



مقالة رأي وتحليل

النحلة التي أوقفت الرقصة

من أسرار ذكاء الأسراب إلى مختبر «كون» السعودي: لماذا قد تصنع العقول الصغيرة المتخصصة مستقبلاً تعجز عنه العمالقة؟

في عام 2012، نشر عالم الأحياء الأمريكي توماس سيللي (Thomas D. Seeley) وزملاؤه في مجلة *Science* اكتشافاً مدهشاً عن الطريقة التي تتخذ بها أسراب النحل أحد أهم قراراتها: اختيار منزل جديد.

تخرج النحلات الكشافات بحثاً عن تجويف مناسب لإقامة الخلية. وحين تعثر إحداها على موقع واعد، تعود إلى السرب وتؤدي رقصة اهتزازية (Waggle Dance) تنقل من خلالها معلومات عن الموقع. وكلما كان المكان أجود، ازداد نشاط الدعوة إليه، وخرجت نحلات أخرى لتفحصه بنفسها.

ليس في الخلية مهندس معماري يرسم المخطط، ولا لجنة مركزية تفاضل بين العروض، ولا ملكة تُصدر قرار الانتقال.

ومع ذلك، تستطيع آلاف النحلات أن تتوصل إلى اختيار جماعي بالغ التعقيد.

لكن اكتشاف سيللي لم يكن في الرقصة وحدها.

لقد رصد الباحثون سلوكاً أعجب: بعض النحلات لا تكتفي بالدعوة إلى خيارها، بل ترسل إشارات توقف (Stop Signals) إلى النحلات التي تروج لمواقع منافسة.

تُعرف هذه الآلية في علم الأعصاب باسم التثبيط المتبادل (Cross-Inhibition). وهي تساعد الجماعة على تجاوز التردد بين بدائل متقاربة، وتمنع استمرار المنافسة إلى ما لا نهاية.

كأن الطبيعة أدركت أن الذكاء لا يحتاج فقط إلى من يقول «وجدت الحل»، بل إلى من يعرف متى يقول «كفّ عن اتباع هذا الطريق».

Stop Signals Provide Cross Inhibition in Collective Decision-Making by Honeybee Swarms وقد وثّق الباحثون هذه الظاهرة في دراسة بعنوان

هذه الملاحظة البيولوجية الصغيرة تقود إلى سؤال أكبر بكثير من حياة النحل.

كيف تستطيع مخلوقات محدودة القدرات الفردية أن تنتج سلوكًا جماعيًا متقدمًا، دون أن يمتلك أي فرد منها عقلًا يعرف كل شيء؟

وهل يمكن أن يكون هذا السؤال أحد مفاتيح المستقبل القادم للذكاء الاصطناعي؟

اعتدنا خلال السنوات الماضية النظر إلى تطور النماذج اللغوية من خلال أحجامها: مليارات المعاملات (Parameters)، واتساع السياق (Context Window)، والقدرة على الاستدلال (Reasoning)، والطموح إلى الذكاء الاصطناعي العام (Artificial General Intelligence – AGI)، ثم الذكاء الفائق (Artificial Superintelligence – ASI).

سباقٌ بدا أحيانًا كأنه محاولة لصناعة مخلوق خارق يقرأ كل الكتب، ويتقن كل اللغات، ويفهم كل المهن، ويجيب عن كل سؤال.

لكن الطبيعة، التي سبقت مختبراتنا بملايين السنين في تجريب أنظمة الذكاء، تقدم فلسفة أخرى.

ليس شرطًا أن يصبح كل فرد أذكى لكي تصبح الجماعة أذكى.

وقد تكون العبقرية أحيانًا في هندسة العلاقات بين عقول صغيرة، لا في تضخيم عقل واحد إلى ما لا نهاية.

الروبوت الذي لم يحتاج إلى السحابة

في 11 أغسطس 2026، نشرت مؤسسة رازبري باي (Raspberry Pi) مادة تقنية بدت، للوهلة الأولى، أقل إثارة من إعلانات النماذج العملاقة.

عرضت المؤسسة تشغيل نماذج من عائلة (Gemma) التابعة لـ Google، باستخدام بيئة التنفيذ (LiteRT)، على الحاسوب الصغير (Raspberry Pi 5)، بما في ذلك تجربة لروبوت صغير اسمه (Reachy Mini) يستطيع استقبال الإشارات الصوتية والبصرية والاستجابة محليًا، دون الاعتماد على اتصال دائم بمركز حوسبة بعيد.

وفي اختبارات الأداء المنشورة، استطاع نموذج (Gemma 3 270M)، الذي يضم 270 مليون معامل فقط، فك ترميز النص بسرعة نحو 22.6 رمزًا في الثانية باستخدام (LiteRT-LM)، بملف نموذج حجمه 278 ميغابايت، واستهلاك ذاكرة بلغ نحو 680 ميغابايت في الاختبار.

أما تجربة الروبوت متعددة الوسائط نفسها، فاستخدمت نموذجًا آخر من العائلة هو (Gemma 4 E2B)، ضمن منظومة توّرع معالجة الصوت والصورة واللغة على مكونات متخصصة.

التفاصيل منشورة في [التجربة التقنية الرسمية لرازيري باي](#).

لن يتفوق نموذج من 270 مليون معامل على أعظم النماذج العامة في حل المعضلات العلمية، ولن يستطيع الروبوت الصغير منافسة مراكز الحوسبة في جميع الأعمال.

لكن الخبر المهم يسكن في مكان آخر.

