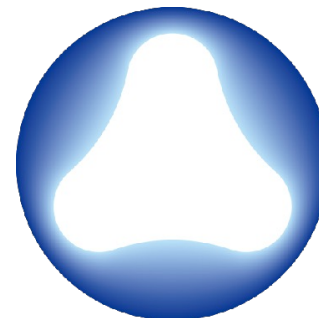




# Knowledge Fusion of Large Language Models

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, Shuming Shi  
Sun Yat-sen University, Tencent AI Lab



# Backgrounds: Large Language Models

## LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron\*, Thibaut Lavril\*, Gautier Izacard\*, Xavier Martinet  
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal  
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin  
Edouard Grave\*, Guillaume Lample\*

Meta AI

## LLAMA 2: Open Foundation and Fine-Tuned Chat Models

Hugo Touvron\* Louis Martin† Kevin Stone†  
Peter Albert Amjad Almahairi Yasmine Babaei Nikolay Bashlykov Soumya Batra  
Prajwal Bhargava Shruti Bhosale Dan Bikel Lukas Blecher Cristian Canton Ferrer Moya Chen  
Guillem Cucurull David Esiobu Jude Fernandes Jeremy Fu Wenyin Fu Brian Fuller  
Cynthia Gao Vedanuj Goswami Naman Goyal Anthony Hartshorn Saghar Hosseini Rui Hou  
Hakan Inan Marcin Kardas Viktor Kerkez Madian Khabsa Isabel Kloumann Artem Korenev  
Punit Singh Koura Marie-Anne Lachaux Thibaut Lavril Jenya Lee Diana Liskovich  
Yinghai Lu Yuning Mao Xavier Martinet Todor Mihaylov Pushkar Mishra  
Igor Molybog Yixin Nie Andrew Poulton Jeremy Reizenstein Rashi Rungta Kalyan Saladi  
Alan Schelten Ruan Silva Eric Michael Smith Ranjan Subramanian Xiaoqing Ellen Tan Binh Tang  
Ross Taylor Adina Williams Jian Xiang Kuan Puxin Xu Zheng Yan Iliyan Zarov Yuchen Zhang  
Angela Fan Melanie Kambadur Sharan Narang Aurelien Rodriguez Robert Stojnic  
Sergey Edunov Thomas Scialom\*

GenAI, Meta

## Mistral 7B

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford,  
Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,  
Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux,  
Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix,  
William El Sayed



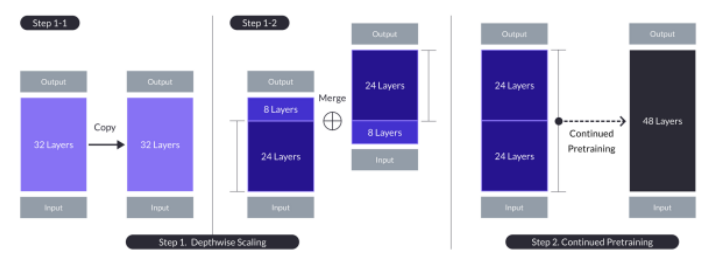
## Single LLM from Scratch

### SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling

Dahyun Kim\*, Chanjun Park\*†, Sanghoon Kim\*†, Wonsung Lee\*†, Wonho Song  
Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyunju Lee, Jihoo Kim  
Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunhyung Park, Gyoungjin Gim  
Mikyung Cha, Hwalsuk Lee†, Sunghun Kim†

Upstage AI, South Korea

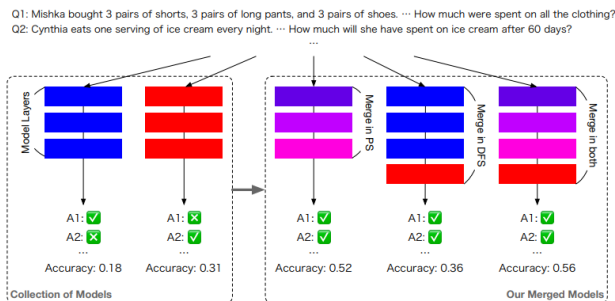
{kdahyun, chanjun.park, limerobot, wonsung.lee, hwalsuk.lee, hunkim}@upstage.ai



