جسر مساعف الهوائي للتدريب الطبقي والاستدلال الحاكمي

توظيف **AIRLLM** والنماذج المفتوحة لبناء وتدريب واستغلال معماريات

MOUSSAIF1B / MOUSSAIF9B / MOUSSAIF-xB للذكاء الفتّاح

ورقة بحثية تأسيسية - هندسة أنظمة، تدريب كفؤ، تقطير حاكمي، واستدلال منخفض الذاكرة

المؤلف: **Zouhair Moussaif** / زهير مساعف

الإصدار البحثي 1.0 - يوليو 2026

ملاحظة تسمية — يُستعمل MOUSSAIF9B في هذه الورقة بوصفه اسم مشروع/إصدار معماري ما لم يكن النموذج المدرَّب فعلياً يضم نحو تسعة مليارات معلمة. أما MOUSSAIF-xB فهو اسم عائلة أحجام قابلة للتدرج، لا ادعاءً مسبقاً بحجم محدد.

الملخص

تقترح هذه الورقة إطاراً هندسياً جديداً يسمى «جسر مساعف الهوائي للتدريب الطبقي والاستدلال الحاكمي»، هدفه تحويل AIRLLM من مجرد وسيلة لتشغيل نماذج ضخمة على ذاكرة GPU محدودة إلى مكوّن بحثي داخل دورة تطوير معماريات MOUSSAIF للذكاء الفتّاح. تنطلق الورقة من حقيقة منهجية حاسمة: AIRLLM، في صورته البرمجية المعروفة، يركز على الاستدلال عبر تفكيك النموذج إلى شظايا طبقية وتحميل طبقة واحدة أو عدد محدود من الطبقات في الذاكرة الرسومية كل مرة؛ ومن ثم فهو لا يلغي تكلفة التخزين، ولا يحول حاسوباً ضعيفاً إلى عنقود تدريب، ولا يقدم بحد ذاته آلية رجوع خلفي كاملة لتدريب نموذج ضخم من الصفر. إلا أن هذا القيد لا ينقص من قيمته؛ بل يحدد موضعه الصحيح في خط الإنتاج: تشغيل معلمين كبار، توليد بيانات تركيبية عالية الكثافة، بناء محاكم تحقق متعددة النماذج، تحليل الفشل، قياس أثر التعديلات المعمارية، وإجراء تقطير معرفي وحاكمي إلى نماذج أصغر يمكن تدريبها فعلياً بواسطة LoRA وQLoRA وPEFT وDeepSpeed ZeRO وFSDP.

تقدم الورقة بنية ثلاثية: مسار MOUSSAIF1B بوصفه مختبر البحث السريع والشفاف، ومسار MOUSSAIF9B بوصفه مستوى القدرة المتوسطة الذي يستوعب مهارات الاستدلال والأدوات والذاكرة، ومسار MOUSSAIF-xB بوصفه عائلة متدرجة أو منظومة خبراء وتوجيه، لا نموذجاً أحادياً بالضرورة. وتقترح خمسة إسهامات نظرية جديدة باسم مساعف: نظرية الجسر الهوائي الحاكمي، ونظرية الأستاذ متعدد الأحجام، ونظرية التقطير التفاضلي الحاكمي، ونظرية سلم الأحجام الوظيفي، ونظرية الذاكرة السيادية العابرة للطبقات. كما تقدم اثنتي عشرة أداة تنفيذية، منها AirTeacher Forge، وFattah Trace Synthesizer، وDifference Court Sampler، وAdapter Lattice، وSovereign Replay Buffer، وxB Curriculum Compiler.

تخلص الورقة إلى أن AIRLLM يمكن أن يخفض عتبة الوصول إلى «القدرة المعرفية الكبيرة» في مرحلة الاستدلال، لكنه لا يخفض بالدرجة نفسها عتبة «التعلم الكبير»؛ ولذلك فإن أفضل استراتيجية لمساعف ليست محاولة تدريب 70B أو 405B عبر AIRLLM، بل استخدام هذه النماذج كهيئة تعليم وتحكيم وتوليد، ثم ضغط نواتجها في نماذج MOUSSAIF أصغر ذات معمارية حاكمية، مع إبقاء التعلم النهائي قائماً على بيانات موثقة، اختبارات مضادة، محكمة هدم، وقياسات مستقلة تمنع تحويل التقطير إلى نسخ للانحياز أو الهلوسة.

Abstract

This paper introduces the Moussaif Air-Governed Layerwise Bridge, an engineering framework that positions AIRLLM as a low-memory teacher, evaluator, and synthetic-data engine inside the MOUSSAIF open-model development lifecycle. AIRLLM's core value is layer-wise inference: model weights are split and streamed so that GPU residency depends largely on the largest active layer rather than the full parameter set. This enables access to large open-weight teachers on constrained hardware, but it does not by itself provide full pretraining or backward propagation for those models. The proposed architecture therefore separates inference democratization from training democratization. Large teachers are executed through AIRLLM to generate diverse candidate traces, counterexamples, critiques, preference pairs, and benchmark judgments; smaller MOUSSAIF1B, MOUSSAIF9B, and MOUSSAIF-xB students are trained through PEFT/LoRA/QLoRA, FSDP, or DeepSpeed ZeRO. Five new Moussaif theories and twelve implementation tools are proposed, together with

model selection, licensing, evaluation, compute planning, and falsification protocols. The central conclusion is that AIRLLM is most valuable not as a substitute for training infrastructure, but as an inference-side cognitive laboratory that allows resource-constrained researchers to borrow the capabilities of large open models and convert them into governed, testable, and compact Moussaif architectures.

الكلمات المفتاحية

AIRLLM، Layer-wise Inference، MOUSSAIF1B، MOUSSAIF9B، MOUSSAIF-xB، الذكاء الفتّاح، QLoRA، PEFT، Knowledge Distillation، FSDP، DeepSpeed ZeRO، التدريب الحاكمي، محكمة الهدم، النماذج المفتوحة.

أهم نتائج الورقة في سبع جمل

- 1. AIRLLM محرك استدلال منخفض الذاكرة قبل أن يكون أداة تدريب، ومن الخطأ مساواة تشغيل نموذج ضخم بتدريبه.
- 2. القيمة الأهم لمساعف هي إنشاء مزرعة معلمين كبار محلياً أو شبه محلياً لتوليد البيانات والتحقق والتقطير.
- 3. النموذج 1B هو أفضل مختبر لتجريب الحاكمية والذاكرة والتوجيه لأن تكلفة التكرار والانهيار أقل.
- 4. النموذج 9B مناسب لدمج المهارات والذاكرة والأدوات، لكنه يحتاج تدريباً حقيقياً بوسائل PEFT أو ZeRO/FSDP لا بـ AIRLLM وحده.
- 5. MOUSSAIF-xB ينبغي أن يكون سلماً وظيفياً أو منظومة خبراء، بحيث يُختار الحجم وفق عمق المهمة لا وفق تسويق الحجم.
- 6. النماذج Apache-2.0 أو MIT هي الأساس الأنظف بحثياً، أما Llama وGemma فينبغي وصفهما بدقة كنماذج مفتوحة الأوزان بتراخيص مخصصة.
- 7. نجاح البرنامج يقاس بقدرة النموذج الأصغر على الاحتفاظ بالفروق الحاكمة، لا بقدرته على تقليد أسلوب المعلم لفظياً.

فهرس تحليلي مختصر

المحتوى	القسم
المشكلة البحثية وحدود الادعاء	1
التشريح التقني لـ AIRLLM	2
موضع AIRLLM داخل دورة التدريب	3
خريطة النماذج المفتوحة المرشحة	4
المعمارية المقترحة لجسر مساعف الهوائي	5
النظريات الخمس الجديدة	6
الأدوات التنفيذية الاثنتا عشرة	7
خطة MOUSSAIF1B	8
خطة MOUSSAIF9B	9
خطة MOUSSAIF-xB	10
تصميم البيانات والتقطير والتفضيلات	11
وصفات التدريب والبنية التحتية	12
التقييم ومحكمة الهدم	13
الاقتصاد الحاسوبي والتخزيني	14
برنامج التجارب وخارطة الطريق	15
الخاتمة والمراجع والملاحق	16

1. المشكلة البحثية وحدود الادعاء

1.1 السؤال المركزي

المشكلة ليست: كيف نشغّل أكبر نموذج ممكن على أصغر بطاقة رسومية فحسب؟ بل: كيف نحوّل القدرة المؤقتة المستعارة من نموذج كبير إلى معرفة مستقرة، قابلة للتدريب، ومطبوعة داخل معمارية مساعف أصغر؟ الفرق بين السؤالين هو الفرق بين العرض التقني وبين برنامج علمي. تشغيل نموذج 70B على ذاكرة 4GB قد يفتح باب التجريب، لكنه لا يصنع تلقائياً نموذجاً جديداً، ولا يثبت أن المعمارية المقترحة تعلّمت، ولا يضمن سرعة مقبولة، ولا يمنع اعتماد الباحث على نواتج معلم واحد.

