Naming Survey Protocol

***Title**: Naming Practices of Pre-Trained Models in Hugging Face*

***Recruitment**: In the study, we will recruit the users of Hugging Face PTMs, including PRO and normal accounts.*

***Compensation**: We will provide financial compensation to 3rd-party team participants. We will incentivize survey participants through a $10 gift card.*

Research Questions

Survey RQs:

- **RQ1**: How is PTM naming different from traditional soft- ware package naming?
- **RQ2**: What elements should be included in a PTM identifier?
- **RQ3**: How do engineers identify naming anomalies?

Discussion:

What **improvements** can be made to model registry infrastructures (e.g. Hugging Face), to enhance searchability and reuse of model names?

Questions

Demographic Questions

1. How many years have you worked on ML?
    a. 1 - 2 years
    b. 3 - 5 years
    c. 6 - 10 years
    d. 11 - 20 years
    e. > 20 years
2. How many years have you worked on SE?
    a. 1 - 2 years
    b. 3 - 5 years
    c. 6 - 10 years
    d. 11 - 20 years
    e. > 20 years
3. How would you rate your expertise in ML:
    a. **Novice**: I'm just starting out, and I usually get stuck in my machine learning projects and ask for help.
    b. **Intermediate**: I've done a few substantial machine learning projects and I usually can complete them without substantial assistance in my work.
    c. **Expert**: Other people often consult me for help on their machine learning projects.
4. How would you rate your expertise in SE:
    a. **Novice**: I'm just starting out, and I usually get stuck in my software engineering projects and ask for help.
    b. **Intermediate**: I've done a few substantial software engineering projects and I usually can complete them without substantial assistance in my work.
    c. **Expert**: Other people often consult me for help on their software engineering projects.
5. What is the size of your organization?
    a. Small (1 - 50 employees)
    b. Medium (51 - 250 employees)
    c. Large (251 - 1000 employees)
    d. Very large (1001+ employees)
6. What is the type of your organization? [https://www.ctc.ca.gov/credentials/leaflets/industry-sectors-chart]
    a. Arts, Media, and Entertainment
    b. Business and Finance
    c. Education

    d. Energy, environment, and utilities
    e. Health science and Medical technology
    f. Information and communication technology
    g. Manufacturing and product development
    h. Marketing, sales, and services
    i. Public services
    j. Transportation
    k. Other (text box)

7. What deployment contexts do you work on?
    a. Web application
    b. Desktop
    c. Cloud and data center
    d. IoT/embedded systems
    e. Mobile devices
8. What requirements do you consider when working with PTMs? [ICSE'23, Ozturk et al. ICML'22, TODO]
    a. Memory consumption
    b. Latency
    c. Accuracy
    d. License
    e. Carbon emitted
    f. Documentation
    g. Task compatibility
    h. Interpretability
    i. Reproducibility
    j. Others *(text box)*
9. What kind of model have you used before? (*The model tasks are based on Hugging Face categorization system.*)
    a. Multimodal
        i. Feature Extraction
        ii. Text-to-Image
        iii. Image-to-Text
        iv. Image-to-Video
        v. Text-to-Video
        vi. Visual Question Answering
        vii. Document Question Answering
        viii. Graph Machine Learning
        ix. Text-to-3D
        x. Image-to-3D
        xi. Others (text box)

b. Computer Vision (CV)

    i. Depth Estimation
    ii. Image Classification
    iii. Object Detection
    iv. Image Segmentation
    v. Image-to-Image
    vi. Unconditional Image Generation
    vii. Video Classification
    viii. Mask Generation

