

# aiDAPTIV+ Pro Suite 2.0 User guide

Version 1.9

**Phison Electronics Corporation** 

Tel: +886-37-586-896 Fax: +886-37-587-868 E-mail: sales@phison.com / suppport@phison.com

ALL RIGHTS ARE STRICTLY RESERVED. ANY PORTION OF THIS PAPER SHALL NOT BE REPRODUCED, COPIED, OR TRANSLATED TO ANY OTHER FORMS WITHOUT PERMISSION FROM PHISON ELECTRONICS CORPORATION.



Phison may make changes to specifications and product description at any time without notice. PHISON and the Phison logo are trademarks of Phison Electronics Corporation, registered in the United States and other countries. Products and specifications discussed herein are for reference purposes only. Copies of documents which include information of part number or ordering number, or other materials may be obtained by emailing us at sales@phison.com or support@phison.com.

©2024 Phison Electronics Corp. All Rights Reserved.



Revision	Draft Date	History	Pro Suite Version	Author	
0.1	2024/10/09	Preliminary release	NXUN_2.0.0J	Sean Liou	
0.1	2024/10/05		(beta version)	Sean Liou	
0.2	2024/10/11	Correct type in p 11 and p 18	NXUN_2.0.0J	Sean Liou	
0.2	2024/10/11		(beta version)	Searr Liou	
		1.Modify p.8 OS version			
1.0	2024/11/15	2.Add example of permission setting on p.36~37	NXUN_2.0.0	Sean Liou	
		3. Add Appendix A and B			
1.1	2024/11/21	Correct typo.	NXUN_2.0.0	Sean Liou	
1.2	2024/12/6	Update Section 3.1	NXUN_2.0.1	Sean Liou	
1.3	2024/12/10	Update Section 4.1	NXUN_2.0.1	Sean Liou	
1.4	2024/12/16	Update Section 4.1.1,& 4.1.2	NXUN_2.0.1	Sean Liou	
1 5	2024/12/25	1.Add Appendix C	NXUN_2.0.2	Soon Liou	
1.5		2.Update Pro Suite verison	NOUN_2.0.2	Seari Liou	
1.6 20	2025/01/10	1.Update Appendix A	NXUN_2.0.2	Soon Liou	
1.0		2.Table 3-1, Table 3-2	NOUN_2.0.2	Seari Liou	
17	2025/01/24		1/24 Undata Pro Suita varsian	NXUN_2.0.3	Soon Liou
1.7		opdate Plo Suite version	NOUN_2.0.3	Sean Liou	
10	2025/02/40			Soon Liou	
1.0	2023/02/18		NOUN_2.0.3	Seall LIOU	
		1. Update Pro Suite version			
1.9	2025/03/24	2. Update Section 3.1.1.1, 3.2.2, 3.7.2.2 and 4.1		Sean Liou	
		3. Update Appendix A	100010_2.0.5		

## **REVISION HISTORY**



## TABLE OF CONTENTS

RE\	/ISIO	N HISTC	<b></b>	
TAE	BLE O	F CONT	NTS	4
LIS	T OF F	FIGURES		
LIS	r of 1	TABLES.		7
1.	EN	VIRONN	ENT PREPARATION	8
	1.1.		upported OS and Nvidi	a driver version
	1.2.		rowser suggestion and	precaution
2.	DES	SCRIPTI	N	
3.	FUI	NCTION	NTRODUCTION	9
	3.1.		ataset	
		3.1.1.	Upload	
		3.1.1.1	Dataset upload	
		3.1.1.2	Dataset Managem	ent
		3.1.1.3	Example of datase	file format
		3.1.2.	aiDAPTIVGuru	
		3.1.2.2	Parameter setting	& file upload
		3.1.2.2	Confirm data pre-p	rocessing results
		3.1.2.3	Generate Dataset.	
	3.2.		ine-tune	
		3.2.1.	Hardware specifico	ition preview
		3.2.2.	Parameter setting	
		3.2.3.	Confirm hardware	configuration and parameter for fine-tuning
	3.3.		Ionitor	
		3.3.1.	Cancel job	
		3.3.2.	Remove job	
	3.4.		alidation	
		3.4.1.	Put questions	
		3.4.2.	Compare result	
	3.5.		enchmark (Option)	
		3.5.1.	Score	

		3.5.2.	Scoring Progress	24
		3.5.3.	Data	25
		3.5.4.	Chart	27
	3.6.	Infe	rence	28
		3.6.1.	Chat	28
		3.6.2.	RAG	30
		3.6.2.1.	Upload new collection	31
		3.6.2.2.	Recommended usage – using with aiDAPTIVGuru	31
	3.7.	Мо	dels	32
		3.7.1.	Model upload	32
		3.7.2.	Model list	33
		3.7.2.1.	Enable model	34
		3.7.2.2.	Set model Inference parameters	34
		3.7.2.3.	Pin the resident inference model	35
		3.7.2.4.	Quantized model	36
	3.8.	Mar	nagement	37
		3.8.1.	Authorization	37
		3.8.1.1.	Features	37
		3.8.1.2.	Roles	38
		3.8.1.3.	Users	39
		3.8.1.3.1.	Create Account	39
4.	APF	LICATION .		40
	4.1.	aiDA	APTIVInbox (Option)	40
		4.1.1.	EWS (Exchange Web Services)	41
		4.1.2.	SMTP (Simple Mail Transfer Protocol)	41
APP	ENDI	X A – MOD	EL AVL FOR FINE-TUNE	42
APP	ENDI	X B – RECO	MMENDED CONFIGURATION	43
APP	ENDI	X C – INBO	X MAIL SERVER TEST	44
	C.1	Pred	cautions before testing	44
	C.2	Exec	cute test script	44
	C.3	Test	result	45

P

5

0



## LIST OF FIGURES

Figure 3-1 Pro Suite main function	9
Figure 3-2 Dataset	. 10
Figure 3-3 Dataset upload	. 10
Figure 3-4 Dataset management	. 11
Figure 3-5 aiDAPTIVGuru pre-processed file management	. 14
Figure 3-6 Generate dataset	. 15
Figure 3-7 Dataset generation progress	. 15
Figure 3-8 Hardware specification preview	. 16
Figure 3-9 Parameters setting for fine-tuning	. 17
Figure 3-10 Final confirmation	. 18
Figure 3-11 Monitor	. 19
Figure 3-12 Cancel job	. 20
Figure 3-13 Remove job	. 20
Figure 3-14 Setting of question	. 21
Figure 3-15 View result of question	. 22
Figure 3-16 Compare results from different models	. 22
Figure 3-17 Parameter setting	. 23
Figure 3-18 Scoring progress	. 24
Figure 3-19 Scoring completed	. 24
Figure 3-20 Scoring data	. 26
Figure 3-21 Bar chart	. 27
Figure 3-22 Model and parameter information	. 27
Figure 3-23 Detail of scoring content	. 27
Figure 3-24 Chat room	. 29
Figure 3-25 RAG	. 30
Figure 3-26 Collection management	. 31
Figure 3-27 Model upload-method 2	. 32
Figure 3-28 Model list description	. 33
Figure 3-29 Inference parameters setting	. 34
Figure 3-30 Pin model failed	. 35

PERSON	ISON	PH
--------	------	----

Figure 3-31 Pin model failed error log	35
Figure 3-32 Setting of model quantization	36
Figure 3-33 Cancel model quantization	36
Figure 3-34 Feature setting of role	37
Figure 3-35 Role management	38
Figure 3-36 User management	39
Figure 3-37 Create account	40
Figure 4-1 EWS Setting	41
Figure 4-2 SMTP Setting	42

## LIST OF TABLES

13
30
34
38
42
43
45



## **1. ENVIRONMENT PREPARATION**

#### 1.1. Supported OS and Nvidia driver version

Category	Detail
OS	Ubuntu 22.04 LTS Desktop
GPU driver	Nvidia driver version 550 or later version

#### **1.2.** Browser suggestion and precaution

- Google Chrome (The recommended default browser for use with the Pro Suite service.)
- Mozilla Firefox

Note : When logging in for the first time, you may log in with the following account

Default system administrator account password Account: admin@aidaptiv.com Password: Admin8299

## 2. **DESCRIPTION**

The aiDAPTIV+ Pro Suite is a web-based GUI program that enables a **No Code** approach to model training. It streamlines the entire process from **Dataset generation**, **Fine-Tuning**, **Validation** to **Inference**. This allows users to quickly convert documents into files that can be used for training their own fine-tuned models, and build their own AI models.