لقد أصبحت قدرات كان الوصول إليها قبل سنوات يتطلب خوادم بعيدة، قابلة للتنفيذ على حاسوب يمكن حمله براحة اليد.

والأهم أن النظام لا يطلب من نموذج واحد أن يفعل كل شيء. هناك مكوّن للرؤية، وآخر للتعرف على الكلام، ونموذج لغوي لاتخاذ الاستجابة، ثم نظام لتحويل النص إلى صوت.

أربعة اختصاصات تلتقي لتنتج سلوكًا واحدًا يبدو للمستخدم طبيعيًا.

وفي يونيو 2026، كانت أبل (Apple) قد أعلنت الجيل الثالث من نماذجها التأسيسية (Apple Foundation Models)، متضمنًا نماذج تعمل على الجهاز نفسه (On-Device Models)، من بينها (AFM 3 Core) بحجم يقارب ثلاثة مليارات معامل، إلى جانب نماذج مخصصة لخوادمها الخاصة. لم يكن ذلك استغناءً عن السحابة، بل إعادة توزيع مدروسة للأعمال بين الجهاز والخوادم. [Apple Machine Learning Research](#)

حين أضع الخبرين معًا، لا أرى مجرد نجاح في تصغير النماذج.

أرى بداية تحول في السؤال الهندسي نفسه.

بدل أن نسأل كيف نرسل كل مهمة إلى أكبر نموذج متاح، سنسأل بصورة متزايدة: ما أصغر قدر من الذكاء يكفي لإنجاز هذه المهمة، وأين ينبغي أن يعمل؟

قد تكون الإجابة داخل الهاتف، أو في جهاز صناعي، أو على خادم المؤسسة، أو في نموذج عالمي ضخم حين تستدعي الحاجة ذلك.

وليس الانتصار لأحد هذه المواقع على الآخرين، بل في القدرة على الاختيار بينها بذكاء.

لماذا لا تحتاج مستعمرة النمل إلى مدير تنفيذي؟

قبل ظهور الذكاء الاصطناعي بزمان طويل، كان عالم الحيوان يواجه مشكلة التنسيق بين أعداد هائلة من الكائنات ذات القدرات المحدودة.

كيف تجد مستعمرة نمل طريقها إلى الغذاء؟ وكيف تنظم حركة آلاف العاملات؟ وكيف تتكيف حين ينقطع مسار أو يظهر مصدر جديد؟

في عام 1959، قدم عالم الحشرات الفرنسي بيير بول غراسيه (Pierre-Paul Grassé) مفهومًا أصبح مهمًا في دراسة الأنظمة المعقدة: التنسيق عبر الأثر البيئي (Stigmergy).

الفكرة أن الفرد يستطيع التأثير في سلوك الآخرين دون أن يرسل إليهم تعليمات مباشرة.

تترك النملة أثرًا كيميائيًا في الطريق، فتستجيب له نملات أخرى، وتساهم حركتها بدورها في تغيير قوة المسارات. وتتأثر هذه الآثار بمرور الزمن وتغير ظروف البيئة.

تنشأ أنماط منظمة من عدد كبير من التفاعلات المحلية.

ليس لدى النملة الواحدة خريطة كاملة لكل ما يحدث في المستعمرة، لكن المستعمرة تستطيع إنجاز أعمال تتجاوز بكثير ما يقدر عليه الفرد.

هذا ما يسميه الباحثون الذكاء الجمعي (Collective Intelligence)، أو السلوك الناشئ (Emergent Behavior) حين تظهر خصائص على مستوى الجماعة لا توجد بالصورة نفسها عند أفرادها.

ولم تكن هذه الأفكار استعارات أدبية عابرة. فقد انتقلت إلى علوم الحاسب والروبوتات (Swarm Robotics)، وأسهمت في تطوير أساليب لحل مسائل التحسين (Optimization)، والبحث عن المسارات، وتوزيع المهام.

ومن أشهرها خوارزميات مستعمرات النمل (Ant Colony Optimization)، التي ارتبط تطويرها باسم الباحث ماركو دوريجو (Marco Dorigo).

وأنا لا أزعم أن الشبكات العصبية ينبغي أن تقلد مستعمرات النمل حرفيًا؛ فالاستعارة البيولوجية ليست تصميمًا هندسيًا جاهزًا.

لكن المبدأ العميق يستحق التأمل:

قد تنشأ القدرة الكبرى من وحدات محدودة، إذا كانت طريقة التعاون بينها متفوقة.

وهنا تحديدًا تستوقفني تجربة سعودية أرى أن قراءتها من منظور الأنظمة الحية أكثر خصوبة من وضعها في سياق تقليدي لأحجام النماذج.

إنها تجربة مسراج (Misraj AI) ومعملها البحثي كَوْن (Kawn Lab).

مسراج: حين لا يكون المطلوب عقلًا يعرف كل شيء

ما يثير اهتمامي في توجه مسراج ليس مجرد امتلاكها نموذجًا لغويًا عربيًا صغيرًا.

الأمر الأعمق أنها بدأت تبني عائلة من القدرات المتخصصة، لكل منها وظيفة واضحة، ومجال معرفة، وطريقة تدريب وتقييم تتناسب مع طبيعة المهمة.

في مركز هذه العائلة يوجد نموذج كُوَيْن (Kuwait 1.5B)، وهو نموذج لغوي صغير (Small Language Model – SLM) بحجم 1.5 مليار معامل، استُخدم أساسًا لتطوير نماذج أكثر تخصصًا.

