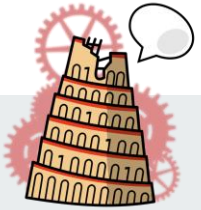
CounteR

PR[AI]RIE
PaRis Artificial Intelligence Research InstitutE

Inria



From Text to Source:
Results in Detecting Large Language Model-Generated Content

Wissam Antoun
Benoît Sagot
Djamé Seddah

ALMAnaCH, INRIA-Paris

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021607

Motivation

- Real-life scenarios lack knowledge of specific text generation models.
- "Cross-Model Detection" investigates if a classifier trained for one model can identify text from another without retraining.
- Aim to discern text generated by different language models without fine-tuning or additional training.

Contribution

- Prior works limited exploration to few model sizes and families.
- Comprehensive Study. We systematically examine the impact of:
  - LLM sizes (from 125M to 70B)
  - Model Families (GPT-2, LLaMA, Pythia, OPT and others)
  - Conversational Finetuning
  - Watermarking
  - Quantization
- We study both cross-model generated text detection, and model attribution.

Methodology

- Cross-Model Detection
  - Objective: Evaluate whether a classifier, initially trained to distinguish text produced by a source LLM from human-written text, can also detect text generated by a target LLM
- Model Attribution
  - 5 Sub-Tasks
    - Source Model Identification
    - Model Family Classification
    - Model Size Classification
    - Quantization Detection
    - Watermark Detection

Experimental Protocol: LLM Choice

We chose the following model families and sizes for our experiments for a total of 55 models:

- **BLOOM** *(Scao et al., 2022)*: 560M, 1.1B, 1.7B, 3B, 7.1B.
- **Cereberas-GPT** *(Dey et al., 2023)*: 111M, 256M, 1.3B, 2.7B, 6.7B, 13B.
- **Falcon, Falcon-Instruct** *(Almazrouei et al., 2023; Penedo et al., 2023)*: 7B and 40B. **Alfred-0723**: 40B
- **GPT-2** *(Radford et al., 2019)*: 124M, 355M, 774M, 1.5B.
- **LLaMA** *(Touvron et al., 2023a)*: 7B, 13B, 30B, 65B. **Vicuna-v1.3** *(Zheng et al., 2023)*: 7B, 13B, 33B
- **LLaMA-v2, LLaMA-v2-Chat** *(Touvron et al., 2023b)*: 7B, 13B, 70B.
- **MPT, MPT-Chat** *(MosaicML, 2023)*: 7B, 30B.
- **OPT** *(Zhang et al., 2022)*: 125m, 350m, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B.
- **OpenLLaMA** *(Geng and Liu, 2023)*: 3B, 7B, 13B.
- **OpenLLaMA-v2** *(Geng and Liu, 2023)*: 3B, 7B.
- **Pythia** *(Biderman et al., 2023)*: 70m, 160m, 410m, 1B, 1.4B, 2.8B, 6.9B, 12B

Experimental Protocol: Data Generation

- **Prompting LLMs:**
  - Use first 10 words of documents from OpenWebText dataset
  - For conversational models, instruct with: "Give the best continuation of the following text:" followed by the 10 words
- **Model Loading:**
  - HuggingFace Text Generation Inference (TGI) server
  - up to 4 48GB NVIDIA GPUs, with float16 precision
- **Hyperparameters** (*Consistent hyperparameters across models*):
  - Maximum 256 tokens per generation
  - Beam-search size: 5
  - Repetition penalty: 1.0
  - Temperature: 1.0
  - Top-k: 10, Top-p: 0.9
  - Typical sampling: 0.9
- **Model Optimization**:
  - 4-bit GPTQ quantization.
  - Watermark text using "red/green" token algorithm by Kirchenbauer et al. (2023)