### Depth Up-Scaling Merging

### Evolutionary Optimization of Model Merging Recipes

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, David Ha  
Sakana AI  
Tokyo, Japan  
{takiba, mkshing, yujintang, qisun, hadavid}@sakana.ai

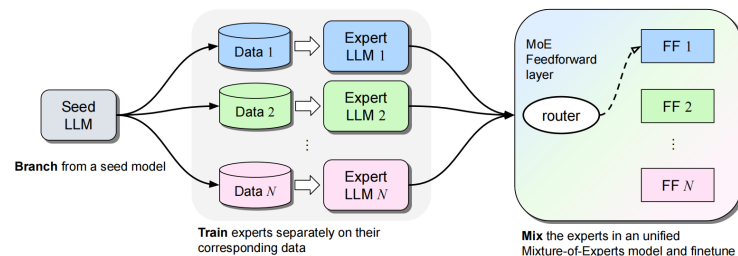


### Evolutionary Model Merging

### Branch-Train-MiX: Mixing Expert LLMs into a Mixture-of-Experts LLM

Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, Xian Li

FAIR at Meta



### Branch-Train-Mix

## Multiple LLMs Fusion

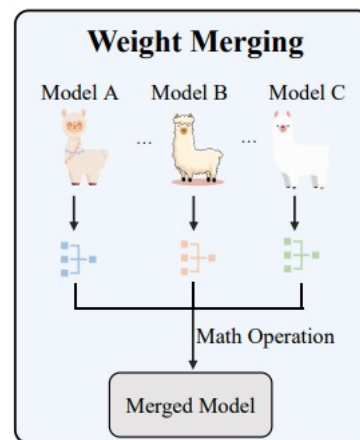
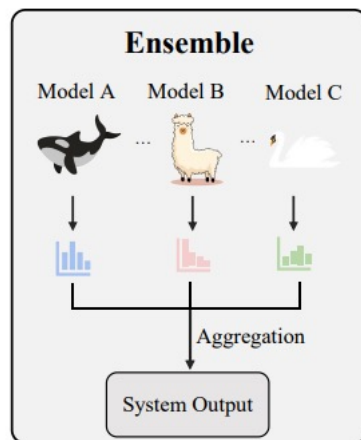
# Backgrounds: Knowledge Fusion

## ❑ Knowledge Fusion of LLMs

- ❑ Combining the capabilities of existing LLMs and transferring them into a potent LLM
- ❑ Model Ensemble: aggregate the **outputs of multiple models**
- ❑ Model Merging: arithmetic operation on the **parameter space** of multiple models

## ❑ Limitations of Existing Methods

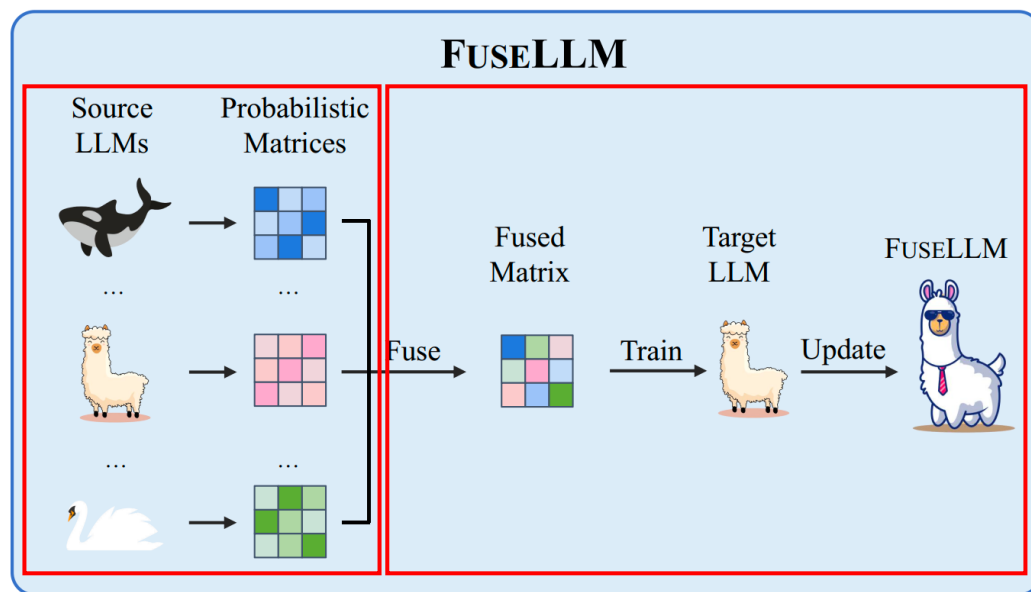
- ❑ Model Ensemble: **parallel deployment** of multiple LLMs
- ❑ Model Merging: **identical architecture** and minor parameter discrepancy



