The Foundation Model Transparency Index Search Protocol

1 Search protocol

In this section, we outline the search process we used to look for evidence that an FM developer satisfies our requirements for a given indicator.

1.1 General search process

1.1.1 Keyword Definitions

Each item under review has associated search keywords available on our GitHub repository (`https://github.com/stanford-crfm/fmti`).

1.1.2 Model-Item Pair Searches

For every model-item pair, we conduct a search using the defined keywords within the centralized resources associated with the respective models listed below.

1.1.3 Search Methodology

We employ the following format for every model-item-keyword tuple while using Google search, and read through the first 10 search results.

```
site:[Refer to developer's website list below] [Refer to model
    name list below] [Enter keyword]
```

For example, for GPT-4's energy efficiency item, the searches would be:

```
site:openai.com gpt-4 energy
site:openai.com gpt-4 efficien
```

1.1.4 Justification

We note the source (e.g., website, company blog post, paper) for each piece of evidence that helped confirm an item is present, alongside the justification. We link to an archive.org URL that contains the justification (instead of linking to developers' pages directly), to maintain records.

1.1.5 Avoid Search Personalization

To minimize the influence of personalized search results, we perform all searches in a private or incognito browser tab.

1.1.6 Determination Criteria

If we find one piece of evidence that fully justifies 1 point - or, in rarer cases, 0 points - for an item, we don't perform other searches.

1.1.7 Distribution Channels

In certain limited cases where the above steps fail to generate any information for indicators related to distribution channels, we interact with the developer's intended distribution channel (if disclosed), such as its API or its preferred deployment partner's API, or the documentation related to this API. We search for the required information via this distribution channel to the extent possible. We also use proxies, such as model playgrounds, if enterprise access is otherwise required.

1.2 Developer website

- AI21 Labs (Jurassic-2): `ai21.com`
- Amazon (Titan Text): `aws.amazon.com/bedrock/titan/`
- Anthropic (Claude): `anthropic.com`
- Cohere (Command): `cohere.com`
- Google (PaLM 2): `ai.google`
- Hugging Face (BLOOMZ): `bigscience.huggingface.co`
- Inflection (Inflection-1): `inflection.ai`
- Meta (Llama 2): `ai.meta.com`
- OpenAI (GPT-4): `openai.com`
- StabilityAI (Stable Diffusion 2): `stability.ai`

1.3 Centralized resources for all models

1.3.1 AI21 Labs (Jurassic-2)

- `https://docs.ai21.com/docs/jurassic-2-models`
- `https://docs.ai21.com/docs/responsible-use`
- `https://uploads-ssl.webflow.com/60fd4503684b466578c0d307/61138924626a6981ee09caf6_jurassic_tech_paper.pdf`
- `https://www.ai21.com/blog/introducing-j2`
- `https://docs.ai21.com/docs/responsible-use#usage-guidelines`
- `https://studio.ai21.com/terms-of-use`
- `https://studio.ai21.com/privacy-policy`
- `https://docs.ai21.com/changelog`

1.3.2 Amazon (Titan Text)

- https://aws.amazon.com/bedrock/titan/
- https://docs.aws.amazon.com/pdfs/bedrock/latest/APIReference/bedrock-api.pdf#API_ListFoundationModels
- https://aws.amazon.com/aup/

1.3.3 Anthropic (Claude 2)

- https://legal.anthropic.com/#aup
- https://vault.pactsafe.io/s/9f502c93-cb5c-4571-b205-1e479da61794/legal.html#aup
- https://console.anthropic.com/docs/api/supported-regions
- https://legal.anthropic.com/#terms
- https://legal.anthropic.com/#privacy
- https://docs.anthropic.com/claude/docs
- https://www.anthropic.com/index/claude-2
- https://www.anthropic.com/earlyaccess
- https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf
- https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

1.3.4 Cohere (Command)

- https://docs.cohere.com/docs/
- https://cohere.com/security
- https://dashboard.cohere.ai/playground/generate
- https://cohere.com/terms-of-use
- https://cloud.google.com/blog/products/ai-machine-learning/accelerating-language-model-training-with-cohere-and-google-cloud-tpus
- https://cohere.com/data-usage-policy
- https://cohere.com/privacy
- https://cohere-inc.secureframetrust.com/

1.3.5 Google (PaLM 2)

- https://ai.google/static/documents/palm2techreport.pdf
- https://developers.generativeai.google/models/language
- https://policies.google.com/terms/generative-ai/use-policy
- https://developers.generativeai.google/guide/safety_guidance
- https://developers.generativeai.google/products/palm
- https://developers.generativeai.google/available_regions
- https://developers.generativeai.google/terms#content_license_and_data_use

1.3.6 Hugging Face (BLOOMZ)

- https://arxiv.org/abs/2211.01786
- https://huggingface.co/docs/transformers/model_doc/bloom
- https://huggingface.co/bigscience/bloom
- https://arxiv.org/abs/2303.03915
- https://arxiv.org/abs/2211.05100
- https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf

1.3.7 Inflection (Inflection-1)

- https://inflection.ai/assets/Inflection-1.pdf
- https://inflection.ai/inflection-1
- https://inflection.ai/assets/MMLU-Examples.pdf
- https://heypi.com/policy#privacy
- https://inflection.ai/safety

1.3.8 Meta (Llama 2)

- https://arxiv.org/pdf/2307.09288.pdf
- https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md
- https://ai.meta.com/static-resource/responsible-use-guide/

1.3.9 OpenAI (GPT-4)

- https://openai.com/research/gpt-4
- https://openai.com/policies/usage-policies
- https://openai.com/form/chat-model-feedback
- https://platform.openai.com/docs
- https://openai.com/customer-stories
- https://status.openai.com/
- https://openai.com/policies/terms-of-use
- https://cdn.openai.com/policies/employee-data-privacy-notice.pdf
- https://cdn.openai.com/papers/gpt-4-system-card.pdf
- https://arxiv.org/pdf/2303.08774.pdf
- https://openai.com/research/triton
- https://openai.com/pricing
- https://platform.openai.com/docs/deprecations
- https://openai.com/waitlist/gpt-4-api
- https://openai.com/our-structure
- https://openai.com/api-data-privacy

1.3.10 StabilityAI (Stable Diffusion 2)

- https://huggingface.co/stabilityai/stable-diffusion-2
- https://openreview.net/forum?id=M3Y74vmsMcY
- https://huggingface.co/terms-of-service
- https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL
- https://platform.stability.ai/legal/terms-of-service
- https://stability.ai/use-policy