تنطلق الورقة من فصل ثلاث عمليات كثيراً ما تختلط: الاستدلال Inference، والتكييف Fine-tuning، والتدريب المسبق Pretraining. الاستدلال يحتاج أوزاناً ثابتة وتمريراً أمامياً؛ التكييف يحتاج تدرجات وذاكرة تنشيطات وحالة محسّن، ولو كانت المعلمات المتعلمة قليلة؛ أما التدريب من الصفر فيحتاج كل ذلك إضافة إلى حجم بيانات وتكرارات وحوسبة هائلة. AIRLLM يحل جزءاً مهماً من المشكلة الأولى، ويمكن أن يخدم الثانية بصورة غير مباشرة عبر توليد البيانات أو التحكيم، لكنه ليس بديلاً آلياً عن منظومات التدريب.

الفرضية المؤسسة هنا هي أن الوصول إلى AGI لا يتطلب دائماً امتلاك أكبر كتلة معلمات، بل امتلاك دورة تعلم أعلى جودة: سؤال حاكمي، توليد بدائل، استخراج فروق، محكمة هدم، تغذية راجعة من أثر العالم، وذاكرة تعيد بناء نفسها. النماذج الكبيرة المفتوحة تصبح في هذا السياق «بنية مختبرية خارجية» تمد MOUSSAIF بإجابات واختلافات ومحاولات، بينما تبقى الهوية المعمارية والحاكمية في النموذج الطالب وخط تدريبه.

1.2 أطروحات يجب رفضها

- «إذا اشتغل النموذج على 4GB فقد صار رخيصاً بالكامل»: خطأ، لأن التخزين، وقراءة القرص، وذاكرة النظام، ووقت التوليد تبقى قيوداً مركزية.
- «AIRLLM يدرّب 70B على بطاقة صغيرة»: لا يثبت المشروع ذلك؛ الواجهة الأساسية ووثائقه موجهة للاستدلال الطبقي [1].
- «النموذج المفتوح يبيح كل استعمال»: خطأ؛ ينبغي التمييز بين Apache-2.0 وMIT وبين تراخيص مجتمع أو شروط مخصصة.
- «التقطير ينسخ الذكاء كله»: خطأ؛ التقطير ينتقي السلوك الذي تغطيه البيانات، وقد ينقل الأخطاء والثقة الزائفة واللغة النمطية.
- «ارتفاع نتائج benchmark يعني ذكاءً فتّاحاً»: خطأ؛ يلزم قياس القدرة على كشف الفروق، وتغيير المسار عند الدليل المضاد، واستخدام الأدوات، وحفظ قيود الصلاحية.

1.3 الإسهام الجديد

الإسهام الجديد ليس تعديل شفرة AIRLLM فقط، بل إعادة تعريف وظيفته داخل نظام بحثي: من «محمّل طبقات» إلى «بوابة معلمين حاكمية». ويقوم جسر مساعف الهوائي على أن كل استدعاء لنموذج كبير يجب أن ينتج واحداً أو أكثر من الأصول القابلة لإعادة الاستخدام: عينة تدريب، زوج تفضيل، اعتراض، مسار أداة، تقدير ثقة، اختباراً مضاداً، أو قاعدة قرار. فإذا انتهى الاستدعاء إلى نص يُقرأ ثم يُنسى، فقد استُهلكت الحوسبة بلا تراكم معرفي.

2. التشريح التقني لـ AIRLLM

2.1 المبدأ الطبقي

يصرح مستودع AIRLLM بأنه يجزئ النموذج ويحفظه طبقةً طبقة، ثم يحمّل ما يلزم للاستدلال، بحيث لا تقيم الأوزان الكاملة في VRAM في الوقت نفسه. وتشرح الوثائق أن الحاجة إلى الذاكرة الرسومية ترتبط تقريباً بحجم الطبقة النشطة لا بالحجم الكلي للنموذج، مع دعم عائلات مثل Qwen وLlama وDeepSeek وMistral وPhi وGemma [1]. هذه الفكرة شبيهة بمبدأ streaming weights: القرص أو ذاكرة المضيف يحتفظان بالكتلة الكبرى، والGPU يعمل كمحرك حساب عابر.

يمكن تقريب الذاكرة المقيمة في التمرير الأمامي بالعلاقة المقترحة: $M_GPU \approx \max_l(W_l + A_l + K_l + B_l)$، حيث W_l أوزان الطبقة النشطة، وA_l تنشيطاتها، وK_l حصة KV cache ذات الصلة، وB_l مخازن التنفيذ. أما الاستدلال التقليدي فيميل إلى $M_GPU \approx \Sigma_l W_l + \text{activations} + \text{KV}$. هذه الصيغة ليست ضماناً رقمياً، لكنها تفسر لماذا يمكن تخفيض VRAM جذرياً مع بقاء حجم النموذج على القرص كبيراً.

وتشير وثائق المشروع إلى أن النموذج يُفكك أول مرة إلى شظايا طبقية وأن على المستخدم توفير مساحة قرص كافية. كما تنص على أن عنق الزجاجة في نمطها هو تحميل الأوزان من القرص، ولذلك تضيف ضغط أوزان bit-4 أو bit-8 لتقليل حجم القراءة وتسريعها [1]. هذا يوضح المفاضلة: نحن لا نخو تكلفة النموذج؛ نحن ننقلها من إقامة متزامنة في GPU إلى حركة متكررة عبر طبقات التخزين والنقل.

2.2 ما الذي خُفّض وما الذي لم يُخفّض؟

النتيجة البحثية	أثر AIRLLM	المورد
يفتح اختبار معلمين كبار على بطاقات صغيرة	يُخفض بقوة عبر إبقاء طبقات محدودة فقط	VRAM
يلزم NVMe واسع ومنظم	لا تختفي؛ قد تزيد مؤقتاً بسبب الشظايا	مساحة القرص
غير مناسب افتراضياً للخدمة السريعة عالية الطلب	قد يصبح عنق الزجاجة الرئيسي	زمن الإدخال/الإخراج
يجب قياس الذاكرة الفعلية لا الاكتفاء بـVRAM	قد تستخدم بحسب الإعداد والنموذج، لكنها ليست بديلاً عن القرص	RAM
السياقات الطويلة قد تقلب معادلة الذاكرة	يبقى مرتبطاً بالسياق والتوليد	KV cache
التدريب يحتاج PEFT/ZeRO/FSDP أو منظومة جديدة	لا يعالج التنشيطات والتدرجات وحالات المحسّن لكل تدريب كامل	ذاكرة التدريب
يجب قياس أثر الضغط لكل مهمة	النسخة غير المضغوطة لا تحتاج quantization؛ ويمكن تفعيل ضغط اختياري	الدقة

2.3 دعم النماذج والنسخة الراهنة

بحسب تحديث المستودع في يونيو 2026، أضاف AIRLLM v3.0 دعماً لـFP8 وعائلات أحدث، وذكر أمثلة تشغيل ضخمة جداً بذاكرة GPU محدودة. هذه الأرقام ينبغي التعامل معها كادعاءات مشروع وتقارير تشغيل، لا كقياس مستقل للسرعة أو الكلفة الكلية [1]. وفي برنامج مساعف، لا نحتاج بناء النظرية على أقصى رقم تسويقي؛ يكفينا أن المبدأ يتيح تشغيل معلم أكبر بكثير من ذاكرة البطاقة، ثم نقيس throughput وlatency والطاقة ومساحة القرص على عتادنا الفعلي.

2.4 تصادم الاسم AirLLM

يوجد بحث آخر صدر سنة 2025 باسم AirLLM يتعلق بسياسة LoRA تكيفية للضبط الدقيق عن بُعد فوق قناة اتصال، وهو مختلف عن مكتبة lyogavin/airllm موضوع هذه الورقة. لذلك تستخدم الورقة كتابة AIRLLM للإطار البرمجي للاستدلال الطبقي، وتذكر البحث الآخر فقط عند الحاجة منعاً للخلط.

3. موضع AIRLLM داخل دورة التدريب

3.1 القاعدة الذهبية: AIRLLM قبل التدريب وبعده، لا مكانه

أفضل موضع لـAIRLLM هو جانبا خط التدريب: قبله لتوليد المنهج والبيانات والأحكام، وبعده لتقييم الطالب ومقارنته بمعلمين متعددين. أما قلب الخط -حساب الخسارة والرجوع الخلفي وتحديث الأوزان- فيُنفذ بأدوات تدريب مصممة لهذا الغرض. PEFT يدرب عدداً صغيراً من المعلمات الإضافية مع تجميد النموذج الأساسي [2]؛ وbitsandbytes يوفر QLoRA عبر نموذج أساسي مكمم bit-4 ومحولات LoRA قابلة للتدريب [3]؛ وFSDP يشطر المعلمات والتدرجات وحالات المحسّن عبر الأجهزة [8]؛ وZeRO-Offload ينقل حالات المحسّن أو المعلمات إلى CPU/NVMe لتوسيع الحجم الممكن [9].