Users can access Pro Suite main functions through the tabs at the top of the webpage. Below are detailed instructions for each function.

1. Dataset

D

- aiDAPTIVGuru
- 2. Fine-tune
- 3. Monitor
- 4. Validation
- 5. Benchmark (Option)
- 6. Inference
- 7. Models
- 8. Management



Figure 3-1 Pro Suite main function

#### 3.1. Dataset

There're 2 main functions in the Dataset tab: Upload and aiDAPTIVGuru. The Upload function allows users to upload an existing dataset to Pro Suite and manage the uploaded datasets. After clicking the Upload tab, users will see the page below. This page is divided into the upload area on the left and list area on the right.



Figure 3-2 Dataset

#### 3.1.1. Upload

#### 3.1.1.1. Dataset upload

- Field description:
  - File upload location : Support JSON, JSONL and Parquet file format. (Upload one file at a time)
  - Function description:
    - o Upload
    - **m** : Remove temporary files from the storage area

Import Dataset						
Drop file here or click to upload JSON, JSONL and Parquet files are supported, and JSON file must comply with one of the following formats. After you upload, the key in the file will be converted to "question", "cot_answer". *JSONL and Parquet files will generate JSON file.						
1. Instruction Tuning:         - Data file name should not contain "pretrain".         - 1. key: "question", "cot_answer"         - 2. key: "instruct", "output"         [         {         "question": "",         "cot_answer": "",         }.	<ul> <li>2. Pretrain:</li> <li>Data file name should contain "pretrain". ex: pretrain.json</li> <li>key: "text"</li> <li>[ <ul> <li>text": "",</li> <li>,</li> <li></li> </ul> </li> </ul>					
Upload						

Figure 3-3 Dataset upload

- JSONL and Parquet dataset upload
  - Step1: Upload JSONL or Parquet dataset.
  - Step2: Select the corresponding Key values in the Question and Answer fields.

Quantization				
Question:				
prompt				
Answer				
prompt				
			No	Yes

#### 3.1.1.2. Dataset Management

- Function description:
- 1. 💽 : View dataset content
- 2. 🛃 : Download dataset
- 3. 💼 : Delete dataset

Dataset List				
Name	Size	Туре	Upload Time	
sample_instruction_data_1k_pu blic.json	311.28 KB	User	2024/07/10 19:31:37	⊙ ± 前

Figure 3-4 Dataset management



#### 3.1.1.3. Example of dataset file format

- Instruction Tuning:
  - Data file name **Should NOT** contain "pretrain".
  - Key: "instruct", "output"

```
[
   {
        "instruct": "What is t
```

```
"instruct": "What is the more nutrient food in the convenience store?",
    "output": "I think that it might be a big ol chocolate bar."
},
{
```

```
"instruct": "Where could I get the best Italian food in town?",
    "output": "In my neighborhood, the food truck right next to the cross street."
}
]
```

#### • Pretrain:

- The data file name **Should** contain "pretrain". (For example: pretrain.json)
- Key: "text"

[

{
 "text": "Regular exercise and a balanced diet are important for maintaining
good health."

},

{

}

"text": "Drinking enough water and getting an adequate amount of sleep can contribute to overall well-being."

]

#### 3.1.2. aiDAPTIVGuru

Dataset preparation is a very labor-intensive process. aiDAPTIVGuru is a feature of Pro Suite that enhances the dataset generation. It transforms user-provided documents or files with domain-specific knowledge (such as product manuals, technical documents, specifications...etc) into Q&A sets and will automatically create a training dataset.

For the best usage of aiDAPTIVGuru, please refer to <u>Section 3.6.2.2</u>

# Item aiDAPTIVGuru\_Entry aiDAPTIVGuru\_Pro (Option) File Format pdf \ docx \ txt pdf \ docx \ txt Supported Model Llama-3.1-8B-Instruct Llama-3.1-8B-Instruct Upload multiple files of same/different Y Y

#### Table 3-1 Specification of aiDAPTIVGuru

**Note**: aiDAPTIVGuru\_Pro is an additional value-added service. For more information on enabling this service, please contact Phison's Sales account.



aiDAPTIVGuru parameter settings and document upload.

- Field description:
- 1. Dataset Name
- 2. Model: The model needs to be pinned first in order to appear in the model list.
- 3. Embedding model: Retrieval model.
- 4. **QA Pairs Count**: Number of Instruction Dataset data.
- 5. **Training QA ratio(%)**: The proportion of data in the data set that is used as training data (the remaining proportion is used as scoring data).
- 6. **Chunk Size**: Split size of the data file.
- 7. **Overlap**: The amount of data overlap when the data file is divided into chunks.
- 8. **Number of reference chunk per question**: Number of reference Chunk for each data instruction.
- 9. **Chunk Shuffle**: Mix the data chunks or distribute the data evenly.
- Language Evaluation: After activation, the content generated by the Dataset through Guru will refer to the original document's language. Only supports zh-TW, zh-CN, en-US.
   Note: Please upgrade to the aiDAPTIVGuru\_pro version to support this feature.
- 11. Upload File Area: Domain file upload area when generating a dataset.
- Function description:
  - 1. **Remove all**: Remove temporary files.
  - 2. **Upload**: Upload the file for pre-processing.

#### 3.1.2.2. Confirm data pre-processing results

Confirm Inspect the data pre-processing results. The user can edit and adjust data online.

- Function description:
- 👩 : View the pre-processed txt file content and edit online.
- **I** : Download the temporary txt document.
- 💼 : Delete the temporary txt document.
- Remove all: Delete all temporary txt files.

± Upload	$\rangle$	E G	enerate
I	Drop file here or click to uploa	ad	
Only pdf, csv, xlsx, docx, pptx, txt formats an	e allowed.		
Tx1 Uploaded TXT files			
Phison aiDAPTIV ProSuite 2.0_User gui	de_Preliminary v0.3.txt		⊙ ± ā
	Remove all		

Figure 3-5 aiDAPTIVGuru pre-processed file management



#### 3.1.2.3. Generate Dataset

Execute aiDAPTIVGuru.

- Function description:
  - **Generate :** After confirming the pre-processing results, click "Generate".

Ø Parameter			> 🔒 Generate
Dataset Name			
Input dataset name			$\sim$
Model			$(\uparrow)$
Select Model		Drop file	here or click to upload
Embedding model		Only DDF TYT and DOCY file formats are allowed	for unload
multilingual-e5-large		S1-Phison integration Phase 1.pdf	
QA Pairs Count	- 2000 +	l	n Remove all
	•		
Training QA ratio (%)	- 80 +		
Chunk Size	- 256 +		
	- 230 +		
Overlap			
•			
Number of reference chunk per question			
-•			
Chunk Shuffle 🗹 Language Evaluation 🔳			
		Upload	

Figure 3-6 Generate dataset

• Cancel

Monitor the dataset generation progress. Pressing "Cancel" will terminate the operation.