# FuseLLM: Knowledge Fusion of Large Language Models

## □ FuseLLM Design

- Motivation: **different probabilistic distributions** for the same text, originating from various LLMs, can be used to represent the **diverse knowledge** embedded within these models
- Step1: leverage **generative distributions** of source LLMs to externalize individual knowledge
- Step2: fuse the distributions and transfer to the target LLM through **lightweight training**



# FuseLLM: Knowledge Fusion of Large Language Models

## Key Factors

- Fusion Function: opt for distributions with **minimal ppl** or **weighted average based on ppl**
- Token Alignment: use **dynamic programming** to recursively minimize the total cost of editing one sequence of tokens to match the other

LLM 1 Tokenization		LLM 2 Tokenization		
<b>now</b>	token aligned			<b>now</b>
now	0.90	distribution aligned	0.91	now
current	0.05	distribution aligned	0.04	current
immediate	0.04	distribution not aligned	0.03	immediately
<b>get</b>	token not aligned			<b>gets</b>
get	0.50			
gains	0.38	degeneration to one-hot	1.00	get
obtains	0.10			

EM

LLM 1 Tokenization		LLM 2 Tokenization		
<b>now</b>	token aligned			<b>now</b>
now	0.90	distribution aligned	0.91	now
current	0.05	distribution aligned	0.04	current
immediate	0.04	mapping to 'immediate'	0.03	immediately
<b>get</b>	mapping to 'get'			<b>gets</b>
get	0.50	mapping to 'get'	0.60	gets
gains	0.38	distribution aligned	0.22	gains
obtains	0.10	distribution aligned	0.13	obtains

MinED

# FuseLLM: Knowledge Fusion of Large Language Models

## ❑ Experimental Setup

- ❑ Source LLMs: Llama-2 7B, OpenLLaMA 7B, MPT 7B
- ❑ Training Corpus: MiniPile (1M Documents, <2B Tokens, 22 Domains)
- ❑ Evaluation Details
  - ❑ Multiple-Choice Tasks: Big-Bench Hard, ARC-easy, ARC-challenge, BoolQ, HellaSwag, OBQA
  - ❑ Generation Tasks: MultiPL-E, TrivialQA, DROP, LAMBADA, IWSLT2017, SciBench
- ❑ Baselines
  - ❑ Original LLMs: Llama-2 7B, OpenLLaMA 7B, MPT 7B
  - ❑ Continual Trained LLMs: Llama-2 CLM (Casual Language Modeling)

# FuseLLM: Knowledge Fusion of Large Language Models

## ❑ Main Results

### ❑ Multiple Choice Tasks

Model	BBH	ARC-easy	ARC-challenge	BoolQ	HellaSwag	OpenBookQA
OpenLLaMA-7B	33.87	69.70	41.38	72.29	74.53	41.00
MPT-7B	33.38	70.12	42.15	74.74	76.25	42.40
Llama-2-7B	39.70	74.58	46.33	77.71	76.00	44.20
Llama-2-CLM-7B	40.44	74.54	46.50	76.88	76.57	44.80
🧡 FuseLLM-7B	41.75	75.04	47.44	78.13	76.78	45.40