3.2 الوظائف الست لـAIRLLM في مشروع مساعف

التطبيق	الوظيفة
تشغيل نموذج 8B/14B/32B أو أكبر لإنشاء مسائل متدرجة، ونسخ عربية، وحالات مضادة، ومعايير تقييم.	1. مولد المنهج
إنتاج عدة حلول بدرجات حرارة وبذور مختلفة ثم استخراج الفروق الحاكمة بينها.	2. الأستاذ متعدد الإجابات
تشغيل عائلات مختلفة لمراجعة إجابات الطالب وعدم ترك نموذج واحد حكماً نهائياً.	3. محكمة التحقق
صناعة chosen/rejected مع أسباب تفصيلية لتدريب DPO أو خوارزميات تفضيل أخرى.	4. مولد أزواج التفضيل
مقارنة تأثير tokenizer أو prompt protocol أو tool schema قبل إنفاق حوسبة تدريب كبيرة.	5. مختبر الهندسة المعمارية
تحويل معرفة المعلم إلى traces مختصرة، قواعد قرار، بيانات أدوات، واختبارات قابلة للتعلم.	6. عامل التقطير

3.3 ما لا ينبغي استعماله فيه

- خدمة محادثة كثيفة الطلب دون قياس throughput؛ لأن قراءة الطبقات قد تجعل زمن كل token مرتفعاً.
- التدريب الكامل من الصفر لمجرد أن النموذج صار قابلاً للتحميل؛ فالرجوع الخلفي يغير معادلة الذاكرة جذرياً.
- توليد بيانات بلا محكمة جودة؛ لأن البيانات التركيبية قد تضخم انحيازات المعلم.
- استبدال التقييم البشري أو الاختبارات الموثقة بحكم نموذج أكبر وحده.
- التعامل مع نموذج ضخم بطيء كذاكرة دائمة؛ الأفضل استخراج أصول مضغوطة يمكن فهرستها وإعادة استخدامها.

3.4 خط الإنتاج المقترح

المرحلة	المكوّن	المخرج
A	AIRLLM + معلم/معلمون كبار	مسائل، أجوبة، نقد، traces، تفضيلات
B	Difference Court Sampler	فروق حاكمة وعينات خلافية
C	dedup + provenance + فلترة بشرية/آلية	مجموعة تدريب نظيفة قابلة للتتبع
D	Full/Continued Pretraining أو QLoRA/PEFT	MOUSSAIF student checkpoint
E	اختبارات أدوات/DPO/RLVR	نسخة محاذاة وأقدر على التنفيذ
F	AIRLLM teacher court + benchmarks	تقرير فجوات وإعادة توليد المنهج
G	Sovereign Replay Buffer	تراكم الأخطاء ذات القيمة العالية للدورة التالية

4. خريطة النماذج المفتوحة المرشحة بالاسم

في هذه الورقة نستخدم «مفتوح» بدقة. النموذج المفتوح المصدر أو المفتوح بالكامل يتيح أوزاناً وكوداً وتفاصيل تدريب بترخيص معروف؛ النموذج المفتوح الأوزان قد يتيح الاستخدام والتعديل ضمن شروط مخصصة. هذا الفرق مؤثر في النشر التجاري، وإعادة التوزيع، والتقطير، وبناء مشتقات باسم MOUSSAIF.

النموذج	الحجم	الترخيص/الانفتاح	الدور المقترح
allenai/OLMo-2-0425-1B	1B	Apache-2.0؛ كود ونقاط تحقق وسجلات وتفاصيل تدريب منشورة	أفضل قاعدة علمية لمسار MOUSSAIF1B ودراسة ديناميات التدريب
Qwen/Qwen3-0.6B-Base	0.6B	Apache-2.0	مختبر سريع متعدد اللغات؛ baseline دون 1B
Qwen/Qwen3-8B-Base	8B	Apache-2.0	قاعدة عملية لمسار 8-9B أو معلم متوسط قوي
Qwen/Qwen3-14B-Base	14B	Apache-2.0	معلم للمنطق والبرمجة والتوليد متعدد اللغات
microsoft/Phi-4-mini-instruct	3.8B	MIT	معلم متخصص في الاستدلال الكثيف وfunction calling والسياق الطويل
microsoft/Phi-4-mini-reasoning	3.8B	MIT	أستاذ رياضي/استدلالي لإنتاج مسائل وحلول محكومة
mistralai/Mistral-7B-v0.3	7B	Apache-2.0	baseline بسيط وخفيف ومعروف للضبط والتقطير
deepseek-ai/DeepSeek-R1-0528-Qwen3-8B	8B	MIT؛ يصرح بدعم الاستخدام التجاري والتقطير	أستاذ reasoning لتوليد traces واختبارات مضادة
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	7B	MIT	مرجع لتأثير التقطير من نموذج reasoning أكبر
google/gemma-2-9b	9B	Gemma terms؛ gated في بعض المنصات	مرشح دقيق لحجم 9B، لكن بشروط ترخيص مخصصة
meta-llama/Llama-3.2-1B	1B	Llama Community License	baseline مشهور لمسار 1B، لكنه ليس Apache/MIT
HuggingFaceTB/SmolLM2-1.7B	1.7B	Apache-2.0	نموذج صغير مع نقاط وسطية مفيدة لدراسة التعلم

4.1 التوصية الأساسية

MOUSSAIF1B — ابدأ بـ OLMo 2 1B عندما يكون الهدف بحثاً معمارياً قابلاً للتفسير وإعادة الإنتاج؛ واستخدم Qwen3-0.6B

أو Llama 3.2 1B للمقارنة. OLMo 2 يعلن نشر الكود ونقاط التحقق والسجلات وتفاصيل التدريب تحت [4] Apache-2.0.

لـ**MOUSSAIF9B** — إن كان المقصود حجماً فعلياً قريباً من 9B، فهناك طريقان: انطلاق مشتق من Gemma 2 9B مع قبول شروطه، أو بناء نسخة 8B/9B انطلاقاً من Qwen3-8B/بنية مفتوحة ثم توسيعها أو تدريبها. بحثياً وقانونياً، Qwen3-8B Apache-2.0 أنظف كنقطة بداية [5].

لـ**MOUSSAIF-xB** — لا تربط العائلة بقاعدة واحدة. اجعل 0.6B/1B للرقابة والتوجيه، 3.8B للمهارات المتخصصة، 7-9B للعامل العام، و14B+ للمعلم أو الخبير البطيء. AIRLLM يسمح باستدعاء الأحجام الأكبر عند الحاجة بدل إبقائها مقيمة دائماً.

4.2 لماذا لا نختار نموذجاً واحداً؟

لأن كل عائلة تحمل تحيزاً معمارياً وبيانياً. Qwen3 مفيد للتعدد اللغوي والتحول بين التفكير وغير التفكير حسب بطاقة النموذج [5]؛ Phi-4-mini مصمم لبيئات محدودة الحوسبة ويقدم سياقاً طويلاً واستدلالاً كثيفاً [6]؛ OLMo 2 هو الأكثر فائدة لدراسة التدريب لأنه ينشر نقاطاً وسيطة ومكونات المسار [4]. تعدد المعلمين ليس ترفاً: إنه وسيلة لكشف ما يعتقده نموذج واحد حقيقة عامة بينما هو أثر بيانات أو tokenizer أو post-training.

5. المعمارية المقترحة: جسر مساعف الهوائي

5.1 تعريف المعمارية

جسر مساعف الهوائي هو طبقة وسيطة بين مستودع النماذج المفتوحة وبين خط تدريب MOUSSAIF. وظيفته ليست إخفاء اختلافات النماذج، بل تحويلها إلى إشارات تدريب. يتكون من: سجل نماذج وترخيص؛ مخطط موارد يحدد GPU/RAM/NVMe؛ منفذ AIRLLM؛ مولد مهام؛ جامع traces؛ محلل فروق؛ محكمة تحقق؛ مصنع بيانات؛ ومدير تجارب يربط كل عينة بالمعلم والإصدار والإعداد.