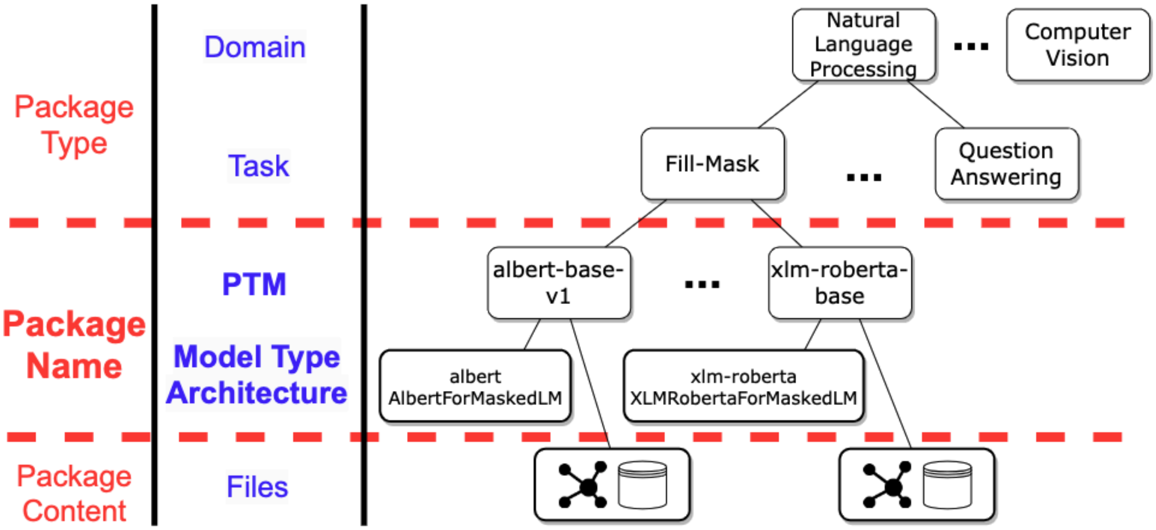
        ix. Others (text box)
    c. Natural language processing (NLP)
        i. Text Classification
        ii. Token Classification
        iii. Question Answering
        iv. Translation
        v. Summarization
        vi. Conversational
        vii. Text Generation
        viii. Text2Text Generation
        ix. Fill-Mask
        x. Sentence Similarity
        xi. Others (text box)
    d. Audio
        i. Text-to-Speech
        ii. Text-to-Audio
        iii. Automatic Speech Recognition
        iv. Audio-to-Audio
        v. Audio Classification
        vi. Voice Activity Detection
        vii. Others (text box)
    e. Reinforcement Learning (RL)
    f. Tabular
    g. Others (*text box*)

10. How many pre-trained model (PTM) packages have you used from model registries: (e.g. *Hugging Face, Pytorch hub):*
    a. 0
    b. 1 - 5
    c. 5 - 10
    d. 10 - 20
    e. > 20
11. Check all the forms of reuse you have done: *checkboxes*
    a. No modification
        i. Download and use without any modification
    b. Fine-tuning approaches
        i. Fully-finetuned, without additional changes
        ii. Fully-finetuned and extended (e.g. adding heads or layers)

        iii. Partially-finetuned (i.e. only parts of the model were fine-tuned)
    c. Parameter efficient tuning
        i. Prompt tuning
        ii. LoRA
    d. Advanced adaptation techniques
        i. Knowledge distillation
        ii. Few-shot learning
    e. Others: *text box*

12. How many PTM packages have you created (contribute + give names) in model registries (e.g. Hugging Face, Pytorch hub):
    a. 0
    b. 1 - 5
    c. 5 - 10
    d. 10 - 20
    e. > 20
13. (if answer to Q11 is not 0) Do you follow naming conventions when creating PTM packages?
    a. Yes (Organization practice): Please describe
    b. Yes (Personal practice): Please describe
    c. No
14. How many traditional packages have you used from software registries (e.g. NPM, PyPI, Nuget, Maven)?
    a. 0
    b. 1 - 5
    c. 5 - 10
    d. 10 - 20
    e. > 20
15. How many traditional packages have you created (contribute + give names) on software registries (e.g. NPM, PyPI, Nuget)?
    a. 0
    b. 1 - 5
    c. 5 - 10
    d. 10 - 20
    e. > 20
16. (If answer to Q13 is not 0) Do you follow any naming conventions when creating traditional packages?
    a. Yes (Organization practice): Please describe
    b. Yes (Personal practice): Please describe
    c. No