لكن قصة العائلة تصبح أكثر إثارة حين نبتعد عن النموذج العام وننظر إلى فروعها.

هناك مُترجم (Mutarjim)، المتخصص في الترجمة بين العربية والإنجليزية، والمطور فوق كُوَيْن.

وفي مسودته البحثية المنشورة على arXiv عام 2025، عرض الباحثون نتائج تفوق فيها مُترجم على نماذج أكبر منه بكثير في اختبارات محددة، ومنها تفوقه على (GPT-4o mini) في الترجمة من الإنجليزية إلى العربية على معيار (Tarjama-25).

هذا ليس ادعاءً بأنه أفضل من جميع نماذج GPT في كل اللغات والمهام.

لكنه يعرض مفارقة تقنية بالغة الأهمية: نموذج صغير متخصص يستطيع، في نطاق اختباري محدد، تجاوز نموذج تجاري عام يفوقه حجمًا وقدرةً في مجالات كثيرة أخرى.

ثم هناك سديد (Sadeed)، المتخصص في التشكيل العربي (Arabic Diacritization).

وقد يبدو التشكيل مهمة هامشية لمن لا يعرف تعقيد العربية.

لكن كلمة مثل «علم» قد تشير إلى العلم أو العلم أو الفعل عِلِمَ، ولا يستطيع النظام اختيار القراءة الصحيحة من شكل الحروف وحده. عليه أن يفهم تركيب الجملة وسياقها والعلاقات بين كلماتها.

وفي النصوص التراثية والأدبية، قد تغيّر حركة صغيرة وظيفة الكلمة أو معنى العبارة.

لهذا فإن التشكيل ليس تمرينًا زخرفيًا على إضافة الضمات والفتحات، بل مسألة تجمع التحليل الصرفي (Morphological Analysis)، والفهم النحوي (Syntactic Analysis)، والتمييز الدلالي (Semantic Disambiguation).

وفي مسودته البحثية المنشورة على arXiv، استخدم الفريق التدريب المتخصص على بيانات عربية عالية العناية، وقدم معيار التقييم (SadeedDiac-25) لمقارنة أداء التشكيل عبر أنواع مختلفة من النصوص.

ثم تأتي عائلة لهجوي (Lahjawi)، للترجمة بين اللهجات العربية وتحويلها إلى العربية الفصحى المعاصرة (Modern Standard Arabic – MSA).

وهذا اختصاص مختلف عن الترجمة بين الإنجليزية والعربية. فاللهجات لا تتباين في المفردات وحدها؛ إنها تختلف في التراكيب والتعبيرات والإيحاءات الاجتماعية، وأحياناً في المعنى الذي تحمله الكلمة بحسب البلد والسياق.

أما نُبر (Nabr)، فهو نموذج صوتي متخصص في تلاوة القرآن الكريم، والتعرف على القراءة ورصد بعض أخطاء النطق والتجويد (Tajweed-Aware Mispronunciation Detection).

لا يُطلب منه أن يكون مساعداً عاماً يجب عن تاريخ الإمبراطورية الرومانية، ولا أن يحلل ميزانية شركة.

مهمته أن يقترب من مجال صوتي شديد التخصص، تتداخل فيه المدود والغنة والقلقلة والتفخيم وخصائص الأداء.

وفي مجال آخر، نجد بصير (Baseer)، وهو نموذج رؤية ولغة (Vision-Language Model – VLM) بحجم ثلاثة مليارات معامل، متخصص في فهم الوثائق العربية وتحويل محتواها إلى نصوص منظمة، مع مراعاة البنية البصرية للصفحات.

ثم نصل إلى عائلة أرى أنها تحمل واحدة من أهم مفاتيح المستقبل: نماذج التمثيل الدلالي (Embedding Models).

أبرزها (Kawn Embed Light) بحجم 300 مليون معامل، و(Kawn Embed Islamic) و(Kawn Embed Medical)، بحجم 139 مليون معامل لكل منهما، والمتخصصة في استرجاع المعرفة ضمن مجالات مختلفة.

هذه النماذج لا تعمل بوصفها كتاب مقالات أو روبوتات محادثة. إنها تحول النصوص إلى تمثيلات رياضية (Vector Embeddings)، تساعد الحاسوب على العثور على المقاطع الأقرب إلى المعنى المطلوب، حتى حين تختلف الألفاظ المستخدمة.

وفي بحث تصفه صفحة مسراج بأنه قُدّم ضمن أعمال (LaTell 2026)، وصف باحثو مسراج طريقة لتخصص النماذج الدلالية العربية عبر تقطير المعرفة من عدة معلمين (Multi-Teacher Knowledge Distillation).

وهنا تنشأ صورة تكاد تطابق الفكرة التي بدأت بها المقالة.

لدينا عائلة لا تتطلب من كل عضو أن يعرف كل شيء.

نموذج يسمع. وآخر يترجم. وثالث يشكل. ورابع يفهم الصفحة المصورة. وخامس يبحث عن المعنى في ملايين الوثائق.

لم تُوزّع المعرفة هنا لأن الفريق عجز عن صنع نموذج واحد يفعل كل شيء، بل لأن التخصص نفسه قد يكون أفضل قرار هندسي عند تصميم نظام يخدم مهمات محددة.

وهذه، في رأيي، إحدى أكثر الأفكار التي تستحق الانتباه في أعمال مسراج.

ليست القيمة في أن كل نموذج صغير سيهزم جميع العمالقة، وإنما في أن النظام المتكامل يستطيع أن يختار مَنْ يتولى المهمة، ومتى ينتقل العمل إلى غيره.