5.2 مخطط التدفق

المستوى	العقدة	العملية الحاكمة
0	سجل النماذج	الترخيص، الحجم، tokenizer، نوع المعمارية، تكلفة التشغيل
1	AIRLLM Runtime	تقسيم طبقي، تحميل متتابع، ضغط اختياري، قياس I/O
2	Teacher Ensemble	توليد N إجابات ومسارات أدوات واعتراضات
3	MCD/Q-MCD Scoring	قياس الفرق، الدليل، التكامل، الأثر، الضجيج، وكثافة السؤال
4	Falsification Court	اختبارات مضادة، إعادة صياغة، تعارض مصادر، حدود الصلاحية
5	Dataset Forge	SFT، preference pairs، tool traces، negative examples، curricula
6	Training Engine	PEFT/QLoRA/FSDP/ZeRO/continued pretraining
7	Student Evaluation	اختبارات قدرة، calibration، robustness، world-effect
8	Self-Reconstruction Memory	إعادة إدخال الفشل المهم في المنهج التالي

5.3 معادلة العائد من استدعاء المعلم

نقترح مقياس «عائد المعلم الهوائي» ATRY: $ATRY = (D \times V \times R \times U) / (T \times I \times S)$، حيث D تنوع العينات، V نسبة العينات المجتازة للتحقق، R قابلية إعادة الاستخدام، U مقدار التحسن المقاس في الطالب، T زمن التوليد، I كلفة الإدخال/الإخراج، وS مساحة التخزين. لا يهم أن يعمل 235B على 3GB إن كان ينتج token كل زمن طويل ولا يضيف تحسناً للطالب. المقارنة الصحيحة بين المعلمين تكون بالعائد التدريبي النهائي، لا بالانبهار بالحجم.

5.4 أنماط التشغيل الثلاثة

- النمط الاستكشافي: عينات قليلة عالية التنوع لاكتشاف أنواع الحلول والأخطاء، يقبل بطئاً أكبر.
- النمط المصنعي: دفعات كبيرة متكررة، يحتاج caching ،pre-sharding، ضغطاً محسوباً، وجدولة ليلية.
- النمط القضائي: يُستدعى النموذج الكبير فقط عند خلاف المعلمين الصغار أو انخفاض الثقة، لتقليل الكلفة.

6. النظريات الخمس الجديدة باسم مساعف

6.1 نظرية مساعف للجسر الهوائي الحاكمي (M-AGB)

تنص النظرية على أن الاستدلال منخفض الذاكرة يكتسب قيمة AGI فقط حين يتحول من إجابة عابرة إلى مورد تعلم محكوم. أي أن نجاح AIRLLM لا يقاس بأكبر نموذج تم تحميله، بل بنسبة المخرجات التي تدخل في دورة: توليد ← فرق ← تحقق ← تدريب ← أثر ← ذاكرة. وتفرق النظرية بين «إتاحة القدرة» Access to Capability وبين «امتلاك القدرة» Internalized Capability. الأولى تتحقق عندما يستطيع النظام استدعاء معلم كبير، والثانية عندما يظهر السلوك في الطالب دون المعلم أو بميزانية أقل.

- قانون العبور: لا يُقبل استدعاء معلم بلا artifact محفوظ.
- قانون الاستقلال: يجب قياس الطالب دون المعلم بعد كل دورة.
- قانون الأثر: لا تُحتسب العينة مفيدة إلا إذا غيرت metric أو كشفت فشلاً.

6.2 نظرية مساعف للأستاذ متعدد الأحجام (M-MST)

ترفض النظرية صورة الأستاذ الواحد. المعرفة التي نحتاجها في MOUSSAIF موزعة بين أحجام وعائلات وتخصصات: نموذج صغير قد يكون أكثر انضباطاً في التنسيق، وآخر متوسط أفضل في الأدوات، وكبير أفضل في تركيب الفرضيات. لذلك يُبنى Ensemble غير متساوٍ، ويُعطى لكل معلم نطاق سلطة. نموذج 1B لا يحكم على برهان فلسفي نهائي، لكن يمكنه كشف مخالفة schema، نموذج reasoning لا يحكم وحده على الوقائع الزمنية، ونموذج multilingual يراجع العربية والمعنى.

- قانون السلطة المحدودة: لكل معلم مجال حكم ودرجة ثقة.
- قانون الخلاف المنتج: الخلاف بين المعلمين يولد عينات أعلى قيمة.
- قانون الحكم المركب: لا تُقبل الإجابة الحساسة إلا بتوافق دليل ونقد.

6.3 نظرية مساعف للتقطير التفاضلي الحاكمي (M-GDD)

التقطير التقليدي يركز على مطابقة logits أو تقليد الإجابة. أما التقطير التفاضلي الحاكمي فيركز على نقل «الفروق التي غيرت القرار»: لماذا رُفض مسار؟ أي قيد كان أعلى؟ ما الدليل الذي قلب الحكم؟ ما الأداة التي أزالت عدم اليقين؟ تُجزأ عينة المعلم إلى: حالة، بدائل، فروق، ترتيب حاكمي، حكم، أثر، وحدود صلاحية. بهذا لا يتعلم الطالب نصاً طويلاً فقط، بل يتعلم مواضع التحول.

- قانون فرق القرار: الأولوية للعينات التي يتبدل فيها الحكم بتغير صغير ذي معنى.
- قانون السالب المفسر: العينة المرفوضة يجب أن تحمل سبب رفض قابل للتعلم.
- قانون الاختصار الأمين: يُضغط trace دون حذف الفرق الذي حسم النتيجة.

6.4 نظرية مساعف لسلم الأحجام الوظيفي (M-FSS)

المجم في MOUSSAIF ليس رتبة شرف، بل وظيفة. تقترح النظرية توزيع العمل: 1B-0.3 للمراقبة والتصنيف والتوجيه، 4B-1 للمهارات المحدودة والذاكرة المحلية، 9B-7 للوكيل العام، 32B-14 للمعلم المتخصص، والأحجام الأكبر للقضاء النادر أو توليد المناهج. يمكن أن تكون منظومة xB أذكى اقتصادياً من نموذج واحد ضخم إذا كان التوجيه صحيحاً.

- قانون أصغر كافٍ: اختر أصغر نموذج يحقق threshold المهمة.
- قانون التصعيد: انتقل للحجم الأعلى عند فشل قابل للقياس لا عند الشعور.
- قانون الاسترداد: يجب أن يعيد المعلم الكبير معرفة قابلة لضغط كلفة استدعائه لاحقاً.

6.5 نظرية مساعف للذاكرة السيادية العابرة للطبقات (M-TSM)

AIRLLM يمرر التنشيط عبر طبقات النموذج، لكن مشروع مساعف يحتاج طبقة ثانية من الذاكرة خارج الأوزان: ذاكرة تحفظ أسباب القرار، مصادر العينة، نسخة المعلم، إعدادات التوليد، نتائج المحاكم، وأثر التدريب. تسمي النظرية هذه البنية «الذاكرة السيادية العابرة للطبقات» لأنها تربط طبقة النموذج بطبقة التجربة وبطبقة الحاكمية. وهي تمنع ضياع المعرفة عندما يتغير checkpoint أو يُحذف معلم.

- قانون الأصل: كل عينة تحمل provenance كاملاً.
- قانون القابلية للعكس: يمكن إعادة بناء سبب إدخال العينة أو حذفها.
- قانون الذاكرة الحية: الفشل المتكرر يرفع أولوية التدريب تلقائياً.

7. الأدوات التنفيذية الاثنتا عشرة

الأداة	الوظيفة
1. AirTeacher Forge	منسق تشغيل AIRLLM يقرأ سجل النماذج، يختار المعلم، ينفذ pre-sharding، ويسجل الزمن والذاكرة والقرص لكل job.
2. Fattah Trace Synthesizer	يطلب من المعلم إخراج مسارات منظمة: فرضيات، فروق، أدوات، اعتراضات، حكم، وثقة؛ ثم يحولها إلى JSONL.
3. Difference Court Sampler	ينتقي الأسئلة التي تختلف فيها النماذج أو تتبدل إجاباتها بتغيير صغير، لأنها أغنى للتعلم من الأسئلة السهلة.
4. Cross-Model Consensus Judge	يجمع أحكام نماذج متعددة مع وزن حسب المجال والترخيص والإصدار، ويمنع نموذجاً واحداً من فرض الحقيقة.
5. Adapter Lattice	شبكة محولات LoRA منفصلة للحاكمية، الأدوات، العربية، البحث، المنطق، والذاكرة، مع تركيب/فصل بحسب المهمة.
6. xB Curriculum Compiler	يحول مجموعة المهارات إلى منهج متدرج موزع على أحجام الأسرة، ويحدد ما يتعلمه كل حجم وما يصعّد.
7. Sovereign Replay Buffer	مخزن للفشل عالي القيمة: هلوسة، تجاهل قيد، خطأ أداة، ثقة زائدة، تعارض أدلة؛ يعاد حقنه في التدريب.
8. Layer Budget Profiler	يقيس أكبر طبقة، زمن قراءة shard، VRAM، RAM، KV cache، وtokens/s لتحديد هل AIRLLM مفيد فعلاً للنموذج.
9. License-Aware Model Registry	يفرض قواعد آلية: Apache/MIT للمشتقات الحرة؛ تراخيص مخصصة للتجارب المقيدة؛ منع خلط بيانات أو أوزان غير متوافقة.
10. Falsification Pack Generator	ينشئ counterfactuals، paraphrases، اختبارات تناقض، إزالة دليل، وتغيير أولوية لمعرفة هل الحكم حقيقي أم هش.
11. World-Effect Evaluator	يقيس نجاح تنفيذ الأداة أو الكود أو البحث لا جمال النص فقط؛ يربط الإجابة بنتيجة قابلة للتحقق.
12. AIRLLM Benchmark Gate	بوابة قرار تقارن AIRLLM بـ llama.cpp/vLLM/Transformers offload من حيث السرعة والطاقة والجودة قبل اعتماده.