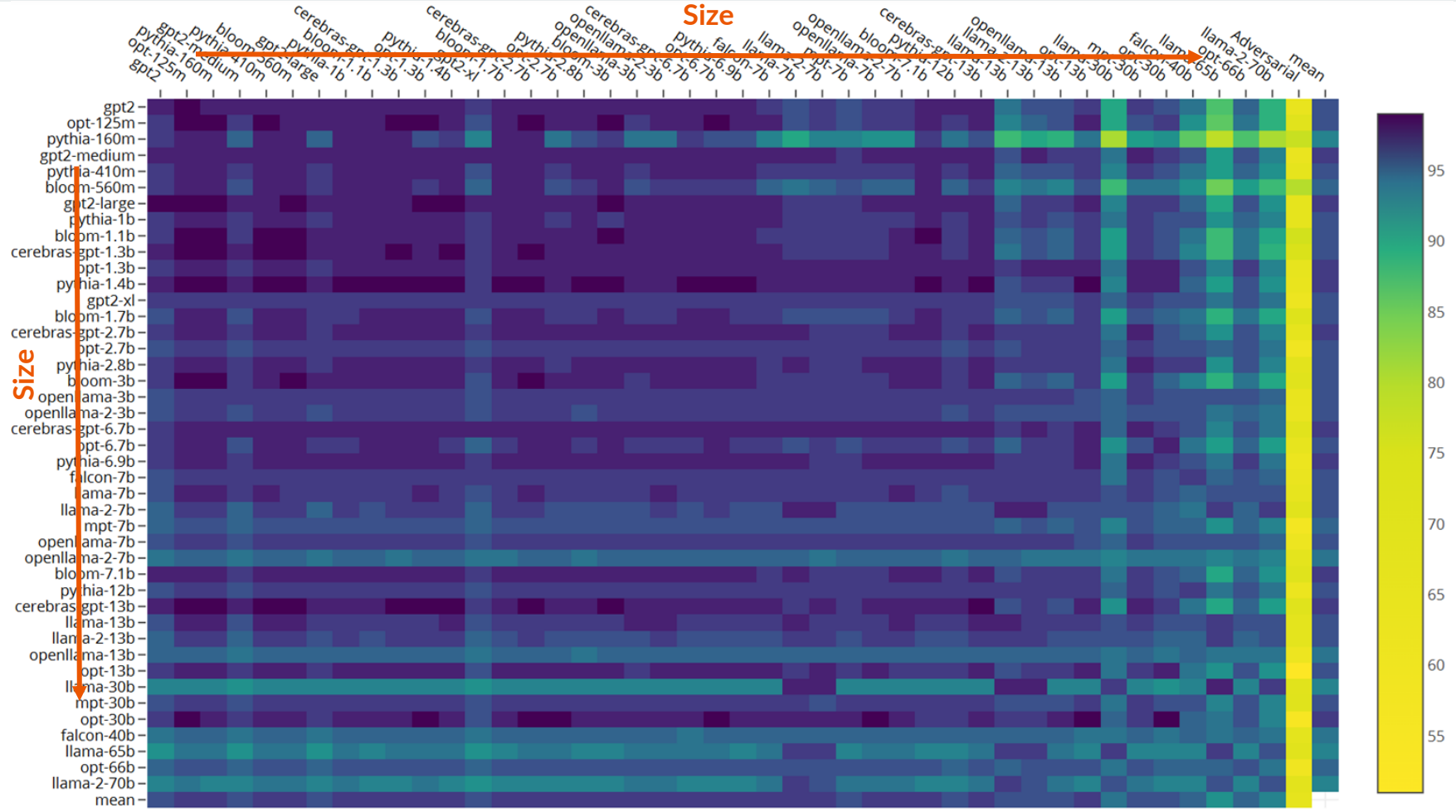
Experimental Protocol: Data Splitting and Filtering

- **Initial Split:**
  - 80% for training, 20% for validation.
- **Filtering:**
  - Remove bad generations:
    - Too short.
    - Repetitive.
    - Contain apologies or "As an AI language model" sentences.
- **Fair Comparison:**
  - Sample 800 training and 200 validation samples from all models.
  - Discard some models unable to generate enough valid examples.
- **Negative Human-Generated Samples:**
  - Sample 800 training and 200 validation samples from OpenWebText dataset for negative human-generated samples.

Experimental Protocol: Classifier

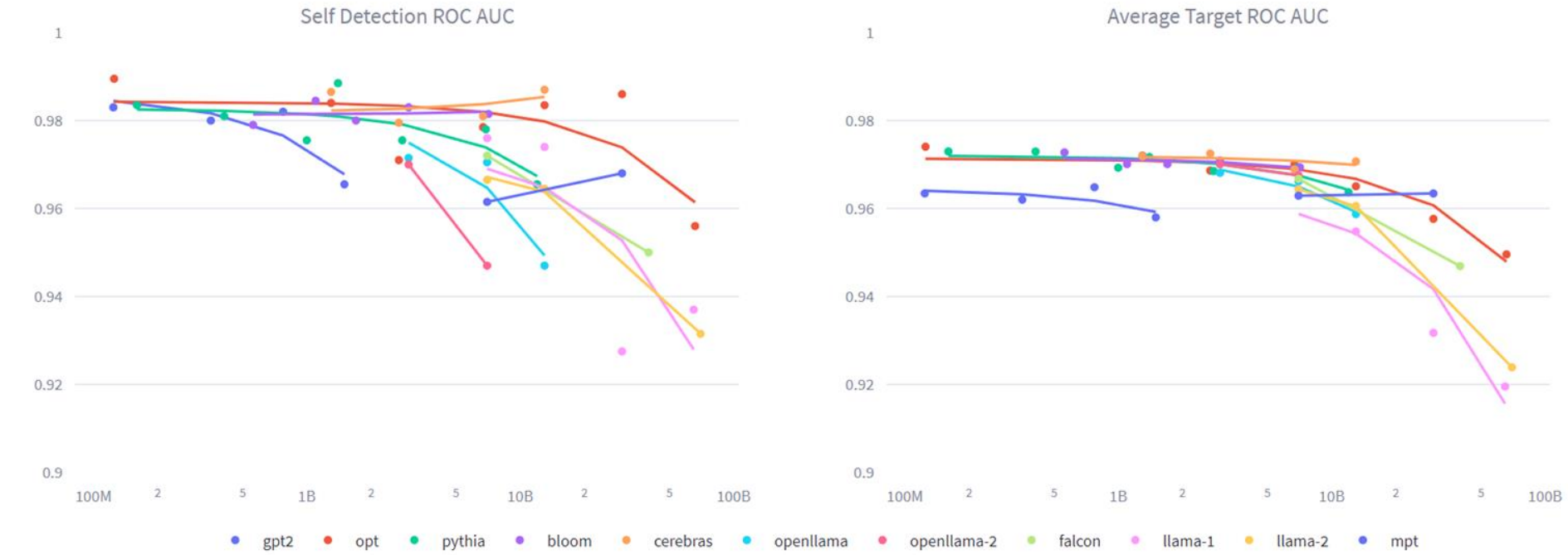
- **Encoder Finetuning:**
  - Popular approach for AI-generated text detection.
  - DeBERTaV3-base.
- **Training Details:**
  - Batch size: 32.
  - Learning rate: 2e-5 for 5 epochs.
- **Robustness Enhancement:**
  - Conduct experiments with five different random seeds.
  - Average resultant AUC scores to mitigate seed-specific variations.