(Definition of PTM Naming Provided here)



The user-controllable names are PTM identifier, model type and architecture. In this study, we define a PTM **package name** as the combination of a package ***identifier*** (e.g., albert-base-v2, facebook-llama/Llama-2-7b-chat-hf) and the ***model type*** or ***architecture*** indicated in the metadata (e.g., albert, AlbertForMaskedLM)

Naming practices

1. (RQ1a) When selecting or searching for PTMs from model registries, how does your search use PTM names (remember, that means the stated *architecture* + *identifier*)?
    a. Text box
2. (RQ1a) Which naming convention do you prefer when reusing PTM from model registries like Hugging Face? [Henninger 1994] (*multi-checkbox*)
    a. Named by task: high-level category (e.g. question-answering)
    b. Named by application: what a PTM does (e.g. fake-news-detector, text2image-prompt-generator)
    c. Named by implementation: what a PTM is (e.g. bert-base-uncased)
    d. Named by "Implementation + task" (e.g. *Llama-2-7b-chat-hf*)
    e. Named by "Implementation + application" (e.g. *distilroberta-base-finetuned-fake-news-detection*)
    f. Others (text box)

3. (RQ1a) Please evaluate the importance of the following **factors** in the naming of PTMs. For each factor, indicate your level of agreement with its importance [Jiang et.al (ICSE’23, MSR’24)]: (for each) *strongly agree, agree, disagree, strongly disagree*
    a. Architecture (e.g. xxx):
    b. Model size
    c. Dataset
    d. Model versioning
    e. Language
    f. Task
    g. Adaptation method
    h. Training regime
    i. Application goal
    j. Number of (hidden) layers (e.g. L-12, H-128)
    k. Number of parameters
    l. Dataset characteristics
    m. Others (please specify)
4. (RQ1b) In the context of adapting machine learning models for improved performance or specific needs, at what point should modifications (e.g., changes to architecture, training regime, or dataset) necessitate a new name for the model?
    a. Input/Output layers
        i. Modified tensor shape in Input/Output layers
        ii. Addition/deletion of layers in Input/Output layers
    b. Main body of the architecture
        i. Modified layers in the main body of the architecture (e.g. dropout rates, activation functions)
        ii. Addition/deletion of layers in the main body of the architecture
    c. Modified trianing regime of the PTM
    d. Changed training dataset
    e. Other (*text box*)

Naming Challenges

5. (RQ2b) In your experience, do the PTMs available in model registries accurately describe their behavior/content? What discrepancies have you experienced? Please explain [Montes et al. FSE’22, ICSE’23]
    a. Yes (short answer for each attribute)
        i. Architecture (e.g. bert, resnet):
        ii. Model size (e.g. base, large, 50, 101)

        iii. Dataset (e.g. squad, imagenet)
        iv. Model versioning (e.g. v1, v2)
        v. Language (e.g. en, English, Arabic)
        vi. Task (e.g. question-answering, qa)
        vii. Adaptation method (e.g. finetine, distill, fewshot)
        viii. Training regime (e.g. pretrain, sparse)
        ix. Number of (hidden) layers (e.g. L-12, H-128)
        x. Number of parameters (e.g. 100M, 8B)
        xi. Dataset characteristics (e.g. case, uncased, 1024-1024)
        xii. Others (please specify)
    b. No (Which component/factor could be inaccurate based on your experience?)

