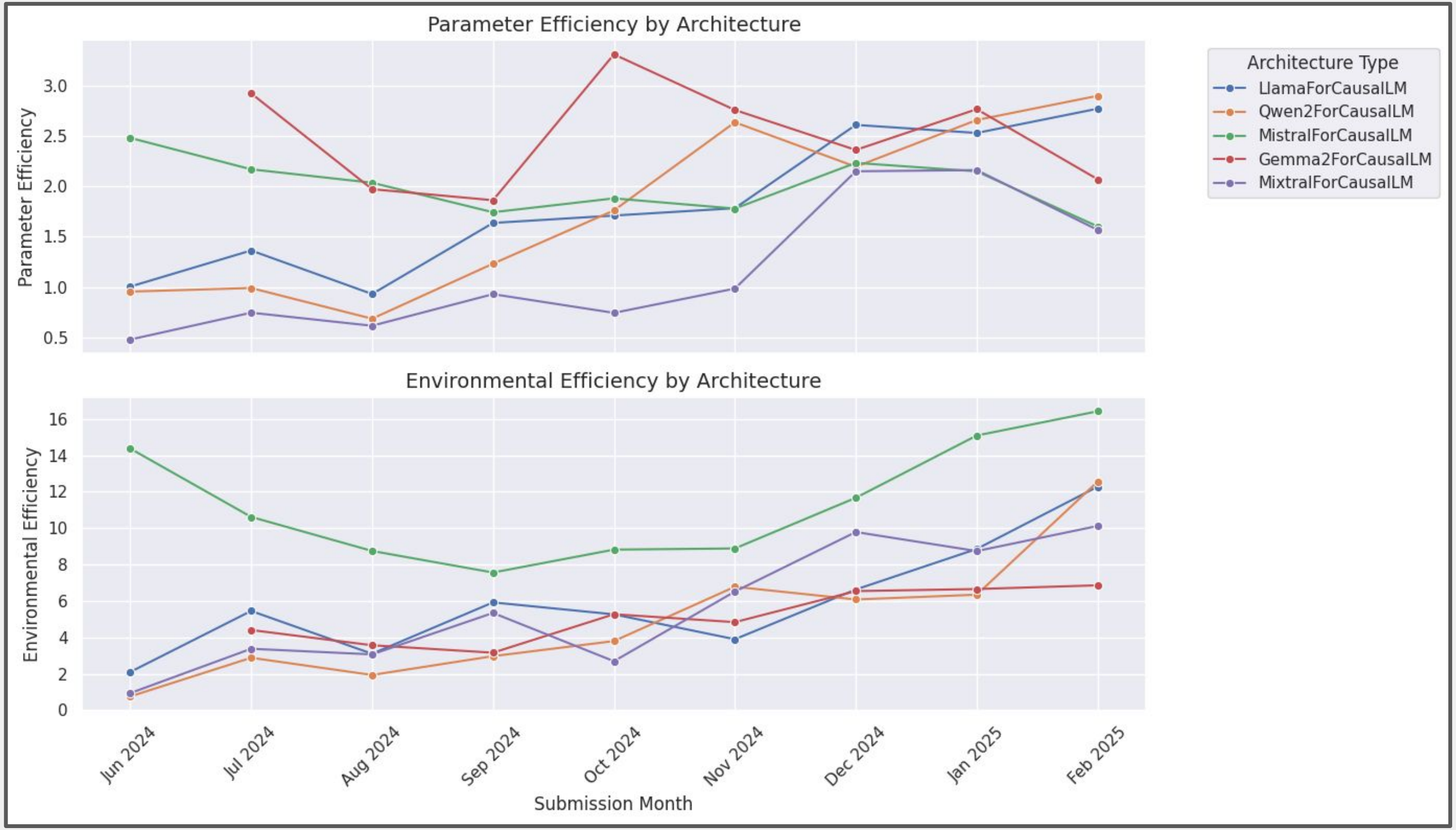
Maximum Metric Scores by Architecture (Top 10 Architectures, with Max Average)