Results: Cross-Model Detection Results

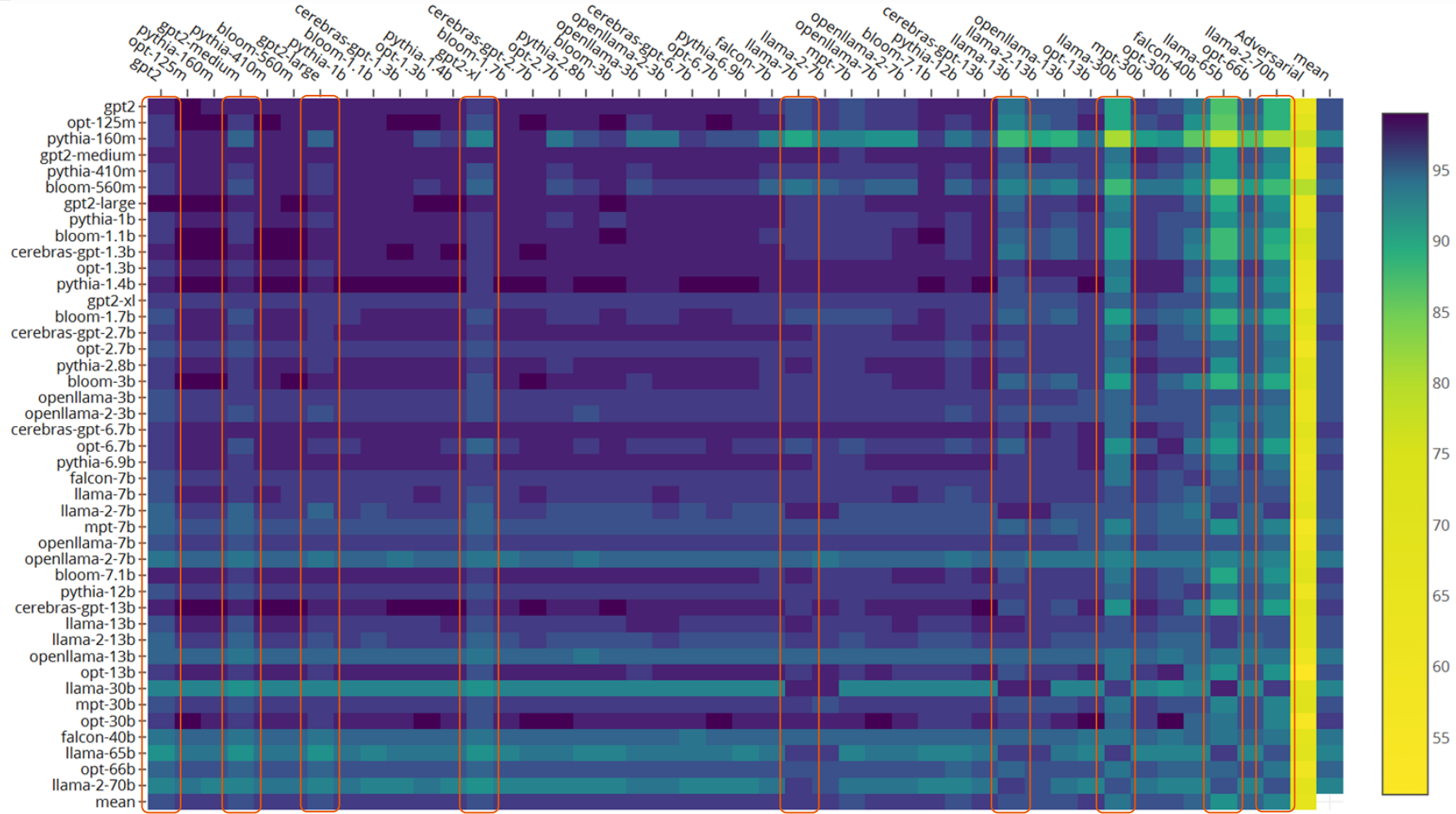


Results: Cross-Model Detection Results - Model Size Influence

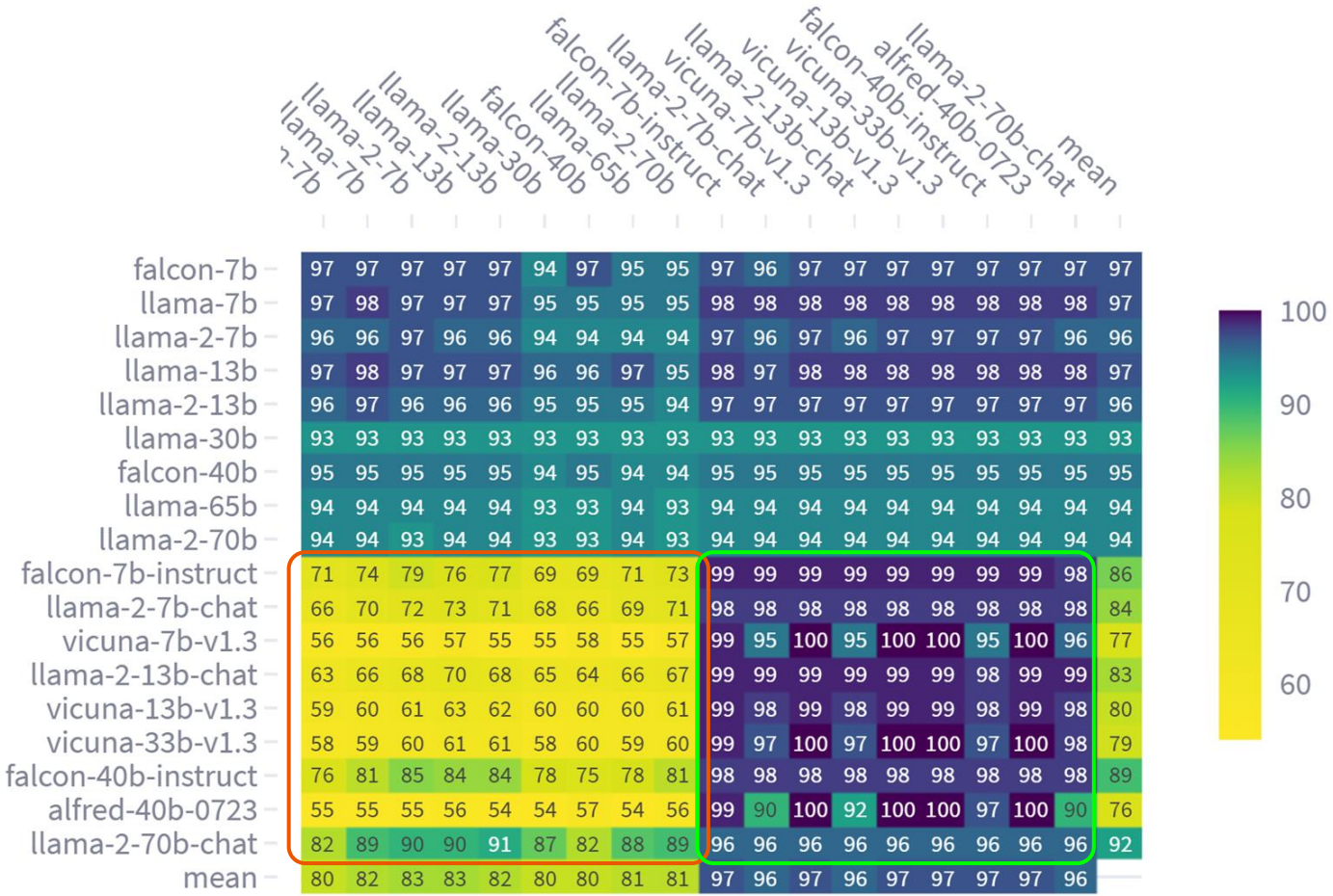
Results: Cross-Model Detection Results - Model Size Influence