7.1 مواصفة البيانات الموحدة

```
{
  "sample_id": "moussaif-000001",
  "task": "...",
  "teacher_model": "Qwen/Qwen3-14B",
  "runtime": "AIRLLM-v3",
  "license_class": "apache-2.0",
  "candidates": [...],
  "governing_differences": [...],
  "tools": [...],
```

```
  "verifiers": [...],
  "chosen": "...",
  "rejected": "...",
  "rejection_reason": "...",
  "confidence": 0.0,
  "provenance": {...},
  "student_eval_before": {...},
  "student_eval_after": {...}
}
```

هذه البنية تجعل AIRLLM جزءاً من MLOps معرفي. دون الحقول الخاصة بالترخيص والمصدر والإعداد والتحقق، يصبح من المستحيل لاحقاً تفسير تحسن أو تراجع النموذج، أو حذف بيانات معلم معين، أو إعادة إنتاج التجربة.

8. خطة MOUSSAIF1B: مختبر الذكاء الفتّاح المكثف

8.1 لماذا 1B استراتيجي؟

النموذج 1B ليس نسخة فقيرة من 9B، بل منصة قياس. يمكن إعادة تدريبه أو تكييفه مراراً، تشغيل ablations أكثر، فحص نقاط وسطية، واكتشاف هل التحسين نابع من النظرية أو من ضخ البيانات. OLMo 2 1B مناسب خصوصاً لأنه ينشر نقاط تحقق وتفاصيل تدريب، ويتيح fine-tuning من checkpoint نهائي أو وسيط [4].

هدف MOUSSAIF1B ليس تخزين كل العالم، بل إتقان البنية الحاكمية: تحديد نوع السؤال، استخراج القيود، طلب الأداة، تمييز المعرفة من التخمين، التراجع عند الدليل، واختيار متى يصعّد المهمة إلى 9B أو معلم AIRLLM. إذا حقق 1B هذه الوظائف، صار Kernel ذا قيمة حتى لو احتاج RAG أو نموذجاً أكبر للمعرفة الواسعة.

8.2 الوصفة المقترحة

المكوّن	التنفيذ
القاعدة	OLMo 2 1B أو Qwen3-0.6B/1B-class للمقارنة.
المرحلة A	Continued pretraining على Corpus عربي/إنجليزي منظم حول الفروق والحاكمية والأدوات، لا تبدأ بتعليمات فقط.
المرحلة B	SFT على 100k-500k عينة عالية الجودة: تحليل سؤال، استخراج قيود، اختيار أداة، جواب موثق.
المرحلة C	تقطير من Qwen3-8B/14B وPhi-4-mini وDeepSeek-R1- عبر distill AIRLLM.
المرحلة D	DPO على أزواج تفضيل يكون سبب الرفض فيها واضحاً: هلوسة، تجاوز صلاحية، تجاهل فرق.
المرحلة E	RLVR أو تحقق برمجي للمهام التي لها جواب قابل للحساب/التنفيذ.
المرحلة F	تكميم ونشر، ثم جمع failure telemetry وإعادة التدريب.

8.3 نسب البيانات المقترحة

فئة البيانات	النسبة الابتدائية	مثال
اللغة والمعرفة العامة	25%	نص موثوق، تلخيص، سؤال جواب
الاستدلال والفروق	20%	تمييز مسارين متقاربين، counterfactuals
الحاكمية والثقة	15%	حدود الصلاحية، طلب دليل، تراجع
الأدوات والبحث	15%	function calling، retrieval، code execution
العربية متعددة المستويات	10%	فصحى، مصطلح تقني، شرح مبسط

تلخيص state، استعادة قرار سابق	10%	الذاكرة والسياق
رفض مسبب مع بديل مفيد	5%	السلامة والرفض الدقيق

8.4 معيار النجاح

لا يُعلن نجاح MOUSSAIF1B لأنه اقترب من 8B في MMLU فقط. النجاح هو أن يحقق: دقة عالية في اختيار الأداة، انخفاض hallucination under abstention، calibration أفضل، ثباتاً عند إعادة الصياغة، وقدرة على تصعيد المهمة بدلاً من اختلاق جواب. هذا النموذج يصبح «حاكماً صغيراً» أو router حتى عندما لا يكون أفضل مولد نص.

9. خطة MOUSSAIF9B: نموذج القدرة الوسطى الحاكمية

9.1 ثلاثة معانٍ لـ9B

- اسم إصدار معماري: قد يكون النموذج الفعلي 8B أو 1B، ويجب التصريح بذلك.
- حجم فعلي قريب من تسعة مليارات معلمة: يتطلب اختيار بنية وقاعدة وتدريباً يتوافقان مع هذا الرقم.
- منظومة قدرة تعادل 9B: نموذج أصغر مع RAG وأدوات وخبراء، لا ينبغي تسويقها كـ9B parameters.

9.2 المسارات الثلاثة

المسار	الميزة	المخاطرة	التوصية
مشتق Qwen3-8B	Apache-2.0؛ قريب حجماً؛ متعدد اللغات	ليس 9B حرفياً؛ يحتاج تكييفاً عميقاً للهوية	المسار الأسرع والأكثر نظافة
مشتق Gemma 2 9B	حجم 9B حقيقي وbaseline قوي	ترخيص Gemma وشروط الوصول/التوزيع	للبحث المقيد أو إن كانت الشروط مناسبة
تدريب 9B جديد من الصفر	ملكية معمارية كاملة	كلفة بيانات وحوسبة وهندسة عالية جداً	مرحلة لاحقة بعد إثبات 1B و8B

9.3 أين يفيد AIRLLM؟

قبل امتلاك 9B، يمكن تشغيل Qwen3-14B أو 32B أو نماذج reasoning أكبر كمعلمين لتوليد منهج 9B. وبعد ضبط 9B، يمكن تشغيل نماذج أكبر كقضاة على حالات الخلاف. لكنه لا يجعل full fine-tuning لـ9B ممكناً على 4GB تلقائياً. عملياً، QLoRA قد يسمح بضبط 8-9B على بطاقة استهلاكية ذات VRAM مناسبة لأن الأوزان الأساسية مجمدة ومكممة، بينما ZeRO-Offload أو FSDP يدعمان التدريب الأوسع حسب العتاد [3][8][9].

9.4 بنية مهارات 9B

الهدف	المجال
ترتيب القيود والقيم، تفسير سبب الحكم، تصحيح المسار	الحاكمية العليا
سلاسل قصيرة موثوقة + بحث شجري عند الحاجة	الاستدلال المركب
بحث، كود، ملفات، قواعد بيانات، APIs، ذاكرة	الأدوات
انتقال دقيق دون فقد المصطلح أو السياق	اللغة العربية/الإنجليزية
خطة، تنفيذ، مراقبة، recovery، إغلاق المهمة	Agentic planning
ضغط السياق، تمييز الثابت من المؤقت، provenance	الذاكرة
كشف الجهل، طلب تحقق، تقدير ثقة calibrated	التقييم الذاتي

9.5 فرضية المعمارية

نقترح أن يكون MOUSSAIF9B dense decoder بذاكرة سياقية خارجية وتوجيه مهارات sparse، بدلاً من محاولة بناء MoE ضخم مبكراً. يمكن إدخال LoRA/IA3 adapters لكل مجال ثم دمج ما يثبت فائدته. وعند الانتقال لاحقاً إلى xB، تتحول بعض المحولات إلى خبراء أو وحدات منفصلة. هذا يقلل خطر تعقيد المعمارية قبل امتلاك قياسات ثابتة.

10. خطة MOUSSAIF-xB: عائلة أحجام لا صنم حجم

10.1 التعريف

MOUSSAIF-xB هو بروتوكول بناء عائلة: قواعد تسمية، schemas، tokenizer، ذاكرة، محاكم، وواجهات أدوات مشتركة بين نماذج مختلفة الحجم. x لا يعني رقماً مجهولاً فقط؛ بل يعني أن النظام يقرر الحجم المناسب في لحظة التنفيذ والتدريب.