| Architecture | IFEval | BBH | MATH Lvl 5 | GPQA | MUSR | MMLU-PRO | Average (max) |
|---|---|---|---|---|---|---|---|
| LlamaForCausalLM | 89.98 | 76.70 | 52.42 | 29.42 | 25.88 | 59.35 | 45.74 |
| Qwen2ForCausalLM | 86.62 | 64.05 | 71.45 | 22.15 | 38.53 | 70.03 | 52.08 |
| MistralForCausalLM | 84.01 | 52.74 | 49.55 | 24.94 | 23.82 | 57.26 | 46.52 |
| Gemma2ForCausalLM | 81.46 | 51.09 | 28.17 | 18.01 | 21.82 | 41.66 | 38.36 |
| MixtralForCausalLM | 78.50 | 51.42 | 41.24 | 19.13 | 21.61 | 46.02 | 36.67 |
| Qwen2Model | 26.47 | 7.73 | 6.50 | 4.36 | 5.54 | 6.95 | 8.32 |
| Phi3ForCausalLM | 70.40 | 55.67 | 56.19 | 20.81 | 24.43 | 49.06 | 42.42 |
| PhiForCausalLM | 36.91 | 28.50 | 4.08 | 5.59 | 13.84 | 20.03 | 17.17 |
| GemmaForCausalLM | 50.39 | 28.74 | 10.88 | 5.93 | 16.91 | 24.48 | 20.73 |
| GPTNeoXForCausalLM | 28.48 | 7.91 | 2.34 | 2.46 | 10.68 | 3.03 | 6.46 |