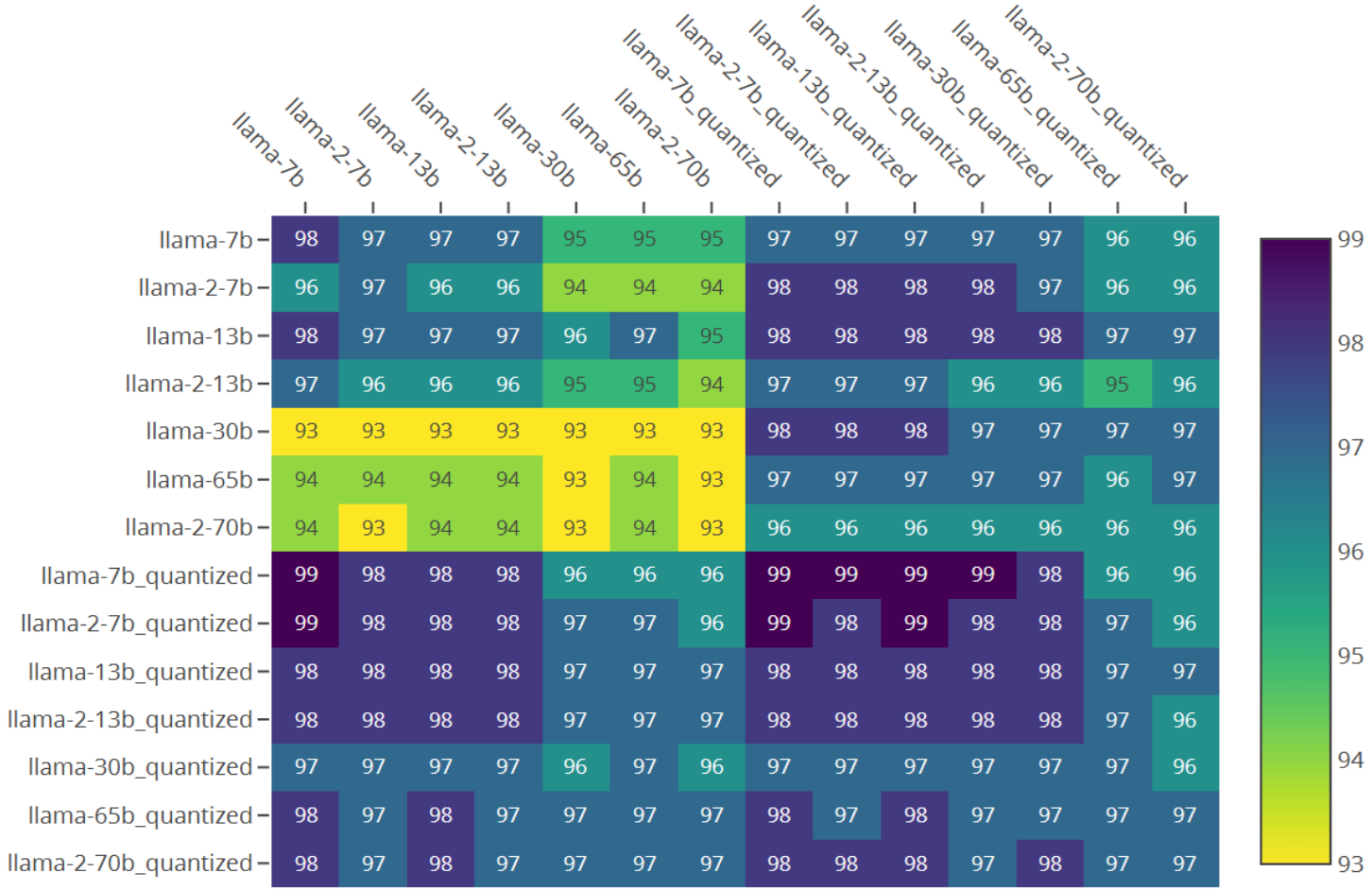
Results: Cross-Model Detection Results - Model Family Influence