<u></u>		📑 Generate	
			Cancel
	10%		
(2 / 20) Generating	gPhison Nand Flash Spec.	txtnow2total 20	

Figure 3-7 Dataset generation progress

#### 3.2. Fine-tune

Until aiDAPTIV+, small and medium-sized businesses have been limited to small, imprecise training models with the ability to scale beyond Llama-2 7b.

Phison's aiDAPTIV+ solution enables the training of significantly larger models, giving you the opportunity to run workloads previously reserved for data-centers.

Pro Suite's fine-tune feature is integrated with Phison's aiDAPTIV+ technology, reducing hardware resources.

The function will be divided into three stages, hardware specification preview, parameter setting and final confirmation.

- Number of GPUs = 2<sup>n</sup> (n=0,1,2,3,4, GPUs = 1,2,4,8)
- When selecting the number of GPUs, make sure there are enough GPU resources to perform the finetuning.
- Please refer to <u>Appendix A</u> for model support list.
- Please refer to <u>Appendix B</u> for recommended hardware configuration.

#### 3.2.1. Hardware specification preview

System hardware configuration (GPU, VRAM, system memory, aiDAPTIVLink, aiDAPTIVCache, OS...).

Item	Information
GPU	1-NVIDIA RTX 4000 Ada Generation 2-NVIDIA RTX 4000 Ada Generation 3-NVIDIA RTX 4000 Ada Generation 4-NVIDIA RTX 4000 Ada Generation
GPU Count	4
VRAM	80 GB
System Memory	503 GB
aiDAPTIVLink	aidaptiv:vNXUN_2_01_00
aiDAPTIVCache : life remaining	/dev/nvme0n1 (1907.73GB) : 100.00% /dev/nvme1n1 (1907.73GB) : 100.00%
OS	Ubuntu 22.04.4 LTS
OS Disk	439 GB

Figure 3-8 Hardware specification preview

#### 3.2.2. Parameter setting

- Field description:
  - 1. **Model** : Select the model to fine-tune. (Only Pre-training and Fine-tune models will be displayed in the list. AWQ quantified models will not be included in this list.)

Models					
Name	State	Туре	Create Time	Available	
Meta-Llama-3.1-8B-Instruct-gp_64-bit_4-AWQ		Pre_Training_AWQ	2024/11/12 22:09:53		<b>\$</b>

Note: The model needs to be available first in order to appear in the model list.

- 2. Dataset: Select the dataset for fine-tuning.
- 3. Available GPU: GPU model and number to be used for training
- 4. **Epoch**: The number of epochs to train the model. (Range 1 ~ 5, default=1)
- 5. **Per Device Train Batch Size**: Batch size for each GPU.
- 6. Per Update Total Batch Size: Set the total batch size for one update. For example, if you are running on 4 GPUs with per\_device\_train\_batch\_size=4 and want to update the model every 80 batches, then you should set the per\_update\_total\_batch\_size to 80. The machine will run 80/4/4 = 5 iterations and update the model once. If not divisible, round up to the next whole number.
- Max Seq Length: Define the maximum sequence length.
   Note: Click the *Advice* button to automatically calculate the appropriate Max Seq Length value.
- 8. Learning Rate: Set the learning rate.
- 9. **Triton**: Trigger triton training procedure. It can shorten the model training time. (Please refer to Appendix A for the applicable model list.)

**Note**: If the user selects a model that does not support Triton for training, the following error message will appear after the training begins: "Phison Accelerator does not support," and the training process will be terminated.

- 10. Job Name: Allow users to identify different training tasks.
- Function description:
  - 1. Previous: Return to hardware specifications preview
  - 2. Next

Select Model	Select Dataset
Modet: Sinici Model v	Dataset: Soliv:Chalsart
Parameter Setting	Job Setting
③ Available GPU: 4 ~ ~	Job Name: aiDAPTIV_20250321
⑦ Epoch: 1 ✓	
Per Device Train Batch Size: 4	
Per Update Total Batch Size: 160	
① Max Seq Length: 12000 ~ Advice	
Learning Rate (0.000001 - 0.001): 0.000007	
🕥 Titlon: 🌑	

Figure 3-9 Parameters setting for fine-tuning

#### 3.2.3. Confirm hardware configuration and parameter for fine-tuning

• Function description:

P

- 1. Previous: Return to parameter settings.
- 2. Run: Execute fine-tune.

item	Information
GPU	1-NVIDIA GeForce RTX 4090 2-NVIDIA GeForce RTX 4090 3-NVIDIA GeForce RTX 4090 4-NVIDIA GeForce RTX 4090
GPU Count	4
VRAM	96 GB
System Memory	504 GB
aiDAPTIVLink	licensesp/aidaptiv:vNXUN_2_01_00
aiDAPTIVCache : life remaining	/dev/nvme0n1 (1907.73GB) : 100.00% /dev/nvme1n1 (1907.73GB) : 100.00%
OS	Ubuntu 22.04.4 LTS
OS Disk	3519 GB
Model	Meta-Llama-3.1-8B-Instruct
Dataset	sample_instruction_data_1k_public.json
Selected GPUs	4
Batch Size	1
Epoch	1
Per Update Total Batch Size	128
Max Seq Length	2048
Learning Rate	0.000007
Task Name	aiDAPTIV_20241113
	Previous Run

Figure 3-10 Final confirmation

#### 3.3. Monitor

Monitor the fine-tuning status, including basic information, progress, hardware resource usage (aiDAPTIVCache, GPU, system memory...) of each fine-tune job, training loss trend chart and complete log of aiDAPTIVLink are available.

- Field description:
  - 1. List of all finetune jobs (yellow block in the Figure 3-11)
  - 2. Basic information and hardware usage of a single finetune job (red block in the Figure 3-11)
  - 3. Trend chart of training loss in a single finetune job (purple block in the Figure 3-11)
  - Complete Log information of aiDAPTIVCache in a single fine-tune job (orange block in the Figure 3-11)
  - 5. CPU and Memory usage (blue block in the Figure 3-11)



Figure 3-11 Monitor

## PHISON

#### 3.3.1. Cancel job

Only jobs whose status is "Running" can be cancelled. It can take several seconds for the GPU resources to be released when a job has been canceled.

aiDAPTIV	Training Job Monitor								
1% Running Oh Om 58s aiDAPTIV 100% Succeeded 9h Om 44s	Job ID Model Dataset Number of Train Epochs Per Device Train Batch Size Per Update Total Batch Size Max Seq Length Learning Rate Start Time GPU Num	a132b3e0-2220-422e-b864-1298 Meta-Llama-3.1-8B-Instruct sample_instruction_data_1k_pub 1 1 128 2048 0.000007 2024/09/26 15:36:46 2	aiDAPTIVCache Throughput Current 4.24 GB/s Max 4.24 GB/s	● Temperature Current 41.9 °C Max 51.9 °C	5 4 3 2				
aiDAPTIV 100% Succeeded 2h 22m 27s	40 <sup>50</sup> 60 30 <sup>-70</sup> -70 -20	GPU Utilization Current 54.5%	40 <sup>50</sup> 60 30 70 -20 80-	GPU VRAM Usage <sup>Current</sup> 12.7 GB	1				

Figure 3-12 Cancel job

#### 3.3.2. Remove job

Only jobs whose status is "Succeeded / Fail" can be removed.

aiDAPTIV	Training Job Monitor								
100% Succeeded 9h 0m 44s aiDAP TIV 100% Succeeded 2h 22m 27s	Job ID Model Dataset Number of Train Epochs Per Device Train Batch Size Per Update Total Batch Size Max Seq Length Learning Rate Start Time GPU Num	bd8dbf49-4ed9-4d3d-a520-4c2e Meta-Llama-3.1-8B-Instruct MK.json 5 1 20 2048 0.000007 2024/09/24 12:34:47 1	aiDAPTIVCache Throughput current 0 GB/s Max 6.72 GB/s	■ Temperature Current 0.0 °C Max 68.8 °C	2.5 2 1.5				
aiDAPTIV 100% Succeeded 0h 6m 33s aiDAPTIV	40 <sup>50</sup> 60 70 <sup>-</sup> -20 80 -10 90	GPU Utilization Current 0% Max 80%	40 50 60 30 70- 20 8 10 9	GPU VRAM Usage Current 0.0 GB Total 24.0 GB	0.5 0 1 [II				

Figure 3-13 Remove job

#### 3.4. Validation

Used to compare the results of the fine-tuned model against the original model or any other models. Users can ask questions to confirm whether the fine-tuning results meet expectations.