# Architecture - Median Average Score



# Architecture - Efficiency Evolution



# Industry Requirements

## Tech

- Reasoning
- Solve complex problem

## Academic

- Reasoning
- Ability in maths and science

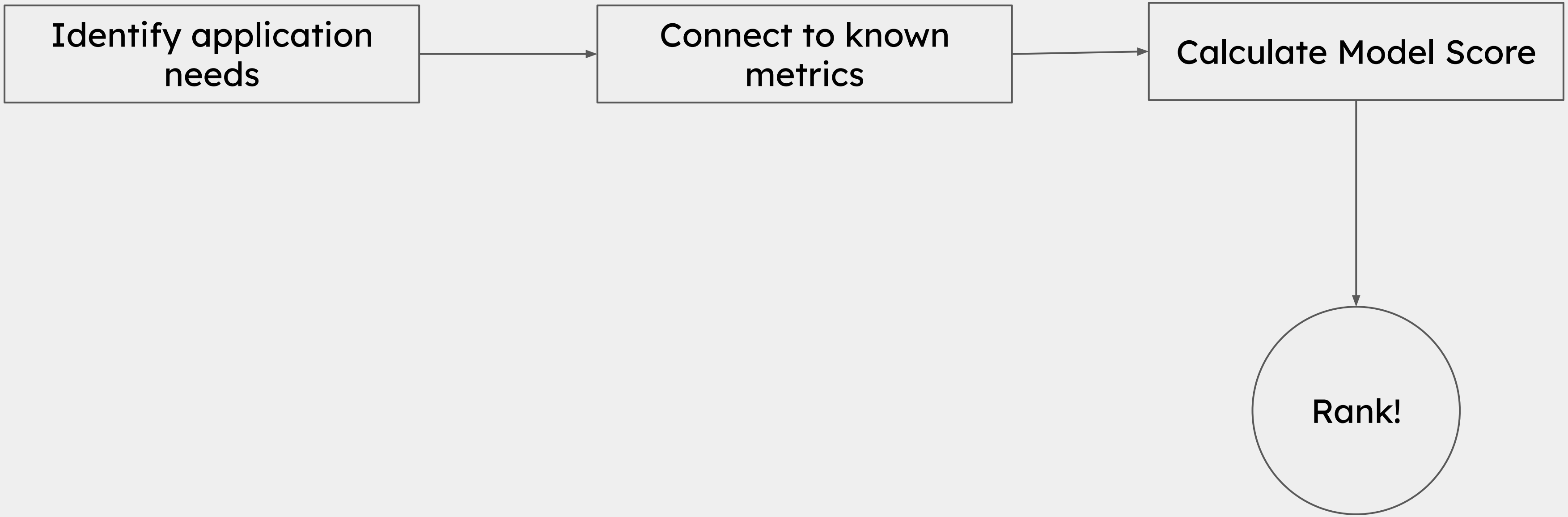
## Customer Service

- Chatbot
- Relevant Conv.
- Knowledge retention

## Manufacturing

- Engineering proficiency
- Instruction adherence

# Real World Use - Workflow



# Real World Use - Linking to Data

Tech

- MuSR
- # of Params
- Architecture

Academic

- MuSR
- Math lv 5
- GPQA

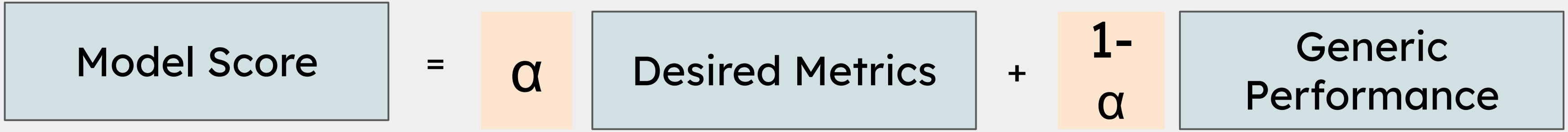
Customer Service

- IFEval
- MMLU-Pro
- Chatbot Template

Manufacturing

- MMLU-Pro
- BBH
- IFEval
- Architecture
- Fine-Tune

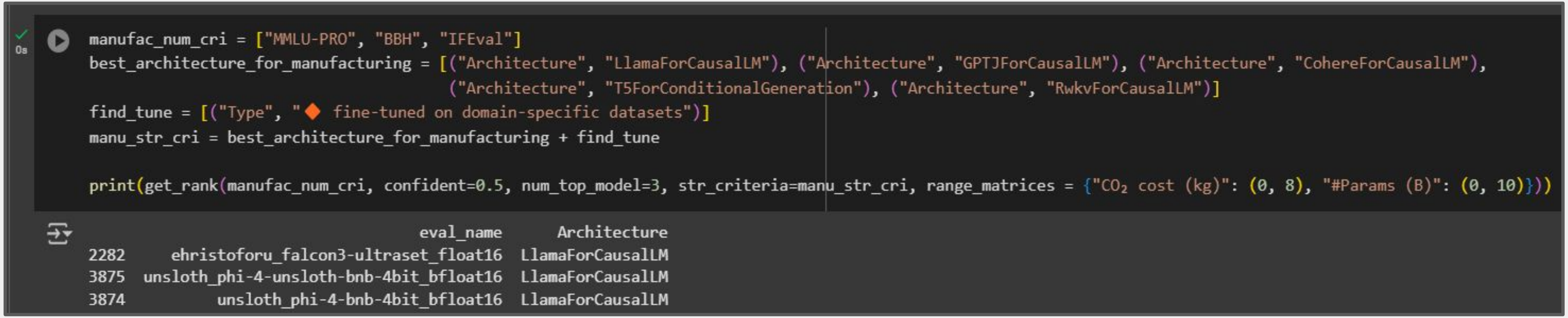
# Real World Use - Some ideas on ranking



Model Score = α Desired Metrics + 1- α Generic Performance

- Desired Metrics
  - Numerical - MMLU-Pro, GPQA
  - Categorical - Architecture, Fine Tuning
- α: A coefficient of confidence (from 0 -1) on Desired Metrics
  - High confidence → Good for specific industry
  - Low confidence → Generally well performing model