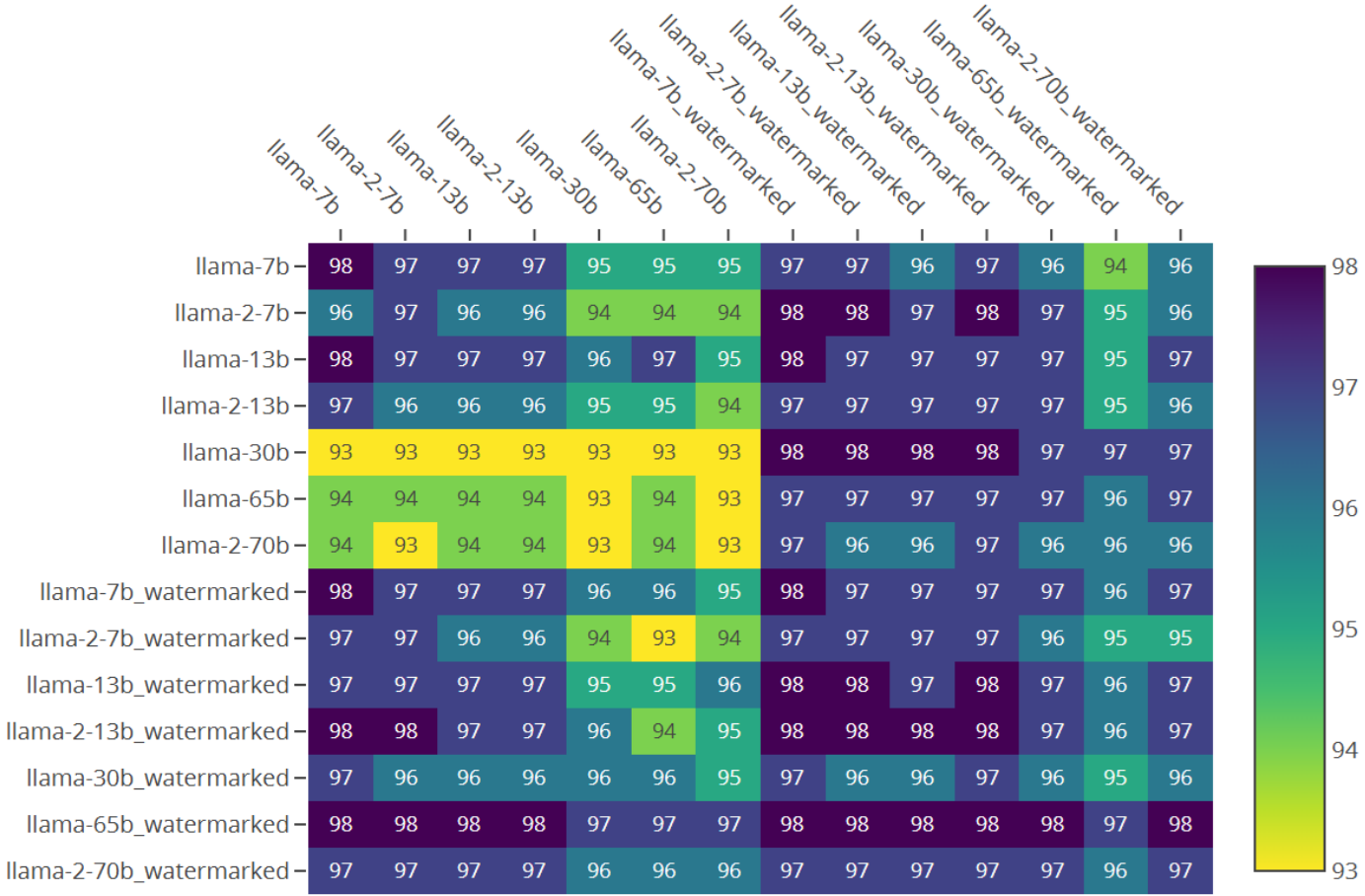
Results: Cross-Model Detection Results - Conversation FT