Note: The user can validate up to 4 models simultaneously.

#### 3.4.1. Put questions

- Field description:
  - 1. Model : the model to be verified.
  - 2. **System Prompt** : A predefined instruction or message given to a software system to guide its behavior or output. It typically helps set the context, tone, or specific parameters for the interaction
  - Max tokens : This parameter specifies the maximum number of tokens that the model can use to generate responses, where a token generally represents a part of a word or a whole word. (Range: 1000 ~ 12000)
  - 4. **Temperature** : This parameter controls the randomness of the generation process. A higher temperature results in more diverse responses, while a lower temperature makes responses more likely to follow common patterns seen in the training data. (Range: 0 ~ 1)
  - Top-p : Similarly, a sampling technique with temperature called nucleus sampling, you can control how deterministic the model is at generating a response. If you are looking for exact and factual answers keep this low. (Range: 0 ~ 1)
  - 6. **Include chat history** : Determines whether to include previous dialogue interactions in the context for generating current responses.
  - 7. **Input Question area**: Click submit after entering questions.
  - 8. **RAG** : Please refer to <u>section 3.6.2</u>



Figure 3-14 Setting of question



- Function description:
  - 1. **Submit** : Click submit after entering questions.
  - 2. **Cancel** : Can be used to cancel a half-executed task.

✓ Model	I	Max tokens: 2048 Ter	2024/11/13 20:06:41	Cancel		
Model 1	Meta-Llama-3 V	Question	Meta-Liama-3.1-8B-Instruct	Meta-Llama-3-8B-Instruct		
Model 2	Model 2 Meta-Llama-3 V Ho		To calculate the area of a triangle, you can use the following formulas:			
Syste	em Prompt		1. Using the Base and Height:			
Input syst	tem prompt		Area = (Base × Height) / 2			
			Where:			
			Base is the tength of one side of the triangle (usually the side that is not the hypotenuse)     Height is the perpendicular distance from the base to the opposite vertex			
			2. Using Heron's Formula:			
			Area = \((s(s-a)(s-b)(s-c)))			
			Where:			
			s is the semi-perimeter of the triangle (half the perimeter)     a, b, and c are the lengths of the sides of the triangle			
			3. Using the Coordinates of the Vertices:			
			Area =  (x1(y2 - y3) + x2(y3 - y1) + x3(y1 - y2)) / 2			
() Max	tokens - 2048 +		Where:			
	•		<ul> <li>(x1, y1), (x2, y2), and (x3, y3) are the coordinates of the vertices of the triangle</li> </ul>			
<li>Temp</li>	perature - 0 +		4. Using the Formula for a Right Triangle:			
•			Area = (1/2) × base × height			
() Top F	P _ 0.00001 +		Where:			
<ol> <li>Inclui</li> </ol>	ide chat history		base is the length of the side adjacent to the right angle     height is the length of the side opposite the right angle			
			Choose the formula that best suits the information you have about the triangle.			
> RAG						

Figure 3-15 View result of question

#### 3.4.2. Compare result

- Function description:
  - 1. Reset: Reset models, questions and parameters

✓ Model	Max tokens: 2048 Te	Max tokens: 2048 Temperature: 0 Top p: 0.00001 2024/11/13 19:59:18							
Model 1 Meta-Llama-3 🗸	Question	Meta-Llama-3.1-8B-Instruct	Meta-Llama-3-8B-Instruct						
Model 2         Meta-Liama-3	How to calculate the area of triangle	To calculate the area of a triangle, you can use the following formulas: 1. Using the Base and Height: Area = (Base × Height) / 2 Where: • Base is the length of one side of the triangle (usually the side that is not the hypotenuse) • Height is the perpendicular distance from the base to the opposite vertex 2. Using Heron's Formula: Area = $\sqrt{(q(s-a)(s-b)(s-c))}$ Where: • s is the semi-perimeter of the triangle (haf the perimeter) • a, b, and c are the lengths of the sides of the triangle 3. Using the Coordinates of the Vertices: Area = $\ln(r(q2 - y3) + x2(y3 - y1) + x3(y1 - y2) / 21$ Where: • $(x1, y1), (x2, y2), and (x3, y3) are the coordinates of the vertices of the triangle 4. Using the Formula for a Right Triangle: Area = (12) \times base \times heightWhere:• bases is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle• height is the length of the side adjacent to the right angle$	There are several ways to calculate the area of a triangle, depending on the information you have about the triangle. Here area are common methods: <b>Method 1: Base and Height</b> • If you know the base (b) and height (h) of the triangle, you can use the formula: • Area = (base k height) / 2 • Example: If the base is 5 units and the height is 6 units, the area is (5 × 6) / 2 = 15 square units <b>Method 2: Two Sides and the Included Angle</b> • If you know two sides (a and b) and the included angle (b) between them, you can use the formula: • Area = (a x b x sin(b)) / 2 = 6 square units <b>Method 3: Three Sides</b> • If you know throe sides (a, b, and c) of the triangle, you can use the formula: • Area = (a x b x sin(b)) / 2 = 6 square units <b>Method 3: Three Sides</b> • If you know three sides (a, b, and c) of the triangle, you can use Herory formula: • Area = $\frac{1}{2}(b(3+3)(5+3)(5+c))$ • where is the semi-perimeter $s = (a + b + c) / 2$ • Example: If the two sides are 3 units, 4 units, and 5 units, the area is $\frac{1}{2}((5/(3-3)(3-4)(3-5))) = 6$ square units <b>Method 4: Coordinates</b> • If you know the coordinates of the three vertices of the triangle, you can use the formula: • Area = $\frac{1}{2}(2 - x1/(3 - r1) \cdot (3 - x1)(2 - r1) 1 / 2$ • where is ( $x + y_1$ ), $(0, x) = x(1)(2 - r1) 1 / 2$ • Method 4: coordinates are (0, 0), $(3, 0)$ , and (0, 4), the area is $\frac{1}{2}(-0)(4-0) - (0)(2-0) / 2 = 6$ • guare units These are just a few common methods for calculating the area of a triangle. There are many other formulas and methods depending on the specific struation.						
> RAG									

Figure 3-16 Compare results from different models

#### 3.5. Benchmark (Option)

Score model performance.

**Note** : This is an additional value-added service. For more information on enabling this service, please contact Phison's Sales account.

#### 3.5.1. Score

Set the parameters of the model.