حين تتحول النماذج إلى فريق عمل

لنتخيل مشهدًا قريبًا من حياتنا.

تجلس أم في منزلها، وأمامها أوراق مدرسية وتقرير يتضمن مصطلحات إنجليزية، وطفل يريد أن يسمع نصًا عربيًا مشكولًا، ورسالة من المدرسة مكتوبة بلغة رسمية تحتاج إلى تفسير.

في نموذج الاستخدام التقليدي، يمكن إرسال كل شيء إلى مساعد لغوي ضخم، ثم طلب ترجمة التقرير، واستخراج المعلومات، وتشكيل النص، وشرح التعليمات.

وقد ينجز المساعد المهمة بكفاءة.

لكننا نستطيع تخيل تصميم آخر، لا يفترض أن كل مشكلة تحتاج إلى استدعاء النموذج نفسه.

إذا كانت الورقة صورة ممسوحة، يتولى نموذج لفهم الوثائق مثل (Baseer) استخراج النص وبنيته.

وإذا احتاج النص إلى ترجمة، تنتقل المهمة إلى (Mutarjim).

وإذا كان المطلوب تشكيل نص عربي للقراءة، يُستدعى (Sadeed).

وإذا كان السؤال عن معنى عبارة وردت في وثائق محددة، يتولى نموذج تمثيل دلالي مثل (Kawn Embed Light) البحث عن المقاطع ذات الصلة، قبل أن يصوغ مكوّن لغوي مناسب الإجابة بالاستناد إليها.

ليس ضروريًا تشغيل كل هذه النماذج في كل مرة. بعض الطلبات تحتاج إلى نموذج واحد، وبعضها إلى سلسلة قصيرة، وبعضها إلى معالجة متوازية.

هذه البنية تسمى، بحسب طريقة تصميمها، توجيه النماذج (Model Routing)، أو سلسلة النماذج (Model Cascading)، وقد تدخل ضمن أنظمة الوكلاء المتعددين (Multi-Agent Systems) إذا أضيفت إليها قدرات اتخاذ القرار واستخدام الأدوات وإدارة العمل.

وهي ليست الشيء نفسه الذي يُعرف معماريًا بمزيج الخبراء (Mixture of Experts – MoE) داخل شبكة عصبية واحدة؛ فذلك تصميم يفعل أجزاء مختارة من النموذج نفسه، بينما نتحدث هنا عن تنسيق بين نماذج مستقلة ذات وظائف مختلفة.

الفارق مهم للمهندس، لكن الفكرة سهلة الفهم.

لا تستدعي طبيب العيون ليصلح جهاز التكييف، ولا تطلب من مترجم محترف أن يحل محل محاسب قانوني، لمجرد أن كليهما متعلم.

فلماذا نصرّ أحياناً على تكليف نموذج لغوي واحد بكل شيء؟

في عام 2023، قدّم باحثون من ستانفورد، بينهم لينغجياو تشين (Lingjiao Chen)، وماتّي زاهاريا (Matei Zaharia)، وجيمس زو (James Zou)، مسودة بحثية على arXiv بعنوان FrugalGPT، درست كيف يمكن توجيه الطلبات عبر نماذج مختلفة وتقليل تكلفة الاستدلال مع المحافظة على جودة الأداء في اختبارات محددة.

ثم جاء مشروع RouteLLM، الذي نُشر في (ICLR 2025)، ليدرس تعلم اختيار النموذج الأنسب للطلب (Learned Model Routing) باستخدام بيانات التفضيلات.

ما يعجبني في هذه الأبحاث أنها تنقل جانباً من الذكاء إلى مستوى أعلى من النموذج نفسه.

فبدل أن نضيف معرفة جديدة إلى كل نموذج، نستطيع تحسين النظام من خلال تعلّم أي نموذج ينبغي أن نستخدم أصلاً.

وفي هذا التحول، قد تصبح براعة من يوزع المهمات بين النماذج مساوية في قيمتها الهندسية لبراعة من يطور أحدها. بل ربما يتبين أن بعض المشكلات التي ظننا أنها تحتاج إلى عقل أعظم، كانت تحتاج في الحقيقة إلى توزيع أفضل للعمل.

التصميم	أين يحدث الاختيار؟	ما الذي يتعاون؟
توجيه النماذج (Model Routing)	بين نماذج مستقلة بحسب المهمة	تخصصات مستقلة
سلسلة النماذج (Model Cascading)	عبر مراحل متتابعة بحسب الحاجة	نماذج تتسلم المهمة تباعاً
مزيج الخبراء (Mixture of Experts)	داخل معمارية نموذج واحد	أجزاء مختارة من الشبكة

مقارنة مفاهيمية مستخلصة من هذا القسم، وليست مقارنة أداء.

أعظم مهارة قد يمتلكها الوكيل: أن يعرف متى يتوقف

هنا أعود إلى نحلة توماس سيللي.

ما يجعل قرار السرب متماسكاً ليس قدرة النحلات على اقتراح مواقع جديدة فقط، بل وجود آليات تمنع استمرار الخيارات الضعيفة وتساعد الجماعة على الوصول إلى قرار.

وهذا يذكرني بنقطة يغفل عنها كثير من المتحمسين للوكلاء الذكيين (AI Agents).

إذا وزعت مهمة على عشرة نماذج ضعيفة، فلن يصبح النظام قويًا بالضرورة لمجرد أن عدد المشاركين ازداد.