10.2 السلم الوظيفي

الطبقة	حجم تقريبي	الوظيفة
Sentinel	0.1-0.6B	كشف schema، أمان، تصنيف، cache، فلترة
Governor	0.6-1.7B	توجيه، تقدير صعوبة، اختيار أداة، ضغط سياق
Specialist	3-4B	رياضيات/كود/لغة/بحث متخصص
General Agent	7-9B	حل عام، تخطيط، أدوات، ذاكرة
Senior Teacher	14-32B	توليد مناهج ونقد معقد وتقطير
Rare Court	AIRLLM عبر +70B	تحكيم نادر، red-team، مسائل صعبة جداً

10.3 التوجيه الحاكمي

يحسب Router المقترح درجة صعوبة H من عمق الاستدلال، طول السياق، حاجة الأداة، حساسية الخطأ، وحداثة المعرفة. ثم يختار أصغر نموذج يحقق احتمال نجاح مستهدف. إذا انخفضت الثقة أو اختلفت نتائج المتخصصين، يصعد إلى General Agent أو Rare Court. هذه البنية تجعل AIRLLM صمام تصعيد لا محركاً دائماً، فتستفيد من النموذج الضخم دون دفع كلفته لكل سؤال.

10.4 تقطير متبادل

العلاقة ليست من الكبير إلى الصغير فقط. النموذج الصغير يجمع telemetry عن الأنماط المتكررة، ويقترح على المعلم أين يركز التوليد، والمتخصص يرسل اختبارات إلى العام، والعام يعيد حالات فشل إلى الأستاذ. نسمي ذلك «التقطير الدوراني»: البيانات تتحرك صعوداً وهبوطاً، بينما القرار النهائي حول ما يدخل الأوزان يخضع للمحكمة.

11. تصميم البيانات والتقطير والتفضيلات

11.1 من الإجابة إلى وحدة تعلم

الإجابة الخام ليست وحدة تعلم كافية. الوحدة المثالية في مشروع مساعف تتضمن: سؤالاً أصلياً، سياقاً، قيوداً صريحة وضمنية، ثلاثة بدائل على الأقل، الفرق الحاكم بين البدائل، أداة أو مصدر عند الحاجة، إجابة مختارة، سبب رفض الآخرين، حدود الثقة، واختباراً مضاداً. هذه البنية تسمح باستخدام العينة في SFT وDPO والتقييم معاً.

11.2 تقطير النواتج مقابل تقطير العمليات

النوع	ما يتعلمه الطالب	الخطر	العلاج
Output distillation	الإجابة النهائية	تقليد أسلوب بلا فهم	إضافة بدائل وأسباب
Logit distillation	توزيع الاحتمالات	صعب مع APIs/أطر مختلفة	يستخدم عند امتلاك logits
Rationale distillation	مسار تفسير	قد يحتوي هلوسة أو إطالة	تحقق واختصار حاكمي
Tool-trace distillation	متى وكيف تُستخدم الأداة	تعلم هشة calls	اختبارات تنفيذية
Preference distillation	ترتيب إجابتين	تحيز القاضي	قضاة متعددون وسبب رفض
Difference distillation	الفروق التي غيرت القرار	يتطلب ترميزاً دقيقاً	MCD/Q-MCD schema

11.3 SFT ثم DPO

تدريب SFT يبني السلوك الأساسي: التنسيق، الأدوات، اللغة، والالتزام بالتعليمات. ثم تستخدم أزواج التفضيل لتعليم النموذج أن إجابة موثقة متحفظة أفضل من إجابة واثقة مختلقة. DPO يقدم طريقة مباشرة لتفضيل chosen على rejected دون بناء reward model مستقل معقد [13]. لكن جودة DPO لا تتجاوز جودة الأزواج؛ لذلك يجب أن تُبنى الأزواج حول فروق حقيقية لا حول أسلوب لغوي.

11.4 QLoRA كمسار عملي

LoRA يجمد أوزان النموذج ويضيف مصفوفات منخفضة الرتبة قابلة للتعلم، ما يقلل عدد المعلمات والتكلفة [11]. QLoRA يمرر التدرجات عبر نموذج أساسي مجمد ومكمم 4-bit إلى محولات LoRA، مع تقنيات مثل NF4 وdouble quantization وpaged optimizers [10]. لمساعف، هذا يعني إمكان بناء adapters مستقلة: Research، Memory، Tools، Arabic، Governance. لا ينبغي دمجها نهائياً قبل تقييم أثر كل واحدة منفردة وتداخلها.

11.5 جودة البيانات التركيبية

- التنوع: معلمون وعينات ودرجات حرارة وصياغات مختلفة.
- التحقق: اختبارات تنفيذية أو مصادر أو قضاة مستقلون.
- التوازن: عدم إغراق corpus بإجابات طويلة reasoning-like على حساب الأجوبة المباشرة.
- التلوث: منع تسرب benchmark إلى التدريب أو تسجيله إن استُخدم.

- التتبع: نموذج/نسخة/إعداد/ترخيص/تاريخ لكل عينة.
- الاقتصاد: حذف العينات المتكررة حتى لو كانت جميلة؛ الأولوية للمعلومة الحدية الجديدة.

12. وصفات التدريب والبنية التحتية

12.1 وصفة AIRLLM للاستدلال

```
from airllm import AutoModel

model = AutoModel.from_pretrained(
    "Qwen/Qwen3-14B",
    # compression="4bit",   # فعّله بعد قياس أثر الجودة
    layer_shards_saving_path="/nvme/airllm_shards/qwen3_14b"
)

prompt = "حلّل المسألة وأخرج: البدائل، الفروق الحاكمة، الأدوات، الحكم، وحدود الثقة."
inputs = model.tokenizer([prompt], return_tensors="pt", padding=False)
out = model.generate(inputs["input_ids"].cuda(), max_new_tokens=512, use_cache=True)
print(model.tokenizer.decode(out.sequences[0], skip_special_tokens=True))
```

الكود توضيحي ويجب مواءمته مع نسخة AIRLLM والنموذج. أهم ممارسة ليست السطر البرمجي، بل حفظ shards على NVMe، تثبيت الإصدار، تسجيل peak VRAM/RAM، وقياس الزمن من prompt إلى آخر token.

12.2 وصفة QLoRA/SFT

```
from datasets import load_dataset
from transformers import AutoTokenizer, AutoModelForCausalLM, BitsAndBytesConfig
from peft import LoraConfig
from trl import SFTConfig, SFTTrainer
import torch

base = "Qwen/Qwen3-8B-Base"
bnb = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_use_double_quant=True,
    bnb_4bit_compute_dtype=torch.bfloat16,
)
model = AutoModelForCausalLM.from_pretrained(base, quantization_config=bnb, device_map="auto")
tok = AutoTokenizer.from_pretrained(base)
peft = LoraConfig(r=32, lora_alpha=64, lora_dropout=0.05,
                  target_modules="all-linear", task_type="CAUSAL_LM")
trainer = SFTTrainer(
    model=model,
    tokenizer=tok,
    train_dataset=load_dataset("json", data_files="moussaif_sft.jsonl")["train"],
    peft_config=peft,
    args=SFTConfig(output_dir="moussaif9b-sft", max_length=4096,
                   gradient_accumulation_steps=16, bf16=True)
```

```
)
trainer.train()
```

توثق TRL وجود SFTTrainer، كما تدعم PEFT تكامل LoRA وQLoRA [7]. الإعدادات أعلاه نقطة انطلاق لا وصفة مضمونة؛ الرتبة والطبقات المستهدفة وطول السياق والbatch يجب أن تُضبط بتجارب صغيرة.

12.3 متى نستخدم FSDP؟

يستخدم FSDP عندما نملك عدة GPUs ونريد شطر المعلمات والتدرجات وحالات المحسّن، أو عند full/continued training يتجاوز بطاقة واحدة [8]. لا يضيف فائدة كبيرة إذا كان التدريب QLoRA صغيراً جداً ويعمل على GPU واحدة، وقد يزيد التعقيد. أما ZeRO-Offload فيفيد عندما تكون ذاكرة CPU/NVMe كبيرة ونقبل مفاضلة السرعة مقابل الحجم [9].