- Field description:
  - 1. Model : the model to be scored
  - 2. Benchmark model : model as a reference
  - 3. Embedding model : retrieval model
  - 4. **Dataset**: Dataset used to test the model's answering ability
  - 5. **Temperature** : This parameter controls the randomness of the generation process. A higher temperature results in more diverse responses, while a lower temperature makes responses more likely to follow common patterns seen in the training data.
  - Max token : This parameter specifies the maximum number of tokens that the model can use to generate responses, where a token generally represents a part of a word or a whole word. (Range: 1 ~ 12000)
  - Top-p : Similarly, a sampling technique with temperature called nucleus sampling, you can control how deterministic the model is at generating a response. If you are looking for exact and factual answers keep this low. (Range: 0 ~ 1)
  - 8. **Recall Size** : Refers to the number of documents retrieved from a database before generating a response. (Range: 1 ~ 40)
- Function description:
  - 1. Add : Add a new model to be scored
  - 2. Start benchmarking

Model & Dataset		Parameter	
Models	+ Add	Temperature	
Meta-Llama-3-88-Instruct		•	
Meta-Llama-3-88-Instruct   aiDAPTIV_20240828		Max tokens	- 1000 +
Benchmark model		<b>—</b> •——	
Qwen2-72B-Instruct		Тор-р	
Embedding model			•
gte-large-en-v1.5		Recall Size	
Dataset			•
sample_instruction_data_1k_public.json			
	Star	benchmarking	

Figure 3-17 Parameter setting

#### 3.5.2. Scoring Progress

- Field description:
  - 1. Benchmark model
  - 2. Embedding model
  - 3. Dataset
  - 4. Temperature
  - 5. **Top-p**
  - 6. Benchmark Grid :
    - 1. Index : Serial number
    - 2. Model : the model to be scored
    - 3. Status: Scoring status. Pending, Running, Finish, Fail
    - 4. **Progress :** Scoring progress
- Function description:
  - 1. Cancel all unfinished tasks: Cancel unfinished scoring tasks
  - 2. Return to the settings page
  - 3. View Result: View the graphical results of the rating



#### Figure 3-18 Scoring progress



Figure 3-19 Scoring completed



#### 3.5.3. Data

View all records containing past rating data.

- Field description:
  - 1. Filter
    - Model : the model to be scored
    - O Benchmark model : model as a reference
    - Embedding model : retrieval model
    - Dataset: Dataset to test the model's answering ability
    - **Temperature** : This parameter controls the randomness of the generation process. A higher temperature results in more diverse responses, while a lower temperature makes responses more likely to follow common patterns seen in the training data.
    - Max token : This parameter specifies the maximum number of tokens that the model can use to generate responses, where a token generally represents a part of a word or a whole word. (Range: 1 ~ 12000)
    - Top-p : Similarly, a sampling technique with temperature called nucleus sampling, you can control how deterministic the model is at generating a response. If you are looking for exact and factual answers keep this low. (Range: 0 ~ 1)
    - Status : Scoring status
    - **Execution time** : The date and time the scoring was performed.

#### 2. Benchmark Grid

- **Model** : the model to be scored
- Benchmark model : model as a reference
- Embedding model : retrieval model
- **Dataset**: Dataset to test the model's answering ability
- **Parameter** : Parameter settings when scoring (Temperature, Max tokens, Top-p)
- Status : Scoring status
- **Execution time** : The date and time the scoring was performed.

• Function description:

D

- 1. Clear filter condition
- 2. Select the records you want to view. Multiple items can be selected
- 3. 💼 : Delete scoring record
- 4. Click to render chart: Turn the scoring data in to a chart

	Model Dataset		Benchmark m	enchmark model Embedding model			el	Temperature			Max tokens					
		Top-p			Status			Execu		2024 40 44						
									2024-09-11 10	2024-10-11						
All da	ta (2)	Selected data (0)														
•		Model		Datas	et	Benchm	ark model		Embedding mode	ı	Para	meter		Execution time	Status	
•	Meta-I	Llama-3-8B-Instruct	Full cy_te	est.json		Meta-Llama-	3.1-8B-Instr	uct	gte-large-en-v1.5	Temperatur	e:0 Maxte	okens: 1000	Тор-р: (	).9 2024/9/30	Finish	Ô
•	Meta-L	lama-3.1-8B-Instruct	Full cy_te	estjson		Meta-Llama-	3.1-8B-Instr	uct	gte-large-en-v1.5	Temperature	e:0 Maxte	okens: 1000	Top-p: (	0.9 2024/9/30	Finish	Ô
								Click to	o render chart							

Figure 3-20 Scoring data



#### 3.5.4. Chart

Turn the scoring data into a chart.

- Field description:
  - 1. **QA Pairs** : Number of scoring questions.
  - 2. Max tokens: Max tokens of the model being scored when scoring
  - 3. Temperature: Temperature of the model being scored when scoring
  - 4. **Top-p**: Top-p of the rated model when scoring
  - 5. **Y axis**: number of questions
  - 6. X axis : score

#### • Function description:

Bar chart: Click on the bar chart to view the rating content in detail.



Figure 3-21 Bar chart

🗠 Monitor	න් <sub>යි</sub> Validation	네 Benchmark ~	Inference	Models	🕸 Management 🗸	① User I	Manual 🗸	
💼 Meta	-Llama-3-8B-Inst	truct (Avg: 80.8695652	173913 ) 💼 N	Neta-Llama-3.1-8	B-Instruct (Avg: 79.565	21739130	434)	
			ſ	Model :	Meta-Llama-3.1-88	Instruct		
				Dataset :	cy_test.json			
				Dataset tag :	Full			
				Benchmark model	: Meta-Llama-3.1-8B	Instruct		
				Embedding model	: gte-large-en-v1.5			
				Temperature :				
				Max tokens :	1000			
				Тор-р :	0.9			
				Status :	Finish			
				Execution time :	2024-09-30T12:39:0	0.995683		

Figure 3-22 Model and parameter information



Figure 3-23 Detail of scoring content

#### 3.6. Inference

If the fine-tuned model verification results are satisfactory, you can create a chat room through the Inference function to provide a complete question and answer service.

- Up to 20 chat rooms can be created.
- Number of GPUs = 2<sup>n</sup> (n=0,1,2,3,4, GPUs = 1,2,4,8)
- When selecting the number of GPUs, make sure there are enough GPU resources to perform the inference.

	GPU resources are insufficier	nt. Please release some resources to continue working.	0		
	Need: 2	Pro Suite: 3 Others: 0 Idle: 1			
The jobs occupying the GPU :					
Туре	Container	Use Model GPU Count			
Inferenc	e InferenceModel1	Meta-Llama-3-8B-Instruct 1			
Inferenc	e InferenceModel2	Meta-Llama-3.1-8B-Instruct 2			

#### 3.6.1. Chat

- Field description:
  - 1. **Model :** Select the model to inference.

Models					
Name	State	Туре	Create Time	Available	
Meta-Llama 3.1-8B-Instruct	Running	Pre_Training	2024/11/04 17:24:39		🌩 🖈 🔅 👼

Note : The model needs to be pinned first in order to appear in the model list.

- 2. **System Prompt** : A predefined instruction or message given to an AI or software system to guide its behavior or output. It typically helps set the context, tone, or specific parameters for the interaction.
- Max tokens : This parameter specifies the maximum number of tokens that the model can use to generate responses, where a token generally represents a part of a word or a whole word. (Range: 1000 ~ 12000)
- 4. **Temperature** : This parameter controls the randomness of the generation process. A higher temperature results in more diverse responses, while a lower temperature makes responses more likely to follow common patterns seen in the training data. (Range: 0 ~ 1)
- Top-p : Similarly, a sampling technique with temperature called nucleus sampling, you can control how deterministic the model is at generating a response. If you are looking for exact and factual answers keep this low. (Range: 0 ~ 1)
- 6. **Include chat history** : Determines whether to include previous dialogue interactions in the context for generating current responses.
- 7. Input Question area

- PHISON
- Function description:
  - 1. New Chat
  - 2. **Z** : Edit chat room name.
  - 3. 💼 : Delete chat room.