### ❑ Generation Tasks

Model	MultiPL-E	TrivialQA	DROP	LAMBADA	IWSLT2017	SciBench
OpenLLaMA-7B	18.11	39.96	22.31	70.31	5.51	0.68
MPT-7B	17.26	28.89	23.54	70.08	5.49	0.88
Llama-2-7B	14.63	52.46	27.25	73.28	6.48	0.14
Llama-2-CLM-7B	14.83	53.14	28.51	73.45	6.91	0.94
🧡 FuseLLM-7B	15.56	54.49	28.97	73.72	6.75	1.65

FuseLLM-7B outperforms each source LLM and the continual training baseline in almost all tasks

The target LLM's performance on the test domain and the construction of training data are crucial



# FuseLLM: Knowledge Fusion of Large Language Models

## □ Detailed Results

### General Reasoning & Commonsense Reasoning

We first show the performance of FuseLLM on Big-Bench Hard and CommonSense benchmarks, which evaluate the general reasoning and commonsense reasoning abilities respectively.

Task	OpenLLaMA	MPT	Llama-2	Llama-2 CLM	FUSELLM
Boolean Expressions	74.40	66.00	68.80	<b>76.00</b> (+10.47%)	71.60 (+4.07%)
Causal Judgement	45.45	<b>50.80</b>	<b>50.80</b>	46.52 (-8.43%)	46.52 (-8.43%)
Date Understanding	43.60	43.60	59.60	59.20 (-0.67%)	<b>62.40</b> (+4.70%)
Disambiguation QA	36.00	47.60	46.80	48.00 (+2.56%)	<b>50.00</b> (+6.84%)
Dyck Languages	5.20	5.20	7.20	6.40 (-11.11%)	<b>8.80</b> (+22.22%)
Formal Fallacies	50.80	<b>52.80</b>	49.20	48.80 (-0.81%)	49.20 (+0.00%)
Geometric Shapes	0.00	0.00	<b>34.40</b>	19.20 (-44.17%)	22.80 (-33.72%)
Hyperbaton	62.80	53.60	54.40	56.40 (+3.68%)	<b>65.20</b> (+19.85%)
Logical Deduction (3 objects)	43.60	40.80	54.00	57.20 (+5.93%)	<b>60.40</b> (+11.85%)
Logical Deduction (5 objects)	24.80	31.60	31.20	<b>35.60</b> (+14.10%)	33.20 (+6.41%)
Logical Deduction (7 objects)	16.80	18.40	24.80	<b>29.60</b> (+19.35%)	25.60 (+3.23%)
Movie Recommendation	39.60	52.00	72.80	71.60 (-1.65%)	<b>73.60</b> (+1.10%)
Multistep Arithmetic Two	0.80	0.40	0.80	4.40 (+450.00%)	<b>4.80</b> (+500.00%)
Navigate	54.00	48.80	56.00	61.20 (+9.29%)	<b>64.40</b> (+15.00%)
Object Counting	49.60	40.40	49.60	51.60 (+4.03%)	<b>55.20</b> (+11.29%)
Penguins in a Table	28.08	28.08	32.19	31.51 (-2.11%)	<b>32.88</b> (+2.14%)
Reasoning about Colored Objects	28.00	31.60	46.40	47.20 (+1.72%)	<b>48.40</b> (+4.31%)
Ruin Names	31.20	23.20	<b>34.00</b>	30.80 (-9.41%)	32.40 (-4.71%)
Salient Translation Error Detection	14.80	0.00	24.80	27.60 (+11.29%)	<b>29.20</b> (+17.74%)
Snarks	44.94	45.51	47.75	<b>49.44</b> (+3.54%)	49.44 (+3.54%)
Sports Understanding	64.40	82.40	90.00	90.00 (+0.00%)	<b>91.20</b> (+1.33%)
Temporal Sequences	<b>32.00</b>	21.20	12.80	16.40 (+28.13%)	16.40 (+28.13%)
Tracking Shuffled Objects (3 objects)	<b>36.40</b>	30.40	33.20	33.20 (+3.61%)	34.40 (+3.61%)
Tracking Shuffled Objects (5 objects)	<b>19.20</b>	14.40	15.60	15.20 (-2.56%)	15.60 (+0.00%)
Tracking Shuffled Objects (7 objects)	10.80	2.00	<b>11.20</b>	9.60 (-14.29%)	10.40 (-7.14%)
Web of Lies	51.60	63.60	50.80	61.60 (+21.26%)	<b>65.60</b> (+29.13%)
Word Sorting	5.60	6.80	<b>12.80</b>	7.60 (-40.63%)	7.60 (-40.63%)
Avg. 27 Tasks	33.87	33.38	39.70	40.44 (+1.86%)	<b>41.75</b> (+5.16%)