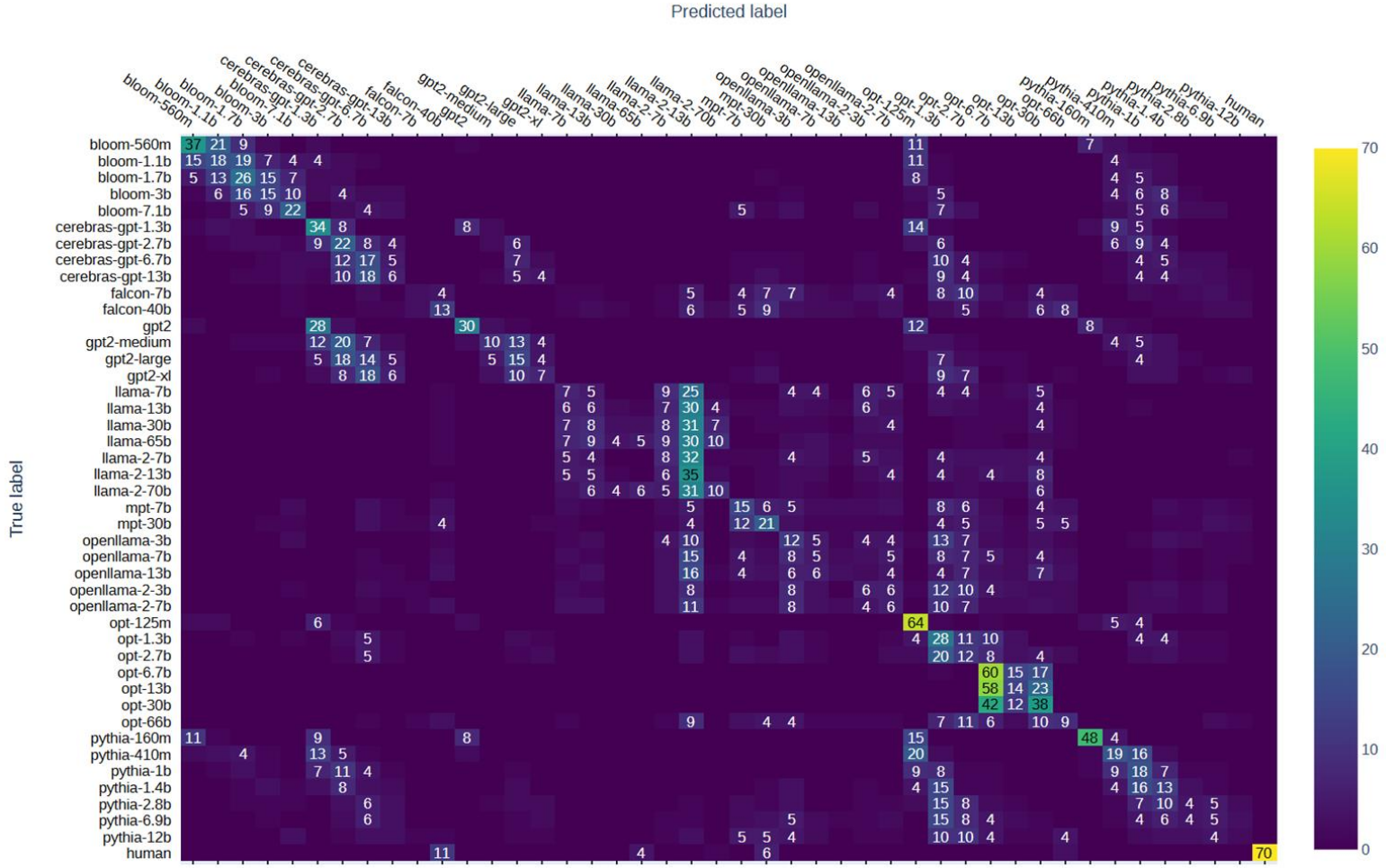
Results: Cross-Model Detection Results - Influence of Quantization