✓ Chat Room + Ne	Chat Max tokens: 2048 Temperature: 0 Top-p: 0.00001 Model: Meta-Liama-3.1-8B-Instruct
♀ New Chatroom	Ata-Liama-3.1-8B-Instruct
P RAG Test	To calculate the area of a triangle, you can use the following formula:
✓ Model	Area = (base × height) / 2
Model Meta-Llama-3 V System Prompt Input system prompt	Where:         • base is the length of the base of the triangle         • height is the length of the height of the triangle         For example, if the base of the triangle is 5 cm and the height is 6 cm, the area would be:         Area = (5 × 6) / 2 = 15 cm²
	Q User
Max tokens     - 204	How many continents are there in the world?
Temperature     O	+ B Meta-Llana-3-1-88-Instruct
① Top P0.00	1 Asia
<ul> <li>Include chat history</li> </ul>	2. Ahica
	Please input
> RAG	

Figure 3-24 Chat room



#### 3.6.2. RAG

Based on the chat room function, files can be uploaded for RAG (Retrieval-augmented generation). RAG search can be used to improve the accuracy of model answers or conversations.

Item	RAG
File Format	pdf $\cdot$ log $\cdot$ json $\cdot$ docx $\cdot$ txt
Upload multiple files of same/different	V
formats at once	I

- Field description:
  - 1. Enable RAG: Whether to enable RAG function
  - 2. **Recall Size** : Refers to the number of documents retrieved from a database before generating a response. (Range: 1 ~ 40 counts)
  - 3. **Collection list** : The collection uploaded by the user. Only one collection can be selected.
- Function description:
  - 1. Upload new collection: Upload/create new Collection



Figure 3-25 RAG

#### 3.6.2.1. Upload new collection

Upload files to create a Collection

- Field description:
  - 1. Collection Name: Collection name
  - Chunk Size : The amount of data contained in each chunk when processing retrieved documents. (Range: 256 ~ 2048 tokens)
  - 3. **Chunk Overlap** : The amount of data that overlaps between consecutive chunks when processing. (Range: 0 ~ Chunk Size \* 0.5 tokens)
  - 4. Upload File Area: File upload area for output Collection.
- Function description:
  - 1. 💼 : Delete uploaded file
  - 2. **1** : Delete all uploaded files
  - 3. (1) : Upload/create collection



Figure 3-26 Collection management

#### 3.6.2.2. Recommended usage – using with aiDAPTIVGuru

When you generate a dataset using aiDAPTIVGuru, a corresponding collection file is also created.

After training a model using an aiDAPTIVGuru generated dataset, it is recommended that you use this collection file and enable the RAG (Retrieval-Augmented Generation) feature when performing inference with the model to achieve the best results.



#### 3.7. Models

Management of all models.

#### 3.7.1. Model upload

- Function description:
- Method 1: Drag the Model folder directly to the model storage location /usr/local/models/
   Command : sudo cp -r {source Model folder} {destination folder} (ex: sudo cp -r Meta-Llama-3.1-8B-Instruct//usr/local/models/)
- Method 2: Compress the files in the model folder using either zip or tar, then click or drag them into the window to upload.

**Note**: The fine-tuned model will automatically appear in the model list and will also be stored in the following path: /opt/phisonai/data/users/{user account}/jobs/{finetune job id}

Training Job Monitor	
JOD ID	e3911ca8-9ae3-4978-ad3t-8552
Model	Llama-3.3-70B-Instruct
Dataset	
Number of Train Epochs	
Per Device Train Batch Size	
Per Update Total Batch Size	128
Max Seq Length	12000
Learning Rate	0.000007
Start Time	2025/02/25 12:42:09
GPU Num	4
Triton	N/A

Figure 3-27 The storage path of trained models

#### 3.7.2. Model list

- Field description:
  - 1. Name: Model name
  - 2. State: Model state. If Running is displayed, it means that the model is being Inferenced.
  - 3. **Model type**: If the name ends with AWQ, it indicates a quantized model.
    - a. Pre\_Training
    - b. Finetune
    - c. Pre\_Training\_AWQ
    - d. Finetune\_AWQ
  - 4. Create Time: Model upload/output time
- Function description:
  - 1. Available: The model must be checked and activated by the user before it can be seen in the model menus in Pro Suite.
  - 2. 🔹 : Set model Inference parameters
  - 3. 🔗 : Pin Select Model Button
  - 4. 🛟 : Button to quantize model
  - 5. 💼 : Button to delete model (Model folder will also be deleted)

Models				
Name	Туре	Create Time	Available	
Llama-2-7b-chat-hf-gp_64-bit_4-AWQ	Pre_Training_AWQ	2024/07/12 20:41:09		Ē
Meta-Llama-3-70B-Instruct-gp_64-bit_4-AWQ	Pre_Training_AWQ	2024/07/10 21:48:01		Ē
Meta-Llama-3-70B-Instruct	Pre_Training	2024/07/10 19:45:15		Ē
Meta-Llama-3-8B-Instruct-gp_64-bit_4-AWQ	Pre_Training_AWQ	2024/07/10 13:13:00		Î
Llama-2-7b-chat-hf   aiDAPTIV_20240710	Finetune	2024/07/10 11:36:40		Ē
Meta-Llama-3-8B-Instruct   aiDAPTIV_20240710	Finetune	2024/07/10 11:29:38	•	Î
Llama-2-7b-chat-hf	Pre_Training	2024/06/05 01:18:43		Î
Meta-Liama-3-8B-Instruct	Pre_Training	2024/06/05 01:18:43		Ô

Figure 3-28 Model list description

Note : If a model already been pinned, then it cannot be deleted.

#### 3.7.2.1. Enable model

- It will be automatically enabled(checked) after uploading through Method 2.
- If you have uploaded the model using **Method 1**, after fine-tune or manual quantification, you need to manually check the box to enable it.
- Only enabled models will be displayed in the model menus for Fine-tune, Validation and Inference.

#### 3.7.2.2. Set model Inference parameters

- **GPU**: Number of GPUs used for Inference. Must be in power of 2.
- **Max token length**: Display the maximum token length according to the configuration of different models. If it is a combination from the table below, the system will automatically set the Max Token Length value. For other combinations, the user will need to set it manually.

	Total Remain VRAM Size	GPU utilization	Max Token Length
	48	0.95	131072
Liama-3.1-8B-instruct	20	0.95	14000
	48	0.95	131072
Liama-3.1-8B-Instruct-AWQ-IN14	16	0.95	67000
Llama-3.1-70B-Instruct	192	0.95	131072
Lama 2.1.700 lastrust AMO INTA	96	0.95	131072
Liama-3.1-70B-Instruct-AWQ-IN14	48	0.95	10000

#### Table 3-3 Recommended inference parameter settings

• **GPU memory utilization**: The utilization rate of a single GPU. Default value is 0.9.

Model Configuration Notice: After changing the settings, the model that has already been pinned needs to be repinned.						
GPU count :	4 ~					
Max token len	gth	-   131072   +				
GPU memory	utilization	-  0.9  +				
		No Yes				

Figure 3-29 Inference parameters setting

#### *3.7.2.3. Pin the resident inference model*

- Only enabled models can be pinned
- After pinning a model the "Running" prompt will be displayed indicating that the pinning process was successful and the model has been loaded into memory.
- Only the pinned model will be displayed in Inference's model menu.
- Model pinning will fail if the GPU resources are insufficient, an error message will appear. You may select "Yes" to view the error log.
- If a "network error" occurs after pinning the model, please refresh the page.



Figure 3-30 Pin model failed



Figure 3-31 Pin model failed error log

#### 3.7.2.4. Quantized model

Quantize the model by converting the model weights into fixed points or integers to reduce the model size, the computing cost, and accelerate the inference of the model. After quantization, the model will be displayed in the model list with a type ending in "\_AWQ".