12.4 مصفوفة القرار

البديل	الأداة الأولى	المهمة
Transformers/Accelerate disk offload	AIRLLM	تشغيل معلم أكبر من VRAM
llama.cpp	vLLM/SGLang	خدمة سريعة لنموذج يناسب الذاكرة
LoRA bf16	QLoRA + PEFT	ضبط 1B-9B ببطاقة واحدة
DeepSpeed ZeRO-3	FSDP2	تدريب متعدد GPUs
FSDP CPU offload	ZeRO-Offload/Infinity	تدريب مع CPU/NVMe offload
أصغر student	GGUF/ONNX/quantized	نشر edge
API/cluster inference	AIRLLM batch jobs	توليد بيانات معلمين

13. التقييم ومحكمة الهدم

13.1 طبقات التقييم

الطبقة	المقاييس
لغة ومعرفة	factuality، العربية، التلخيص، perplexity، QA
استدلال	رياضيات، منطق، counterfactual consistency، coding
حاكمية	constraint adherence، abstention، calibrated confidence، reversal under evidence
أدوات	tool selection، argument correctness، execution success، recovery
سياق وذاكرة	needle tests، compression fidelity، state reconstruction
Agentic	plan success، steps، cost، error recovery، completion
اقتصاد	tokens/s، joules/token، VRAM/RAM، NVMe read، cost per accepted sample

13.2 محكمة الهدم الخاصة بـAIRLLM

- هل المقارنة تعرض VRAM فقط وتتجاهل القرص والوقت والطاقة؟
- هل النموذج يعمل بنفس الدقة أم بضغط FP8/ مختلف؟
- هل الزمن مقاس بعد اكتمال تقسيم الطبقات أم يشمل التحضير الأول؟
- هل يزداد زمن التوليد خطياً/أسوأ مع طول output؟
- هل KV cache أو السياق الطويل يكسر الذاكرة المعلنة؟
- هل throughput مناسب لتوليد مئات آلاف العينات؟
- هل نموذج أصغر في vLLM ينتج عينات مقبولة أكثر لكل ساعة؟
- هل التقطير حسّن الطالب في test غير مستخدم في التوليد؟
- هل حكم المعلم متوافق مع مصادر أو execution؟
- هل يمكن حذف بيانات معلم واحد وإعادة بناء النموذج قانونياً وتقنياً؟

13.3 تجارب الفصل السببي

ما الذي نثبته؟	B	A	التجربة
العائد التدريبي لكل ساعة	معلم 7-8B سريع في vLLM	معلم 14B عبر AIRLLM	AIRLLM vs عادي
أثر التنوع والخلاف	Qwen+Phi+DeepSeek	فقط Qwen	معلم واحد vs متعدد
قيمة التقطير الحاكمي	جواب+فروق+رفض	جواب نهائي	Output vs Difference distillation

أثر أزواج الرفض	إضافة تفضيلات	SFT	SFT vs SFT+DPO
التداخل والنسيان	دمج مبكر	كل مجال منفصل	Adapters منفصلة vs مدمجة
الاقتصاد والنجاح	حل كل شيء بـ8B	Router 1B	1B vs 8B router

13.4 شروط رفض الفرضية

تُرفض أطروحة فائدة AIRLLM لمساعف إذا أظهرت التجارب أن نموذجاً أصغر سريعاً يولد عدداً أكبر من العينات المقبولة لكل وحدة وقت وكلفة، وأن المعلم الضخم لا يضيف تحسناً مستقلاً في الطالب، أو إذا كانت حركة القرص تجعل الدورة غير عملية، أو إذا كان دعم النموذج هشاً وغير قابل لإعادة الإنتاج. النظرية العلمية لا تحمي الأداة؛ بل تسمح بإسقاطها عندما لا تحقق عائداً.

14. الاقتصاد الحاسوبي والتخزيني

14.1 ميزانية لا تعتمد على VRAM وحدها

المعادلة التشغيلية يجب أن تجمع: حجم checkpoint، حجم shards، مساحة cache، سرعة VRAM، RAM، NVMe، استهلاك الطاقة، زمن التوليد، ونسبة العينات المقبولة. في AIRLLM قد يكون NVMe أسرع استثمار من GPU أكبر عندما يكون الهدف تجريباً بطيئاً، أما في مصنع بيانات واسع فقد تصبح بطاقة أكبر تشغل 14B مكمماً بالكامل أسرع وأرخص على المدى الطويل.

14.2 مستويات مختبر مقترحة

المستوى	العتاد التقريبي	الاستعمال
L0	CPU + 32GB RAM + SSD	اختبار 0.6B/1B، بناء البيانات، لا إنتاج كثيف
L1	GPU 8-12GB + 64GB RAM + 2TB NVMe	AIRLLM معلمين متوسطين، QLoRA 1B-4B
L2	GPU 16-24GB + 64-128GB RAM + 4TB NVMe	QLoRA 8-9B، تشغيل 14B/32B بطرق متعددة
L3	GPUs 2-4 أو سحابة مؤقتة	FSDP/ZeRO، continued training واسع، batch inference
L4	عنقود	تدريب 9B من الصفر أو سياقات/بيانات ضخمة

14.3 مقياس الكلفة الحقيقي

نقترح Cost per Verified Learning Unit (CVLU): مجموع كلفة الحوسبة والتخزين والوقت البشري مقسوماً على عدد وحدات التعلم التي اجتازت التحقق وأنتجت تحسناً قابلاً للقياس. هذا المقياس قد يثبت أن 70B عبر AIRLLM مفيد فقط في 2% من الحالات الصعبة، بينما 8B سريع يخدم 98% من الإنتاج. وهنا يصبح التصعيد الحاكمي هو التصميم الأمثل.

14.4 تحسينات عملية

- pre-shard مرة واحدة لكل نسخة نموذج، مع checksum ومسار ثابت.
- استخدم NVMe محلياً لا قرصاً شبكياً بطيئاً عند streaming.
- جمّع prompts ذات طول متقارب، لكن اختبر batching فعلياً لأن دعم الإطار قد يختلف.
- فعّل 4-bit/8-bit فقط بعد اختبار جودة على مجموعة حرجة.
- خزّن نواتج المعلم بصيغة مضغوطة وdeduplicated، لا تعِد التوليد بلا سبب.
- اجعل Rare Court event-driven بدلاً من polling دائم.

- قارن دائماً بـllama.cpp/vLLM/Accelerate لأن الأداة الأفضل تتغير مع الحجم والعتاد.

15. برنامج التجارب وخارطة الطريق

15.1 البرنامج التجريبي ذو الاثني عشر اختباراً

الرمز	التجربة
E1	تشغيل OLMo 2 1B وQwen3-8B وPhi-4-mini عبر runtimes مختلفة، قياس الموارد.
E2	توليد 10k عينة حاكمية من Qwen3-14B عبر AIRLLM.
E3	إعادة تحكيم 20% من العينات بواسطة Phi/DeepSeek؛ قياس الخلاف.
E4	تدريب MOUSSAIF1B-SFT على النسخة الخام مقابل النسخة المحكمة.
E5	إضافة Difference Distillation وقياس robustness.
E6	إنشاء 20k preference pairs وتدريب DPO.
E7	اختبار tool-use على sandbox تنفيذي.
E8	بناء Router 0.6B للتصعيد بين court 1B/8B/AIRLLM.
E9	QLoRA لـ Qwen3-8B كأساس MOUSSAIF9B-architectural.
E10	مقارنة adapter lattice مقابل adapter واحد شامل.
E11	دراسة forgetting بعد العربية والحاكمية والأدوات.
E12	تشغيل محكمة الهدم على 1k حالة خارج التوزيع ثم دورة replay.

15.2 مراحل التنفيذ

المرحلة	المدة البحثية	المخرج الحاسم
0. القياس	أسبوعان	تقرير runtime/VRAM/I/O وترخيص
1. MOUSSAIF1B baseline	4-6 أسابيع	نموذج قابل لإعادة الإنتاج + eval
2. Teacher Forge	3-4 أسابيع	100k عينة محكمة متعددة المعلمين
3. Governance adapters	4 أسابيع	SFT/DPO ablations
4. MOUSSAIF9B pilot	6-10 أسابيع	9B Qwen3-8B derivative أو مسار
5. xB routing	4-6 أسابيع	Router + escalation court
6. نشر بحثي	مستمر	ورقة، كود، data cards، model cards

15.3 بوابات الانتقال

- لا ننتقل من 1B إلى 9B قبل إثبات أن schema الحاكمية يحسن اختباراً خارجياً.
- لا نزيد حجم teacher قبل إثبات أن العائد لكل ساعة أعلى من معلم أصغر.

- لا ندمج adapters قبل اجتياز اختبار عدم النسيان.
- لا نعلن xB قبل وجود router فعلي وقياس تصعيد.
- لا ننشر مشتقاً قبل مراجعة الترخيص وdata provenance.
- لا ندعي AGI؛ نستخدم AGI-oriented أو الذكاء الفتّاح كبرنامج بحثي قابل للتكذيب.

16. المخاطر العلمية والقانونية والأمنية

16.1 المخاطر العلمية

أكبر خطر هو أن يصبح النموذج الطالب نسخة أسلوبية من المعلم. إذا كانت كل البيانات من Qwen، سيتعلم الطالب توزيع Qwen حتى لو سُمّي MOUSSAIF. العلاج هو بيانات أصلية، قضاة متعددون، مهام تنفيذية، واختبارات تقيس الحاكمية لا البلاغة.

الخطر الثاني هو وهم الاستدلال. قد ينتج المعلم trace مقنعاً لكنه غير سببي. لذلك ينبغي تدريب الطالب أحياناً على إجابة قصيرة مع دليل أو execution، واستخدام rationale masking أو اختصار traces، وعدم مساواة طول التفكير بصحته.

الخطر الثالث هو benchmark overfitting. مصنع المعلمين قد يولد أسئلة قريبة من الاختبارات الشائعة أو يعيد محتواها. يجب فصل test، فحص التشابه، وبناء مجموعات خاصة غير منشورة أثناء التدريب.