Table 1: Overall results of FUSELLM and baselines in reasoning evaluations on Big-Bench Hard (BBH), where percentages indicate the rate of improvement/decrease compared to Llama-2.

Task	OpenLLaMA	MPT	Llama-2	Llama-2 CLM	FUSELLM
ARC-easy	69.70	70.12	74.58	74.54 (-0.05%)	<b>75.04</b> (+0.62%)
ARC-challenge	41.38	42.15	46.33	46.50 (+0.37%)	<b>47.44</b> (+2.40%)
BoolQ	72.29	74.74	77.71	76.88 (-1.07%)	<b>78.13</b> (+0.54%)
HellaSwag	74.53	76.25	76.00	76.57 (+0.75%)	<b>76.78</b> (+1.03%)
OpenBookQA	41.00	42.40	44.20	44.80 (+1.36%)	<b>45.40</b> (+2.71%)
Avg. 5 Tasks	59.78	61.13	63.76	63.86 (+0.16%)	<b>64.56</b> (+1.25%)

Table 2: Overall results of FUSELLM and baselines in commonsense evaluations on CommonSense (CS), where percentages indicate the rate of improvement/decrease compared to Llama-2.

### Code Generation & Text Generation

We then evaluate FuseLLM on MultiPL-E, which is a multilingual programming benchmark to assess the code generation performance. We also conduct experiments on several text generation benchmarks, including TrivialQA (question-answering), DROP (reading comprehension), LAMBADA (content analysis), IWSLT2017 (machine translation), and SciBench (theorem application).

Task	OpenLLaMA	MPT	Llama-2	Llama-2 CLM	FUSELLM
C++	<b>14.47</b>	13.11	7.45	9.88 (+32.62%)	9.25 (+24.16%)
Go	<b>68.20</b>	66.96	57.02	54.44 (-4.52%)	59.78 (+4.84%)
Java	<b>14.28</b>	13.42	10.31	10.50 (+1.84%)	10.34 (+0.29%)
JavaScript	<b>17.61</b>	13.01	13.17	14.25 (+8.20%)	14.32 (+8.73%)
PHP	<b>11.24</b>	9.53	9.75	9.04 (-7.28%)	9.41 (-3.49%)
Python	15.96	<b>17.24</b>	13.85	13.07 (-5.63%)	13.91 (+0.43%)
R	<b>7.52</b>	4.53	4.97	5.25 (+5.63%)	5.84 (+17.51%)
Ruby	10.34	<b>12.33</b>	10.37	10.68 (+2.99%)	11.24 (+8.39%)
Rust	6.18	<b>8.29</b>	6.77	6.96 (+2.81%)	7.05 (+4.14%)
TypeScript	<b>15.31</b>	14.13	12.61	14.19 (+12.53%)	14.50 (+14.99%)
Avg. 10 Tasks	<b>18.11</b>	17.26	14.63	14.83 (+1.37%)	15.56 (+6.36%)

Table 3: Overall results of FUSELLM and baselines in code generation evaluations on MultiPL-E (ME), where percentages indicate the rate of improvement/decrease compared to Llama-2.