قد تنتقل الأخطاء بين النماذج، ويبدأ أحدها بتفسير غير صحيح، ثم يبني الثاني عليه، ويعيد الثالث صياغته بثقة، حتى تظهر في النهاية إجابة شديدة الإقناع، لكنها خاطئة منذ خطواتها الأولى.

ولهذا فإن هندسة التعاون لا تقتصر على اختيار المتخصص.

إنها تحتاج إلى آلية لتقدير عدم اليقين (Uncertainty Estimation)، ومعايرة الثقة (Confidence Calibration)، واكتشاف الحالات الخارجة عن نطاق التدريب (Out-of-Distribution Detection)، والتحقق من المخرجات (Output Verification).

وحين يعجز النموذج الصغير عن تحقيق مستوى الثقة المطلوب، ينبغي أن يعرف النظام كيف يصعد المهمة إلى نموذج أكبر (Model Escalation)، أو إلى مراجعة بشرية (Human-in-the-Loop).

وبالمعنى الهندسي، يجب أن تكون لدى المنظومة قدرة على الامتناع عن الإجابة (Abstention)، لا أن تعدّ الإجابة واجبًا مقدسًا مهما كان مستوى الشك.

وفي المؤسسات الصحية والمالية والقانونية، قد تكون هذه الخاصية أهم من الفصاحة أو سرعة توليد النص.

فالنموذج الذي يعرف أن المستند غير مكتمل، أو أن المعلومة المسترجعة لا تكفي، أكثر نفعًا من نموذج ينتج إجابة بليغة من فراغ.

وأتصور أن الأجيال المقبلة من الوكلاء المتخصصين ستتنافس على جودة هذا القرار بقدر تنافسها على جودة الإجابات.

ليس فقط: هل يستطيع الوكيل إنجاز المهمة؟

بل: هل عرف أي نموذج يستدعي، وماذا يتحقق منه، ومتى يطلب المساعدة، ومتى يتوقف؟

هنا يتحول الذكاء من مجرد القدرة على الكلام إلى القدرة على إدارة المعرفة وحدودها.

أحيانًا يكون أكثر أعضاء الفريق ذكاءً هو الذي يمنع الفريق من ارتكاب الخطأ التالي.

أسرع عقل ليس دائمًا أكبر عقل

حين نتحدث عن كفاءة النماذج الصغيرة، من السهل الوقوع في تبسيط آخر: الاعتقاد أن تقليل عدد المعاملات يجعل النموذج بالضرورة أسرع وأقل استهلاكًا للطاقة.

لكن هندسة الحوسبة أعقد من ذلك.

يتوقف الأداء على معمارية النموذج (Model Architecture)، وطريقة تمثيل الأوزان، وكفاءة التنفيذ (Inference Runtime)، وحجم المدخلات والمخرجات، وسرعة الذاكرة، ونوع المعالج، وكلفة تحميل النماذج وإبقائها في الذاكرة.

فمن خلال التكميم (Quantization)، يمكن تقليل عدد البتات المستخدمة لتمثيل الأوزان، مثل الانتقال من تمثيل بدقة 16 بت (FP16) إلى أربعة بتات (INT4)، مما يخفض حجم التخزين اللازم للأوزان بصورة كبيرة.

ونظريًا، يحتاج نموذج مكوّن من 139 مليون معامل إلى نحو 70 ميغابايت لتخزين أوزانه بأربعة بتات، قبل احتساب بيانات التكميم الإضافية والذاكرة المؤقتة وبيئة التنفيذ.

أما نموذج بحجم 1.5 مليار معامل، فقد يحتاج إلى نحو 750 ميغابايت للأوزان وحدها بالدقة نفسها، ثم تضاف إليه متطلبات الذاكرة الأخرى.

لهذا يمكن التفكير بجدية في تشغيل فئات من هذه النماذج على الهواتف والأجهزة الطرفية (Edge Devices)، لكن ذلك لا يعني أن كل نموذج في كتالوج مسراج قد اختُبر وأُثبت أدائه على هاتف بعينه أو على رازبري باي.

فالقدرّة على تحميل النموذج ليست هي نفسها القدرة على تشغيله بسرعة مناسبة.

والتخصص قد يجعل النموذج أقل حاجة إلى توليد إجابات طويلة، أو إلى معالجة سياقات ضخمة، فتتحسن كلفة المهمة كاملة حتى دون تغيير كبير في حجم النموذج.

وفي دراسة نُشرت ضمن المؤتمر الدولي المشترك للذكاء الاصطناعي (IJCAI 2026) بعنوان *Mapping the Efficiency Landscape of Small Language Models*، درس الباحثون أكثر من سبعين نموذجًا صغيرًا، ووجدوا أن الأصغر ليس بالضرورة الأكثر كفاءة في كل حالة، وأن العلاقة بين الطاقة والأداء أعقد من مجرد ترتيب النماذج بحسب عدد المعاملات.

هذه النتيجة مهمة.

لأن المعيار الأذكى ليس عدد الرموز التي ينتجها النموذج في الثانية وحده (Tokens per Second)، ولا الطاقة اللازمة لتوليد رمز واحد (Joules per Token).

المقياس الأقرب إلى حياة الناس والمؤسسات هو كم يستغرق النظام ويستهلك من موارد حتى ينجز مهمة صحيحة مكتملة؟

قد يكون نموذج صغير بطيء في توليد النصوص الحرة، لكنه بالغ الكفاءة في تصنيف جملة قصيرة. وقد يحتاج نموذج أكبر إلى عدد أقل من الخطوات لإنجاز مشكلة معقدة، فيكون الخيار الاقتصادي الأفضل.