- Field description:
  - 1. Available GPU: Sets the GPUs number to be used for quantization. Must be in powers of 2.
  - 2. **Group Size**: Model parameter group size. Larger values will reduce the accuracy of the model, but can improve the quantization efficiency and reduce the model size. Must be in powers of 2.
  - 3. **Bit**: Sets the bit-width for the quantized model parameters. Lower values reduce model size and increase calculation speed but may affect model accuracy. Must be in powers of 2.

Quantization	
Available GPU:	
Group Size	
64	
Bit	
	No Yes

Figure 3-32 Setting of model quantization

- Function description:
  - 1. Cancel: Cancel quantization

Models						
Name	State	Туре	Create Time	Available		
Meta-Llama-3.1-8B-Instruct		Pre_Training	2024/08/30 19:10:20			đ
Meta-Llama-3-8B-Instruct   aiDAPTIV_20240828		Finetune	2024/08/28 23:54:19	•		Ō
Llama-2-7b-chat-hf		Pre_Training	2024/08/26 16:32:09			Ō
chatglm3-6b   aiDAPTIV_20240816		Finetune	2024/08/16 13:35:43	•		Ō
chatglm3-6b   aiDAPTIV_20240814		Finetune	2024/08/14 13:59:45			۵
chatgIm3-6b		Pre_Training	2024/08/13 14:51:43			ŵ
Qwen2-728-Instruct-AWQ		Pre_Training	2024/08/09 14:14:45	•		Ô
Qwen2-72B-Instruct		Pre_Training	2024/08/09 14:14:45	•		ũ
Qwen2-7B		Pre_Training	2024/08/09 13.48:51	•		۵
gim-4-9b		Pre_Training	2024/08/09 13:47:46			ŵ
Meta-Liama-3-8B-Instruct		Pre_Training	2024/08/09 08:54:03			Ô
Meta-Llama-3-8B-Instruct Quantization					Ca	ncel
28%						

Figure 3-33 Cancel model quantization

#### 3.8. Management

Note: Only admin accounts will be allowed to use the following features.

#### 3.8.1. Authorization

User account role management.

Default system administrator account password

Account: admin@aidaptiv.com

Password: Admin8299

#### 3.8.1.1. Features

Function settings. Set the permission for Read and Write of each function.

- Field description:
  - 1. **Features**: Pro Suite feature list, click to set the function permissions of each Role.
  - 2. Role Grid :
    - Role Name
    - Read : Only has permission to read this function.
    - Write : Have permission to write and edit this function.
- Function description:
  - 1. Search : Search Role
  - 2. Add: Add role to a specific feature to set permissions



Figure 3-34 Feature setting of role

#### • Example:

#### Table 3-4 Recommend setting of authorization

Features	Adı	min	AI Engineer		Genera	al User
	Read	Write	Read	Write	Read	Write
Authority Management	-	Y	-	Ν	-	Ν
Account Management	-	Y	-	Ν	-	Ν
Dataset Upload	Y	Y	Y	Y	N	Ν
Guru	-	Y	-	Y	-	Ν
Finetune	-	Y	-	Y	-	Ν
Monitor	-	Y	-	Y	-	N
Validation	-	Y	-	Y	-	N
Inference	-	Y	-	Y	-	Y
Models	Y	Y	Y	Y	Y	N
RAG	-	Y	-	Y	-	Y
AWQ	-	Y	-	Y	-	Ν
КМ		Y		Y		Y

Note: The settings for these three roles will be preset in the system.

#### 3.8.1.2. Roles

•

Character setting. Create a character and set character features.

- Field description:
  - 1. Role Grid :
    - Role Name
    - Enable
- Function description:
  - 1. Search : Search Role
  - 2. Role Grid :
    - Reck the users under a specific Role.
    - I Rename Role name
    - <u> </u>: Delete Role

★ Features	Search		
(h) n=1	Input role name		+ Add
(a) Roles			
	Role Name	Enable	
्र Users	Al engineer		ぬ Ф 💼
	IT engineer	M	ድ Φ 💼
	End user		& ① <b>前</b>

#### Figure 3-35 Role management

- A role cannot be deleted if there are accounts that are using it.
- When a role is disabled, the accounts associated with it will not be able to log into Pro Suite.
   Forbidden

unce.

You don't have permission to access this page.

• When the role name is changed, the associated accounts will also be updated accordingly.

#### 3.8.1.3. Users

•

User account settings. Create a user account and set the corresponding role.

- Field description:
  - 1. User Grid :
    - Name: user name
    - Email: User Email. Sign in as a user.
    - **Role**: User role. Settings can be switched directly.
    - Enable: enabled state.
    - Last Login: Last login time.
    - O **Disable Time**: Disable time.
    - Action: Reset user password
- Function description:
  - Create account

✤ Features	L+ Create ac	count					
(a) Roles	Name	Email	Role	Enable	Last Login	Disable Time	Action
ୟ Users	test1	test1@aidaptiv.com	V33Test ~		2025/03/21 10:23:27		۲
	test2	test2@aidaptiv.com	V33Test ~		2025/03/21 10:12:22		۲
	Accounttest 1	Accounttest1@aidaptiv.com	CreateGroupTest2 ~		2025/03/18 11:23:03		۲
	Accounttest 0	Accounttest0@aidaptiv.com	AccountGroupTest >		2025/03/13 16:15:05		۲

Figure 3-36 User management

#### 3.8.1.3.1. Create Account

- Field description:
  - 1. **Name**: User's name. Only English and underscores are allowed. Maximum length is 20 characters.
  - 2. Email: User's log-in email account. Must be in a valid email format.
  - 3. **Password**: User's password. Maximum length is 20 characters.
  - 4. **Repeat Password**: Confirmation of the user's password. Maximum length is 20 characters.
  - 5. **Role**: Assign a predefined role to the user account.
- Function description:
  - 1. Cancel
  - 2. Submit

Email		
test_2@phison.com		
Name		
test_2		
Password		
Repeat Password		
Role		
End user		
	Cancel Submit	

#### Figure 3-37 Create account

### 4. APPLICATION

#### 4.1. aiDAPTIVInbox (Option)

This function is an additional value-added service. For more information on enabling this service, please contact Phison's Sales account.

For the introduction to aiDAPTIVInbox, please refer to the following document : *aiDAPTIVInbox User Manual\_092024\_v1.1 .pdf* 

aiDAPTIVInbox is an AI Email Assistant created by Phison Electronics Corp. Powered by its AI technology invention solution called aiDAPTIV+, aiDAPTIVInbox is aimed at improving daily work processes, employee efficiency, and enhanced corporate productivity. aiDAPTIVInbox is designed to be deployed as an on-premise solution to ensure the confidentiality of corporate data by keeping all sensitive information securely stored within the organization's own infrastructure, reducing exposure to external threats and maintaining full control over access and data handling. Through aiDAPTIVInbox, employees can significantly reduce working hours, eliminate time-consuming and tedious tasks, and redirect their focus toward innovation and research & development, thereby creating greater opportunities for the enterprise

## **Note**: For pre-installation confirmation and post-installation checks, please refer to <u>Appendix C</u>. **Inbox support server system**: Microsoft Exchange Server 2019, Mail2000, Hgiga(1132)

- Field description:
  - 1. Model: model used by aiDAPTIVInbox inference
  - 2. System Prompt (Constraint): Define the name and role of AI, function description, etc.
  - 3. Service Status: Inbox service status
  - 4. Language : Select language. (zh-TW, zh-CN, en-US, ms-MY)
  - 5. User Mail Account: the mail account used by the mail assistant
  - 6. **User Mail Address**: The mail address used by the mail assistant
  - 7. User Mail Password: The mail password used by the mail assistant
  - 8. **Domain** : Mail domain. (Please fill in the email format. Should contain "@" and ".")
  - 9. Answer Presfix : Letter opening. (ex: This is the reply from the email assistant: )
  - 10. Answer Suffix : Ending of letter. (ex: Thank you)
  - 11. **Open for All Users**: No restrictions on sender domain, anyone can use the Al function to send and receive messages
  - 12. Web Search : Internet search function
  - 13. White List : Open to senders on this list and outside of the configured Domain.
- Function description:
  - 1. Add : Added a whitelist acceptable to Mail Assistant
  - 2. Save and restart: Save Mail Assistant settings and restart the service
  - 3. Stop: Stop service