6. Do you think you would notice if the following PTM naming elements are wrong: impossible - neutral - possible - very possible
    a. Architecture
    b. Model size
    c. Dataset
    d. Model versioning
    e. Lagnauge
    f. Task
    g. Reuse method
    h. Training process
    i. Number of (hidden) layers (e.g. L-12, H-128)
    j. Number of parameters
    k. Dataset characteristics
7. (RQ2c) Do you have any privacy or security [Gu et.al. S&P'23, SpellBound arXiv'20, Kaplan et.al. MLHat'21] concerns relevant to PTM names when you are reusing them?:
    a. Traditional software package attacks (e.g. Squatting, Vulnerabilities)
        i. *Text box*
    b. DL-specific attacks (e.g. backdoor/trojan attack, posoning attack, data stealing attack)
        i. *Text box*
8. Do you think the level of risk of these threats will change over time? Why?
    a. Yes *(text box)*
    b. No *(text box)*

Comparison to Trad. Software

9. (RQ3) Should the PTM package naming convention be different from traditional package naming conventions?
   a. Yes (Why?: Text box)
   b. No (Why?: Text box)
   c. Maybe/Sometimes (Why?: Text box)
10. (RQ3) How is PTM naming similar/different from traditional software naming? Why do you think that is? (*text box*)
11. (RQ3) Assess the importance of establishing naming conventions for PTMs as compared to traditional software packages. For each aspect of naming standards listed below, please indicate your level of agreement with its importance: [Alsuhaibani et.al. (ICSE'21)]:*strongly agree, agree, disagree, strongly disagree. (side by side comparison between PTM and trad. package)*
    a. Naming style
    b. Grammatical structure
    c. Verb phrase
    d. Dictionary terms
    e. Full words (not abbreviations or acronyms)
    f. Idioms and slang
    g. Prefix/suffix
    h. Length

Improvement

12. What improvements can model registry (e.g. Hugging Face) make, to enhance ***searchability and reuse*** of models (e.g. registry infrastructure, naming linter, integration of model cards)?:
    a. *Text box*
13. (Optional) Is there any additional feedback or specific suggestions you have regarding the PTM naming that was not covered in the previous questions? Please specify.
    a. Text box

ONNX Questions

1. Which framework do you use for model development?
    - ○ PyTorch
    - ○ TensorFlow
    - ○ JAX/FLAX
    - ○ MLX
    - ○ Other
2. Do you use ONNX as part of your model development and deployment process?
3. Are there other interoperability tools that you use, if so which ones?
    - ○ MMdnn
    - ○ NNEF
    - ○ Other
4. For what purposes do you use ONNX?
    - ○ Framework-to-framework Model Conversion (e.g., converting from a model from TensorFlow to PyTorch)
    - ○ Model Conversion for Deployment (e.g., converting to ONNX for deployment using ONNXRuntime or TensorRT.
    - ○ Other (please specify)
5. Do you ever deploy directly from a deep learning framework such as PyTorch or TensorFlow? What do you consider when choosing between deploying from a DL framework vs. via ONNX?
6. How often do you encounter the following problems while using ONNX models? (Likert scale, with "Never/Rarely/Occasionally/Regularly)
    - ○ Crashes (e.g., Model does not convert to ONNX.)
    - ○ Performance Differences (e.g., the accuracy of the ONNX model does not match the original model)
    - ○ Other (please specify)
7. When you encounter one of these problems, how do you address it?

Recruitment message

Subject: Survey on Machine Learning Model Naming Conventions

Content:

Dear Hugging Face contributor,

I hope this email finds you well. I am studying engineering practices when re-using pre-trained models (PTMs).

I would appreciate your help in collecting some data. I am conducting a survey to understand engineers' perspectives on PTM naming conventions and on interoperability tools.

- My survey takes ~5 minutes. All data will be anonymized to ensure privacy.
- You would be compensated for your time with a $10 Amazon gift card.

Here is the link to the survey: [link]

If you have colleagues who might have relevant opinions, please feel free to forward to them.

This study is supported by the ANONYMOUS SOURCE, and has been approved by our institution's IRB (#xxxx-xxx).

Questions? Contact me (Author [email]).

Thank you for your consideration.

Best regards,

Author