ولهذا فإن العبقرية الهندسية لا تكمن في استخدام أصغر نموذج دائماً.

إنها في استخدام النموذج الذي يحقق النتيجة المطلوبة بأقل كلفة كلية، ضمن حدود الدقة والزمن والخصوصية.

كأننا نعيد اكتشاف مبدأ قديم في الطبيعة: لا ترسل المستعمرة جميع أفرادها إلى زهرة واحدة، ولا تخصص الموارد نفسها لكل مهمة.

اليوم الذي يصبح فيه الهاتف مكاناً للذكاء لا مجرد نافذة إليه

أرى أن أحد التحولات الاجتماعية المهمة خلال السنوات المقبلة لن يحدث على شاشات العروض التقنية الكبرى، بل في الأشياء اليومية التي اعتدنا استخدامها دون تفكير.

الهاتف، والسيارة، والجهاز الطبي المنزلي، والحاسوب المدرسي، وكاميرا المصنع، كلها مرشحة لاكتساب قدرات ذكاء متخصصة تعمل بالقرب من البيانات التي تحتاج إليها.

ليس مطلوباً أن يمتلك الجهاز معرفة العالم كلها.

قد يحتاج هاتف الأم إلى نموذج يتعرف على الكلام العربي محلياً، وآخر يساعد طفلها في قراءة نص، وثالث يرتب مستنداتها، دون أن يرسل كل تفصيل من حياتها إلى خادم بعيد.

وقد يحتاج مصنع إلى نموذج يستخلص معلومات من تقارير الصيانة، وآخر يصنف الإنذارات، وثالث يسترجع تعليمات التشغيل، بينما تبقى بيانات المعدات داخل شبكة المصنع (On-Premises Deployment).

وفي مدرسة، قد يعمل نموذج متخصص لتصحيح القراءة، وآخر لتبسيط النصوص، وثالث لمتابعة تقدم الطالب وفق سياسة واضحة لحماية بياناته.

هذه ليست دعوة لاستبدال المعلم أو الطبيب أو المهندس.

إنها محاولة لجعل الأعمال الصغيرة التي تستهلك وقتهم أقل مشقة، وأن تتوافر بعض القدرات بالقرب منهم حتى عند ضعف الاتصال أو انقطاعه.

ومن أهم النتائج الممكنة لذلك أن تتحول الخصوصية من خدمة إضافية إلى خاصية تصميمية.

فإذا لم تغادر البيانات الجهاز أصلاً، فإن جزءاً مهماً من المخاطر المرتبطة بنقلها وتخزينها خارجها يتقلص، وإن بقيت ضرورة حماية الجهاز نفسه والنماذج المحلية وسجلات استخدامها.

وفي المؤسسات، تفتح قابلية التشغيل المحلي (Local Inference) والنشر داخل الشبكات المغلقة (Air-Gapped Deployment) مجالاً لتطبيقات يصعب تقديمها من خلال خدمات سحابية عامة.

هنا أرى ميزة مهمة لعائلات النماذج المتخصصة العربية.

فبدل الاعتماد على نموذج عام ضخم لمعالجة كل طلب، يمكن تصميم منظومة تعمل فيها نماذج صغيرة داخل حدود المؤسسة، مع إبقاء الأعمال المعقدة على خوادم داخلية أكبر، أو تصعيد ما تسمح به السياسات إلى خدمات خارجية عند الحاجة.

وتعرض مسراج بالفعل منتجات للبناء على نماذجها، مثل (Seamless API) لتكوين مسارات المعالجة والتوجيه، و(Workforces) للوكلاء التشغيليين، مع خيارات للنشر المحلي والمؤسسي.

هذه لبنات معمارية يمكن تطويرها إلى أنظمة متعددة النماذج. أما مقدار التفوق الفعلي لكل تركيب على بدائله، فيعتمد على المهمة والعتاد واختبار التشغيل.

ولعل المكسب الإنساني الأجل أن الذكاء الاصطناعي سيصبح أقل حاجة إلى أن يعرف المستخدم اسمه.

ستختفي بعض المهام داخل الهاتف أو التطبيق؛ تُترجم العبارة حيث تحتاج ترجمة، ويُفهم المستند حيث يحتاج فهمًا، ويُصحح النص حيث يقع الخطأ.

قد تكون علامة نضج الذكاء الاصطناعي أن نتوقف عن الشعور بأننا نستخدم الذكاء الاصطناعي في كل مرة نستفيد منه.

مثل الكهرباء التي لم نعد نذكرها كلما أضأنا غرفة.

حين تصبح العربية ميزة هندسية

ثمة جانب آخر في النماذج الصغيرة المتخصصة يتجاوز مسألة الذاكرة والطاقة.

اللغة العربية، بما فيها من تنوع صرفي ولهجي ودلالي، لا تتوزع صعوباتها بصورة متساوية على جميع المهمات.

فالمشكلة التي يواجهها نظام استرجاع وثائق طبية عربية تختلف عن مشكلة التعرف على أخطاء التلاوة، وتختلف عن الترجمة بين اللهجتين الخليجية والمغاربية، وتختلف عن استخراج الجداول من وثيقة ممسوحة.

والنموذج العام يحتاج إلى تمثيل طيف واسع من هذه المهام داخل قدرة واحدة.

أما التخصص فيسمح بتوجيه البيانات والتدريب والتقييم إلى المشكلة المحددة.