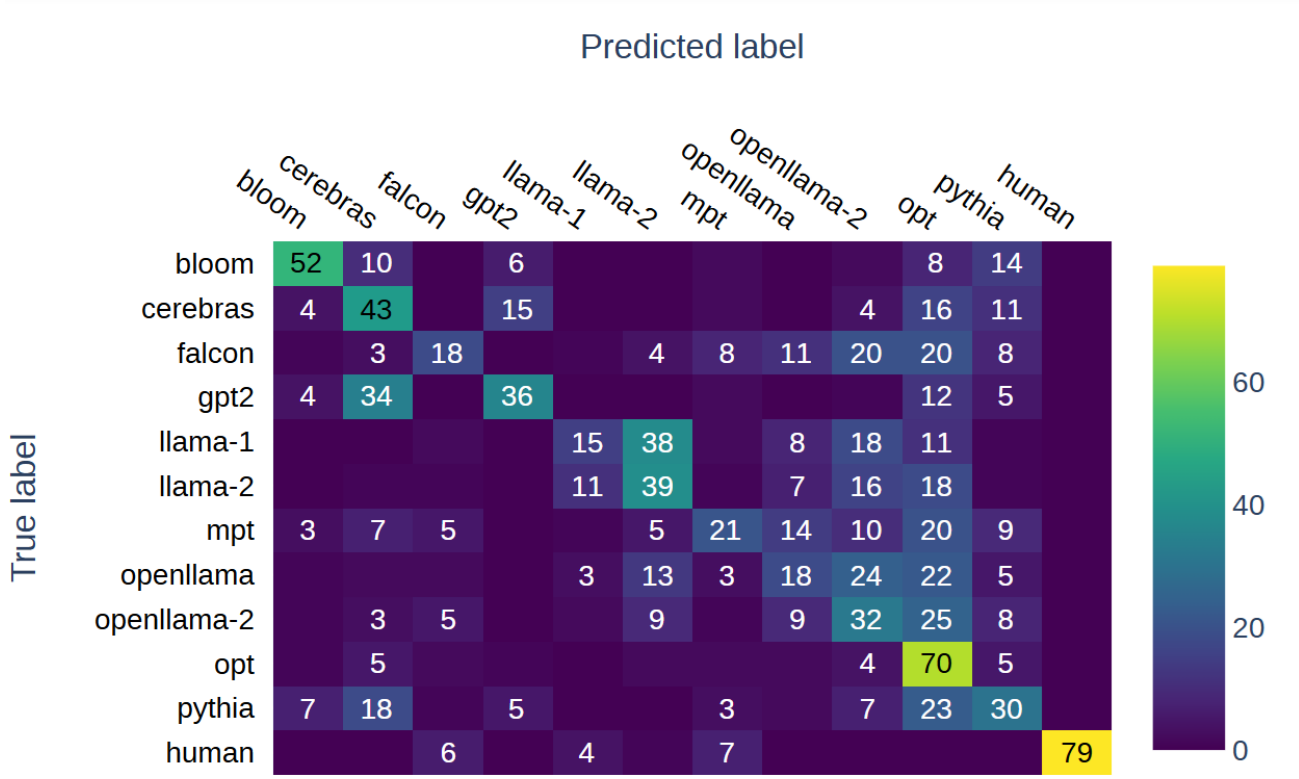
Results: Cross-Model Detection Results - Influence of watermarking