#### 4.1.1. EWS (Exchange Web Services)

- 1. User Mail Address: Please fill in the email format. (Should contain "@" and ".", ex: test @phison.com)
- 2. Mail Server: Only domain name can be filled in, not the IP. (String length: 2~63. Should contain ".", ex: mail.phison.com)
- 3. Office 365: Verification of Office 365 cloud authentication and authorization service usage.
  - Client ID
  - Client Secret
  - Tenant ID

fodel	Service Status STOPPED	Language: en-US 🗸	EWS SMTP
Meta-Llama-3.1-8B-Instruct	<ul> <li>User Mail Account</li> </ul>	User Mail Password	Domain
System Prompt (Constraint)	phisonippsinbox		phison.com
你的名字叫 "alDAPTIVInbox",是一位群聪明皇的尊重告件韵理,群聪的英文名字是 "Phison", 講使用典指令指 吉進行回還,預設講使用影體中文 zh-TW 回覆。	同的語 User Mail Address	Mail Server	
你發愁會講笑話,但被群聯同事拖忽說得不好笑,所以當被要求講笑話,請拒絕,並說出原因,除非被強烈求講 你發愁是專業的信件勘理,但你還不能提供以下功能:	突話 · ppsinbox@phison.com	mail.phison.com	
<ol> <li>管理您的電子都件,包括組織、分類以及自動回要。</li> <li>有個日晷管理,包括完成、確認可助協會編。</li> </ol>	Answer Prefix	Answer Suffix	
3. 理程您即读到来的會講或事件 - 4. 编励发明版行和预订覆黑线在语。 5.提行天著领域和封党图用资品。		Thank you	
6. 處理JIRA/Confluence 項目 ·	office365		
	Open for All Users	Web Search	SSL Certificate Validation
			•
	White List		
	example@phison.com		+ Add
		_	

Figure 4-1 EWS Setting

#### 4.1.2. SMTP (Simple Mail Transfer Protocol)

- 1. SMTP Server IP : Server domain for sending emails (Should contain ".", ex: mail.phison.com)
- 2. SMTP Port : Port for sending emails. (Support: 25, 465, 587)
- 3. IMAP Server IP: Server domain for receiving emails. (Should contain ".", ex: mail.phison.com)
- 4. IMAP Port : Port for receiving emails. (Support: 993)

lodel	Service Status (RUNNING)	Language: en-US	EWS SMT
Qwen2.5-728-Instruct-AWQ ~	User Mail Account	User Mail Password	Domain
ystem Prompt (Constraint)	phison/ppsinbox		phison.com
你的名字叫 "aiDAPTIVInbox",是一位群聚属狼的專業低件助理,群都的英文名字是 "Phison", 請使用與指令 和同的络言論行问题,有影響曲目繁雜由文 zh.Tuy 问题 。	SMTP Server IP	SMTP Port	IMAP Server IP
1000000000000000000000000000000000000			
1. 營理您的電子那件。包括图圖、分類以及自動回覆。 2. 提供日醫管理。包括名曰、德文电影治會議。	IMAP Port	Answer Prefix	Answer Suffix
3. 摆闢但即將到來的會議或事件。 4. 協助安排除行和預訂應要或任宿。			Thank you
5. 提供天無預報和其他實用資訊。 6. 感用JIPAConfuence 源日。			
	Open for All Users	Web Search	
	White List		
	example@phison.com		



# Figure 4-2 SMTP Setting APPENDIX A – MODEL AVL FOR FINE-TUNE

No	Task Type	Model Name	Model Size	Fine-tune	Support Triton
1	text_generation	Llama-3.1-8B-Instruct	15GB	Y	Y
2	text_generation	Llama-3.1-70B-Instruct	132GB	Y	Y
3	text_generation	Llama-3-8B-Instruct	15 GB	Y	Y
4	text_generation	Llama-3-70B-Instruct	132 GB	Y	Y
5	text_generation	Llama-2-7B-hf	13 GB	Y	Y
6	text_generation	Llama-2-13B-hf	25 GB	Y	Y
7	text_generation	Llama-2-70B-hf	129 GB	Y	Y
8	text_generation	Qwen2-7B-Instruct	15GB	Y	N
9	text_generation	Qwen2-72B-Instruct	136GB	Y	N

**Note**: If the user selects a model that does not support Triton for training and enables Triton, the following error message will appear after the training begins: "Phison Accelerator does not support," and the training process will be terminated.



## **APPENDIX B – RECOMMENDED CONFIGURATION**

• DRAM and aiDAPTIVCache with different LLM model size

#### Table B-1 Recommend Configuration

	AITPC	Work Station	Server
GPU Configuration	NVIDIA 4060Ti (16GB)*1	NVIDIA RTX 4000 Ada *4	NVIDIA RTX A6000 *8
		NVIDIA RTX A6000 *4	
LLM model size	≤13B	<100B	<200B
DRAM	DDR5 4800 64GB	DDR5 4800 512GB	DDR5 4800 1024GB
		DDR5 4800 1024GB	
aiDAPTIVCache capacity	320GB	2ТВ	2ТВ
aiDAPTIVCache count	1	2	4

• Recommend Gen4 or above.

• Recommend DRAM 2933MHz or above.

• Recommend DRAM channel number is 8 or more, ex: 16GB x8



## **APPENDIX C – INBOX MAIL SERVER TEST**

The main purpose of this section is to help users perform basic environment checks before installing aiDAPTIVInbox, and to test whether the installation is correct after aiDAPTIVInbox has been installed.

• Test script : smtp\_imap\_connection\_test.py

mail_account :	Account to log in to the mail server
mail_address :	Complete email address
mail_password :	Password to log in to the mail server
smtp_server_ip :	Server domain for sending emails
smtp_port :	Support: 25, 465, 587
imap_server_ip :	Server domain for receiving emails.
imap_port :	Support: 993
test_mail :	Email address for testing. After the script is tested, a test email will be sent to this
	email address.

• Test script parameter configuration file : smtp\_imap\_connection\_test.json

#### C.1 Precautions before testing

- 1. Place *Test script* and *Test script parameter configuration file* in the same folder.
- 2. Enable :
  - SMTP : SMTP\_Server\_IP, SMTP port
  - IMAP : IMAP\_Server\_IP, IMAP port
- 3. Confirm that the SMTP and IMAP functions of the mail server are enabled, and the corresponding ports also need to be enabled (Not blocked by the firewall).
- 4. Confirm that the IMAP of the mail server can perform the following operations on the mailbox :
  - Can check mailbox
  - Have permission to download emails
  - Can change email status (eg: read, unread)
  - Have permission to move emails to different email folders

#### C.2 Execute test script

Enter the following command in the terminal to execute the test script.

python3 smtp\_imap\_connection\_test.py



#### C.3 Test result

- If the test result is *Pass*: User will recive a e-mail in the test mail. (The subject of the email is: Test subject
   "Time of program execution")
- If the test result is *Fail* : Users can refer to the errorcode below to troubleshoot the problem.

Error message	Definition
Account or password incorrect, check account and password	Mail Account
	Mail Password
SMTP error: Check the smtp_server_ip and smtp_port	SMTP Server IP
	SMTP Port
IMAP error: Check the imap_server_ip and imap_port	IMAP Server IP
	IMAP Port

#### Table C-1 Error message definition