لنأخذ نموذج (Kawn Embed Medical).

في الوثائق الطبية، قد يصف المريض عرضًا بعبارة دارجة، بينما يرد المفهوم المقابل في السجل السريري بمصطلح مهني. وظيفة نموذج التمثيل الدلالي المتخصص أن يقرب بين الصيغ المختلفة للمعنى، بما يساعد على العثور على المعلومة ذات الصلة.

لا يشخص المرض لمجرد أنه وجد النص المناسب، لكنه يؤدي وظيفة بحث دقيقة تحتاج إليها التطبيقات الصحية. وفي (Kawn Embed Islamic)، تكون المشكلة مختلفة: ربط الاستفسارات بالنصوص ذات الصلة في القرآن والحديث والتفسير والفقه، مع المحافظة على أهمية الإسناد إلى المصدر.

قد يتشابه النموذجان في الحجم، لكن بيانات التدريب والفضاء الدلالي ومعايير النجاح تختلف.

وأرى أن هذه المرونة تفتح مجالاً واسعاً لتطوير ما يمكن تسميته **الذكاء المحلي الدقيق (Hyperlocal Specialized AI)**.

نماذج تفهم اللهجات والتعبيرات المهنية والوثائق التي تنتجها مجتمعات بعينها، دون أن تضطر إلى إعادة تعلم معارف العالم كلها.

وقد تتطور هذه النماذج أسرع من النماذج العامة في تخصصاتها؛ لأن تحديث بيانات مهمة ضيقة وتقييم أثره يكون أحياناً أكثر قابلية للإدارة من تعديل منظومة معرفية شاملة.

وهذا لا يعني أن العملاق أصبح بلا فائدة.

النماذج العامة المتقدمة ستظل مهمة في الاستدلال المعقد، والتعامل مع المواقف الجديدة، وربط المجالات المعرفية، والقيام بدور المشرف أو المعالج الاحتياطي للحالات الصعبة.

لكنها قد تصبح، في تصميمات كثيرة، القدرة التي نلجأ إليها عند الضرورة، لا القدرة التي نشغلها تلقائياً لأبسط الأشياء. وهذه نقلة كبيرة في فلسفة التصميم.

اقتصاد لا يدفع ثمن العبقرية إلا عند الحاجة إليها

في تقريرها عن الطاقة والذكاء الاصطناعي، قدرت وكالة الطاقة الدولية – (International Energy Agency) IEA) استهلاك مراكز البيانات عالمياً بنحو 415 تيراواط ساعة في عام 2024، وتوقعت في تحديثاتها اللاحقة وصوله إلى قرابة 950 تيراواط ساعة بحلول 2030.

ليس الذكاء الاصطناعي وحده المسؤول عن استهلاك مراكز البيانات، لكنه أصبح أحد أهم محركات نمو الطلب عليها. ومع توسع استخدام النماذج في كل نشاط، تظهر مسألة جديدة.

إذا كان ملايين البشر سيستخدمون الذكاء الاصطناعي في مهام صغيرة طوال اليوم، فهل من الحكمة الحاسوبية إرسال كل مهمة إلى أكثر النماذج قدرة واستهلاكًا للموارد؟

المسألة تشبه استخدام طائرة شحن لنقل رسالة ورقية.

قد تصل الرسالة بسرعة، وقد تكون الطائرة من أفضل ما صنعتته الهندسة، لكن ذلك لا يجعلها الوسيلة المثلى لكل رسالة.

وإذا أمكن إنجاز نسبة كبيرة من المهام اليومية محليًا، أو عبر نماذج أصغر، فإننا نستطيع تخفيف الضغط على البنية الحاسوبية وإعادة تخصيص الموارد الثقيلة للأعمال التي تستحقها فعليًا.

لكن هذا الاحتمال لا يتحول تلقائيًا إلى وفر بيئي. فقد يؤدي انتشار الذكاء الاصطناعي الرخيص إلى زيادة استخدامه بصورة كبيرة، كما أن تشغيل نماذج متعددة دون تنسيق قد يبذل بعض مكاسب تصغيرها.

لهذا أرى أن ميدان المنافسة القادم لن يكون مجرد تقليل المعاملات، بل تحسين **اقتصاد المهمة (Task-Level Economics)**.

كم جولة من الطاقة استهلك النظام حتى وصل إلى إجابة صحيحة؟ كم ميلي ثانية مرت قبل ظهور النتيجة؟ ما نسبة المهمات التي احتاجت تصعيدًا إلى نموذج أكبر؟ وما أثر الخطأ أو التأخير على المستخدم؟

في هذه المقاييس تكمن إمكانات هندسية واسعة.

قد نكتشف أن أفضل نظام يتكون من نموذج صغير يتولى معظم الطلبات، وآخر أكثر تخصصًا للحالات غير المعتادة، ونموذج كبير لا يعمل إلا عندما تصبح المسألة معقدة.

وستصبح وظيفة الموجّه (Router) شبيهة بدور قائد أوركسترا؛ لا يعزف كل الآلات، لكنه يعرف متى يدخل كل صوت ومتى يصمت.

والقطعة الموسيقية لا تتحسن بإجبار جميع الآلات على العزف بأعلى صوتها.

تتحسن حين يصبح لكل آلة مكانها في الزمن.

ما الذي أتوقعه في نهاية هذا العقد؟

أعتقد أن السنوات القليلة المقبلة ستغير علاقتنا بالنماذج اللغوية على نحو أعمق مما توحى به مقارنات الأداء الحالية.