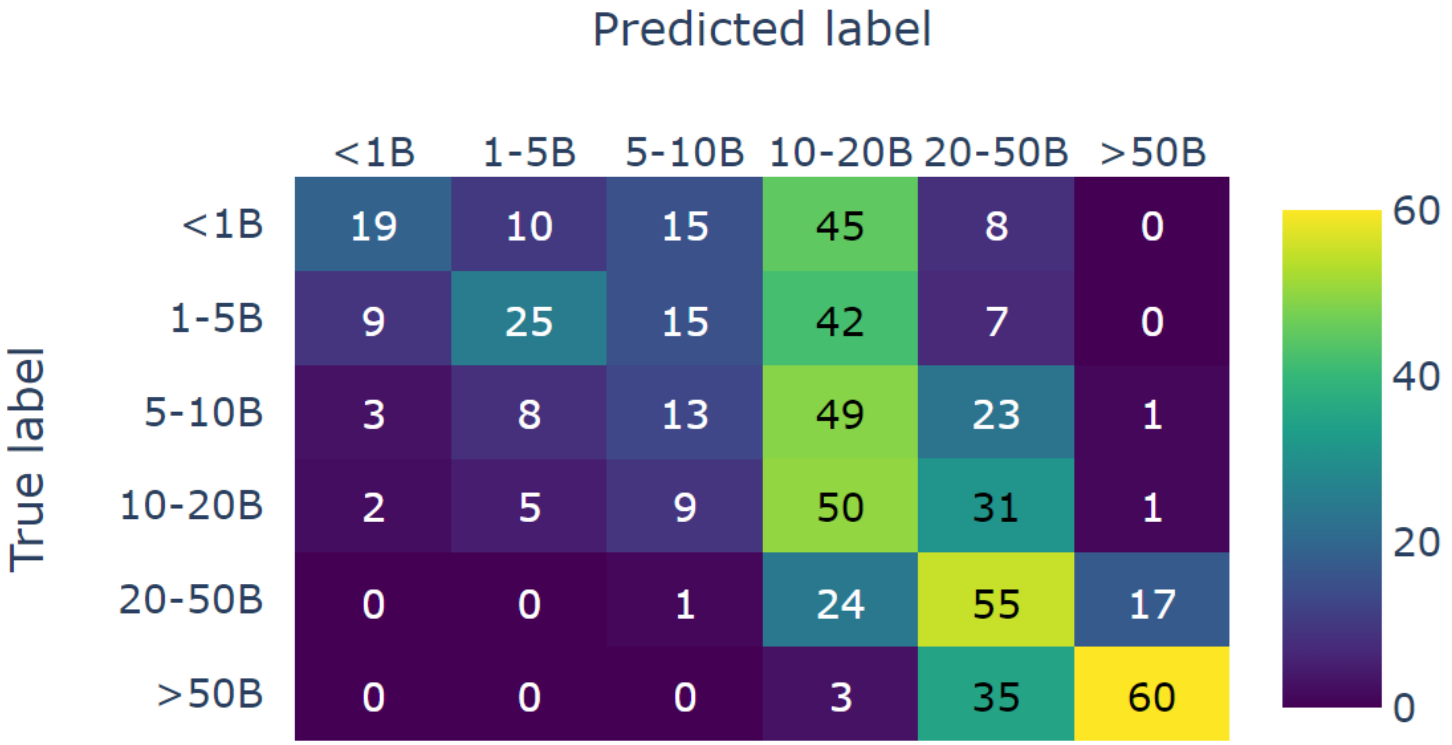
Results: Model Attribution - Source Model Identification



Results: Model Attribution - Model Family Classification



Results: Model Attribution - Model Size Classification



Results: Model Attribution Quantization and Watermarking Detection

- Quantization Detection:
  - Classifier accuracy: 54.5% (2 labels)
  - GPTQ method shows effectiveness without leaving discernible traces.

- Watermark Detection:
  - Classifier accuracy: 82.3%
  - Implication: Watermark signatures identifiable and disclosed through encoder classifier, without access to source model's log probabilities.

Results: Adversarial Content

Conclusion & Limitations

- Key Takeaways:
  - Conducted study in controlled environment to isolate variable influences.
  - Performance demonstrated not indicative of real-world expectations.
  - Envision detectability score as proxy for model quality evaluation.
  - Results highlight complex interplay of model size, family, and training data in LLM detection and attribution.
  - We provide all experiment results in interactive online repository: https://huggingface.co/spaces/wissamantoun/LLM_Detection_Attribution
- Limitations:
  - Did not explore impact of various sampling strategies or parameters like temperature.
  - Study focused only on openly available models, excluding black box models accessible only through APIs.
  - Classification technique constrained to fine-tuning a single model, potentially overlooking alternative approaches.

Thank you