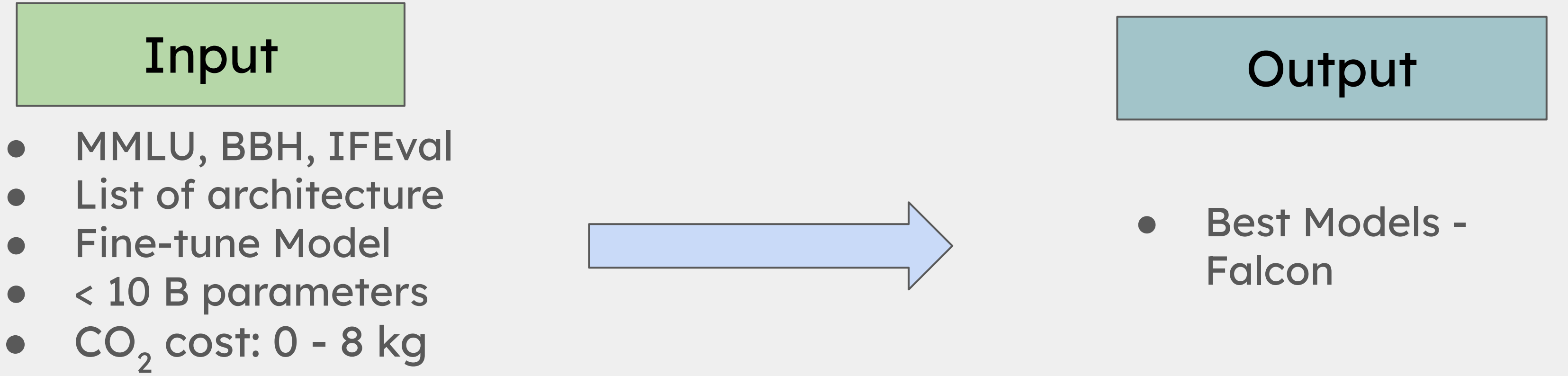
# Demo



```
manufac_num_cri = ["MMLU-PRO", "BBH", "IFEval"]
best_architecture_for_manufacturing = [("Architecture", "LlamaForCausalLM"), ("Architecture", "GPTJForCausalLM"), ("Architecture", "CohereForCausalLM"),
                                       ("Architecture", "T5ForConditionalGeneration"), ("Architecture", "RwkvForCausalLM")]
find_tune = [("Type", "🔶 fine-tuned on domain-specific datasets")]
manu_str_cri = best_architecture_for_manufacturing + find_tune

print(get_rank(manufac_num_cri, confident=0.5, num_top_model=3, str_criteria=manu_str_cri, range_matrices = {"CO₂ cost (kg)": (0, 8), "#Params (B)": (0, 10)}))
```

```
                                    eval_name      Architecture
2282     ehristoforu_falcon3-ultraset_float16  LlamaForCausalLM
3875  unsloth_phi-4-unsloth-bnb-4bit_bfloat16  LlamaForCausalLM
3874          unsloth_phi-4-bnb-4bit_bfloat16  LlamaForCausalLM
```



Input

- MMLU, BBH, IFEval
- List of architecture
- Fine-tune Model
- < 10 B parameters
- $CO_2$ cost: 0 - 8 kg

Output

- Best Models - Falcon

# Demo Results

| Industry Name | Input | Output | Architecture |
|---|---|---|---|
| Tech | "$CO_2$ cost (kg)": (4, 8)<br>"#Params (B)": (20, 30)<br>'MMLU-PRO': (30, 40)<br>["MUSR", "IFEval"] | ● Sumatra-20b<br>● Venti-Blend-sce<br>● Venti-20b | ● LlamaForCausalLM<br>● LlamaForCausalLM<br>● LlamaForCausalLM |
| Academic | "$CO_2$ cost (kg)": (8, 12)<br>"#Params (B)": (0,35)<br>["MUSR", "MATH Lvl 5", "GPQA"] | ● ultiima-32B<br>● Qwen2.5-32B-Instruct<br>● lambda-qwen2.5-32b-dpo-test | ● Qwen2ForCausalLM<br>● QwenForCausalLM<br>● QwenForCausalLM |

# Conclusion

Performance Trends

Efficiency Trends

Real World Use

Performance seems to saturate

Architecture’s choice determines efficiency

Industry breakdown LLM

# Thanks!

Reference:

- [2407.07000] Etalon: Holistic Performance Evaluation Framework for LLM Inference Systems
- The Gap Between Open and Closed AI Models Might Be Shrinking. Here's Why That Matters
- Large Language Models for Manufacturing
- Why Large Language Models (LLMs) are the future of manufacturing | World Economic Forum
- LLM Chatbot Evaluation Explained: Top Metrics and Testing Techniques - Confident AI