Task	OpenLLaMA	MPT	Llama-2	Llama-2 CLM	FUSELLM
TrivialQA	39.96	28.89	52.46	53.14 (+1.30%)	<b>54.49</b> (+3.87%)
DROP	22.31	23.54	27.25	28.51 (+4.62%)	<b>28.97</b> (+6.31%)
LAMBADA	70.31	70.08	73.28	73.45 (+0.23%)	<b>73.72</b> (+0.60%)
IWSLT2017	5.51	5.49	6.48	<b>6.91</b> (+6.64%)	6.75 (+4.17%)
SciBench	0.68	0.88	0.14	0.94 (+571.43%)	<b>1.65</b> (+1078.57%)

Table 8: Overall results of FUSELLM and baselines in additional generative benchmarks, where percentages indicate the rate of improvement/decrease compared to Llama-2.

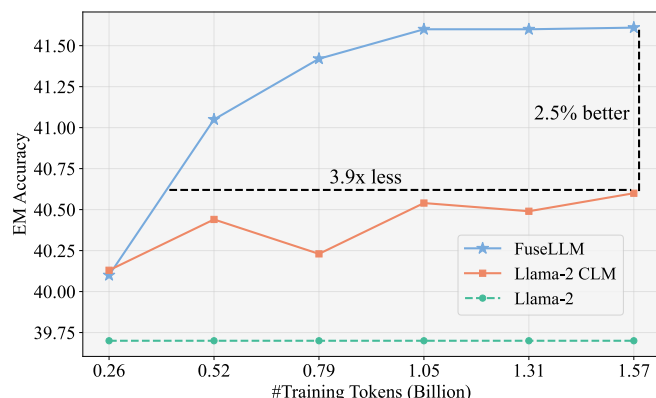


# FuseLLM: Knowledge Fusion of Large Language Models

## ❑ FuseLLM vs. Knowledge Distillation

❑ Similarities: speed up training efficiency (less token, higher performance)

❑ Differences: fusion of **3x7B** LLMs outperforms distillation from a **13B** LLM



Model	BBH	CS	ME
Llama-2 13B	47.92	66.33	18.76
OpenLLaMA	33.87	59.78	18.11
MPT	33.38	61.13	17.26
Llama-2	39.70	63.76	14.63
Llama-2 CLM	40.44 (+1.86%)	63.86 (+0.16%)	14.83 (+1.37%)
Llama-2 KD	40.88 (+2.97%)	64.41 (+1.02%)	15.45 (+5.60%)
FUSELLM	41.75 (+5.16%)	64.56 (+1.25%)	15.56 (+6.36%)

# FuseLLM: Knowledge Fusion of Large Language Models

## ❑ FuseLLM vs. Model Ensemble & Weight Merging

- ❑ Source LLMs: continual training Pythia-1B on different corpus (Phil, NIH, USPTO)
- ❑ Ensemble and merging methods achieve lower average ppl than the source LLMs
- ❑ FuseLLM achieves lowest average ppl

Model	Phil	NIH	USPTO	Average
Pythia	0.9008	0.6740	0.6077	0.7275
Phil	<b>0.8397</b>	0.6861	0.6228	0.7162
NIH	0.9248	<b>0.6215</b>	0.6278	0.7247
USPTO	0.9296	0.6872	<b>0.6017</b>	0.7395
Ensemble	0.8960	0.6647	0.6180	0.7262
Weight Merging	0.8786	0.6496	0.6054	0.7112
FUSELLM	0.8463	0.6569	0.6068	<b>0.7034</b>

# FuseLLM: Knowledge Fusion of Large Language Models

## □ Number of Source LLMs

□ Increasing number of source LLMs → Increased performance.