16.2 المخاطر القانونية

- تراخيص الأوزان: Apache/MIT تختلف عن Llama/Gemma terms.
- بيانات التدريب: حقوق النصوص، الخصوصية، وسياسات إعادة الاستخدام.
- التقطير: بعض التراخيص أو شروط الخدمات قد تقيد استخدام النواتج لتدريب نماذج منافسة؛ يجب فحص كل مصدر.
- إعادة التوزيع: قد يسمح الاستخدام المحلي ولا يسمح نشر أوزان مشتقة أو تضمين علامات تجارية.
- Model/Data Cards: توثيق صريح للمصادر والقيود يحمي النزاهة البحثية.

16.3 المخاطر التشغيلية

- نفاد القرص أثناء splitting أو فساد shard؛ استخدم checksums ومراقبة مساحة.
- تسرب prompt أو بيانات حساسة إلى logs؛ طبق redaction وaccess control.
- تشغيل trust_remote_code بلا مراجعة؛ استخدم بيئة معزولة وثبت commit.
- خلط إصدارات tokenizer/model؛ اربط كل dataset hash بنسخة ثابتة.
- اعتماد قاضٍ آلي في قرار عالي المخاطر؛ أضف human review.

17. الخاتمة العامة

AIRLLM مهم لمشروع مساعف لأنه يغير من يمكنه الوصول إلى النماذج الضخمة في مرحلة الاستدلال. لكنه لا يغير وحده قوانين التدريب. الاستخدام الناضج ليس أن نقول: «شغّلنا 70B على 4GB»، بل: «استعملنا قدرة معلم كبير على عتاد محدود لتوليد وحدات تعلم محكمة رفعت أداء MOUSSAIF1B أو MOUSSAIF9B في اختبارات مستقلة». هذا الانتقال من استعراض الحجم إلى قياس انتقال المعرفة هو جوهر الورقة.

المسار الأكثر واقعية يبدأ بـMOUSSAIF1B فوق قاعدة مفتوحة علمياً مثل OLMo 2 1B، ويستخدم Qwen3 وPhi-4 وDeepSeek وموديلات أخرى كمعلمين متنوعين. بعد إثبات الحاكمية والأدوات والذاكرة، يُبنى MOUSSAIF9B مشتقاً من قاعدة 8-9B مناسبة أو يدرب لاحقاً من الصفر. أما MOUSSAIF-xB فينبغي أن يكون شبكة أحجام وخبراء وتصعيد، لا مجرد رقم أكبر.

تقدم النظريات والأدوات المقترحة تعريفاً عملياً للذكاء الفتّاح: نظام يفتح المسألة عبر السؤال والفروق، يستعير قدرة من معلمين متعددين، يهدم الإجابات الضعيفة، يحول النتيجة إلى تعلم مستقر، ثم يقلل اعتماده على المعلم مع كل دورة. إذا نجح جسر مساعف الهوائي في ذلك، يصبح AIRLLM جزءاً من هندسة تعلم ذات سيادة؛ وإذا لم يحقق عائداً أعلى من البدائل، تسقط الأداة ويبقى المبدأ: كل حوسبة يجب أن تنتج معرفة قابلة للتراكم والتحقق.

17.1 القرار التنفيذي النهائي

القرار – اعتماد AIRLLM كـTeacher/Verifier Runtime اختياري داخل MOUSSAIF ResearcherOS، لا كمحرك التدريب الرئيسي. يبدأ التنفيذ بـOLMo 2 1B + Qwen3-8B/14B teachers، ثم QLoRA/PEFT، ثم DPO ومحكمة هدم، ثم MOUSSAIF9B pilot، وأخيراً xB routing.

الملحق أ: قائمة تحقق تشغيل AIRLLM

- ☐ تثبيت إصدار AIRLLM وcommit أو package version.
- ☐ تسجيل model ID وrevision وtokenizer revision.
- ☐ قراءة الترخيص وتحديد هل يسمح بالمشتق/التقطير/التجارة.
- ☐ توفير مساحة cache + shards + checkpoint + هامش 25%.
- ☐ استخدام NVMe وقياس read throughput.
- ☐ تجربة prompt صغير قبل batch.
- ☐ تسجيل وقت splitting منفصلاً عن وقت inference.
- ☐ قياس tokens/s وtime-to-first-token وpeak VRAM/RAM.
- ☐ مقارنة uncompressed وbit/8-bit-4 على مجموعة حرجة.
- ☐ حفظ provenance لكل output.
- ☐ إعادة تحكيم عينة من النواتج بمعلم آخر أو execution.
- ☐ حساب CVLU قبل توسيع الإنتاج.

الملحق ب: نموذج بطاقة MOUSSAIF checkpoint

- ☐ الاسم الكامل والحجم الفعلي وعدد المعلمات.
- ☐ قاعدة الأوزان وترخيصها.
- ☐ التغييرات المعمارية الحقيقية.
- ☐ بيانات continued pretraining وSFT وDPO ونسبها.
- ☐ المعلمون المستخدمون عبر AIRLLM ونسب مساهمتهم.
- ☐ المحولات/Adapters المضافة والمندمجة.
- ☐ العتاد والزمن والطاقة التقريبية.
- ☐ المقاييس والاختبارات غير المستخدمة في التدريب.
- ☐ القيود واللغات والمخاطر.
- ☐ إرشادات النشر والتكميم والأدوات.

الملحق ج: فرضيات بحثية مستقبلية

الفرضية	الصياغة
H1	يمكن للتقطير التفاضلي أن يحسن robustness أكثر من rationale distillation عند نفس عدد tokens.
H2	معلم 14B متعدد العائلات يحقق CVLU أفضل من 70B واحد عبر AIRLLM في معظم مهام 1B.

H3	Router 1B يقلل كلفة xB بأكثر من نصفها دون هبوط كبير في النجاح.
H4	Adapter lattice يقلل النسيان مقارنة بضبط شامل واحد.
H5	البيانات التي تحمل سبب رفض صريح تحسن calibration أكثر من أزواج تفضيل بلا تفسير.
H6	AIRLLM مفيد أكثر في الاستكشاف والتحكيم من التصنيع الكثيف بسبب I/O.
H7	نقاط التحقق الوسطية في OLMo 2 تسمح بتحديد المرحلة المثلى لإدخال الحاكمية.
H8	MOUSSAIF1B الحاكم قد يرفع أداء منظومة متعددة النماذج أكثر من زيادة حجم المولد وحده.

المراجع

[1] Gavin Li. AIRLLM: Scaling Large Language Models on Low-End Commodity Computers. GitHub software repository, lyogavin/airllm, Apache-2.0, version information consulted July 2026.

[2] Hugging Face. PEFT Documentation: Parameter-Efficient Fine-Tuning.

[3] Hugging Face / bitsandbytes. k-bit quantization, 8-bit optimizers, and QLoRA documentation.

[4] Allen Institute for AI. OLMo-2-0425-1B Model Card: code, checkpoints, logs, training details, Apache-2.0.

[5] Qwen Team. Qwen3-8B and Qwen3 family model cards, Apache-2.0.

[6] Microsoft. Phi-4-mini-instruct Model Card, MIT license, 3.8B parameters, 128K context.

[7] Hugging Face TRL. SFTTrainer and PEFT Integration documentation.

[8] PyTorch. Fully Sharded Data Parallel (FSDP/FSDP2) documentation and tutorial.

[9] Microsoft DeepSpeed. ZeRO-Offload and ZeRO-3 documentation.

[10] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314, 2023.

[11] Hu, E. J., et al. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685, 2021.

[12] Hinton, G., Vinyals, O., & Dean, J. Distilling the Knowledge in a Neural Network. arXiv:1503.02531, 2015.

[13] Rafailov, R., et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290, 2023.

[14] Meta. Llama 3.2 1B Model Card and Llama 3.2 Community License.

[15] Google. Gemma 2 9B Model Card and Gemma Terms.

[16] Mistral AI. Mistral-7B-v0.3 Model Card, Apache-2.0.

[17] DeepSeek AI. DeepSeek-R1 distilled model cards and MIT license statements.

[18] Hugging Face TB. SmolLM2 model cards and intermediate checkpoint release.

بيان الأصالة وحدود الورقة

الأسماء M-AGB وM-MST وM-GDD وM-FSS وM-TSM، وأدوات جسر مساعف الهوائي، هي مقترحات نظرية وهندسية أصلية في هذه الورقة باسم زهير مساعف. أما خصائص AIRLLM والنماذج والمكتبات المشار إليها فمستندة إلى وثائقها ومصادرها، وقد صيغت الادعاءات الرقمية بوصفها تقارير أصحاب المشاريع ما لم تُذكر تجربة مستقلة. الورقة تقدم برنامجاً بحثياً وهندسة قابلة للاختبار، ولا تدعي أن MOUSSAIF قد بلغ AGI أو أن نموذجاً 9B دُرّب فعلياً ما لم يصدر checkpoint وmodel card يثبتان ذلك.