Model	BBH	CS	ME
OpenLLaMA	33.87	59.78	18.11
MPT	33.38	61.13	17.26
Llama-2	39.70	63.76	14.63
Llama-2 CLM	40.44 (+1.86%)	63.86 (+0.16%)	14.83 (+1.37%)
Llama-2 + OpenLLaMA	41.00 (+3.27%)	64.50 (+1.16%)	15.51 (+6.02%)
Llama-2 + MPT	41.16 (+3.68%)	64.51 (+1.18%)	15.47 (+5.74%)
FUSELLM	41.75 (+5.16%)	64.56 (+1.25%)	15.56 (+6.36%)

## □ Alignment Criteria & Fusion Function

□ Strict criteria (EM) ×; More information (MinED) ✓

□ Fusion Func: different is more or accurate is more?

Choice	BBH	ME	CS
<i>Alignment Criteria</i>			
EM	41.57	15.49	64.24
MinED	41.75 (+0.43%)	15.56 (+0.45%)	64.56 (+0.50%)
<i>Fusion Function</i>			
AvgCE	41.04	15.39	63.98
MinCE	41.75 (+1.73%)	15.56 (+1.10%)	64.56 (+0.91%)

# Conclusion

---

- ❑ Training LLMs from scratch incurs substantial costs and may lead to potential redundancy in competencies of LLMs
- ❑ We introduce the notion of knowledge fusion for LLMs, aimed at combining the capabilities of existing LLMs and transferring them into a potent LLM
- ❑ Compared to model ensemble and merging techniques, the field of LLMs fusion appears to be a more promising avenue, especially in light of the diverse structures and scales of LLMs

# FuseLLM: Knowledge Fusion of Large Language Models

Published as a conference paper at ICLR 2024

## KNOWLEDGE FUSION OF LARGE LANGUAGE MODELS


Fanqi Wan<sup>1\*</sup>, Xinting Huang<sup>2‡</sup>, Deng Cai<sup>2</sup>, Xiaojun Quan<sup>1†</sup>, Wei Bi<sup>2</sup>, Shuming Shi<sup>2</sup>

<sup>1</sup>School of Computer Science and Engineering, Sun Yat-sen University, China

<sup>2</sup>Tencent AI Lab

wanfq@mail2.sysu.edu.cn, quanxj3@mail.sysu.edu.cn

{timxthuang, jcykcai, victoriabi, shumingshi}@tencent.com

 **elvis** @omarsar0 · Jan 22

**Knowledge Fusion of LLMs**


Is it possible to merge existing models into a more potent model?

We have already seen a few ways that show the potential to effectively do this using approaches like weight merging and ensembling of models.

This work proposes FuseLLM with the core...

[Show more](#)

18 244 1K 116K

 **AIDB** @ai\_database · Jan 23

既存のLLMを融合させて強力なモデルを作る手法「知識融合」が開発されました。

実験では融合モデルがさまざまな能力（論理や常識、コード生成など）で融合前のモデルを超える結果が得られたということです。

"Knowledge Fusion of Large Language Models", ICLR 2024より...

[Show more](#)

2 254 879 285K

Llama-2+Mistral+MPT=? 融合多个异构大模型显奇效

机器之心 2024-01-27 12:38 北京

机器之心专栏

机器之心编辑部

融合多个异构大模型显奇效！中山大学、腾讯AI Lab推出FuseLLM

PaperWeekly 2024-01-29 21:10 北京



### Knowledge Fusion of Large Language Models

Version 1.0.0 License Apache 2.0 Stars 322 issues 0 open

[FuseLLM Paper @ICLR2024](#) | [FuseChat Tech Report](#) | [HuggingFace Repo](#) | [GitHub Repo](#)



**FuseAI** Community

<https://github.com/fanqiwan/FuseLLM> fanqiwan

Wanfq/FuseLLM-7B

Total downloads

7,515 (all time, tracked internally since January 2021)

🤖 Model Checkpoints: <https://huggingface.co/Wanfq/FuseLLM-7B>



Thank you!