سيظل سباق النماذج الكبرى قائمًا، وسيظل التقدم نحو قدرات عامة أكثر تعقيدًا أحد أهم ميادين البحث.

لكن على الجانب الآخر، ستتسع صناعة أقل صخبًا وأكثر التصاقًا بالحياة اليومية: صناعة النماذج المتخصصة التي تعمل في الخلفية.

وأرى أمامنا ثلاثة تحولات مترابطة.

الأول أن اختيار النموذج سيصبح جزءًا من الذكاء نفسه. لن يكون النظام المتقدم هو الذي يملك أقوى نموذج فحسب، بل الذي يعرف كيف يوزع المهام، ويتنبأ بدرجة صعوبتها، ويقدر كلفتها، ويختار لها الطريق المناسب.

الثاني أن الأجهزة ستتحول تدريجيًا من أطراف تعرض نتائج الذكاء إلى مواقع تنفذ جزءًا معتبرًا منه. الهاتف لن يكون مجرد شاشة متصلة بعقل بعيد؛ ستكون لديه قدرات محلية، تتكامل مع قدرات أكبر حين يحتاج إليها.

أما الثالث، وهو الأهم في نظري، فسيكون صعود ما أسميه **اللياقة المعرفية (Cognitive Fitness)**.

لن تعني اللياقة هنا أن النموذج يعرف أكثر من الجميع، بل أن النظام يستطيع التكيف بسرعة مع مهمة جديدة، وإضافة تخصص جديد، واستبدال مكون بآخر، وتحسين جزء دون الحاجة إلى إعادة بناء المنظومة كلها.

قد تكون شركة تملك عشرات النماذج المتخصصة، موزعة على قطاعات ولهجات ووثائق وأنواع صوتية، أكثر قدرة على تلبية بعض احتياجات السوق من شركة لا تقدم سوى نموذج عام واحد مهما بلغت قوته.

ليس لأن عدد النماذج أكبر، بل لأن المسافة بين المشكلة والأداة المناسبة لها أقصر.

وهنا أرى ما يستحق الاهتمام في مسراج وكون على المدى البعيد.

فإذا استطاع الفريق تحويل عائلته البحثية إلى منظومة متكاملة من النماذج الصغيرة والمتوسطة، مع توجيه ذكي، وتقييم مستمر، وقدرات تشغيل محلية، فسيملك أسلوبًا هندسيًا قادرًا على التوسع بإضافة التخصصات لا بمجرد تضخيم نموذج واحد.

وقد تصبح المرونة أهم من الحجم في أسواق كثيرة.

فالمؤسسات لا تحتاج دائمًا إلى ذكاء يستطيع مناقشة كل شيء؛ تحتاج إلى نظام يفهم عملياتها، ويتصرف ضمن حدودها، ويحافظ على بياناتها، وينجز الأعمال المطلوبة بسرعة وثقة وتكلفة معقولة.

وفي هذه البيئة، قد يصبح امتلاك مجموعة من المتخصصين المتناغمين ميزة أقوى من امتلاك متحدث موسوعي واحد.

أما على المستوى العلمي، فأرى أن أهم مجال يستحق اهتمام الباحثين سيكون دراسة الحدود الفاصلة بين هذه العقول.

كيف يتبادل نموذجان المعرفة دون أن تنتقل بينهما الأخطاء؟ كيف يعرف الموجه أن المهمة خرجت من تخصص النموذج؟ كيف تُختبر منظومة كاملة بينما يتغير أحد مكوناتها؟ وهل تستطيع النماذج الصغيرة تطوير لغة وسيطة للتعاون (Inter-Agent Communication) تقلل الحاجة إلى إعادة شرح السياق كاملاً عند كل انتقال؟

هذه مسائل بحثية لا تقل إثارة عن زيادة عدد معاملات النموذج.

وقد يكون فيها مجال واسع لاكتشافات سعودية وعربية قادمة.

المستقبل الذي يشبه خلية أكثر مما يشبه عملاقاً

أعود إلى سرب النحل.

لو حاولت أن تشرح قدرته على اختيار منزل جديد من خلال دراسة دماغ نحلة واحدة فقط، فلن تصل إلى القصة كاملة.

عليك أن تفهم الرقصة، واستجابة الكشّافات، وتراكم المعلومات، وإشارات التوقف، والعتبة التي يتحول عندها الاحتمال إلى قرار جماعي.

ذكاء السرب ليس مجرد مجموع أدمغة النحلات.

إنه أيضاً الطريقة التي تنتظم بها علاقاتها.

وهكذا أرى جانباً مهماً من مستقبل الذكاء الاصطناعي.

قد لا يكون السؤال الحاسم دائماً: ما أكبر نموذج نستطيع تشغيله؟

بل: كيف نبني منظومة يعرف كل مكون فيها ما يجيده، وما لا يجيده، ومتى يعمل، ولمن يسلم المهمة من بعده؟

إن تجربة رازبري باي الصغيرة في أغسطس 2026، وأبحاث توجيه النماذج، وعائلة مسراج المتخصصة في اللغة والصوت والوثائق والاسترجاع، كلها إشارات متفرقة إلى اتجاه أوسع.

لم يصبح الذكاء الاصطناعي صغيراً بما يكفي لحل كل شيء على جهاز محدود، ولم تُلغ النماذج المتخصصة الحاجة إلى العمالقة.

لكنها بدأت تغير تصورنا عن المكان الذي ينبغي أن نضع فيه الذكاء، وعن المقدار الذي نحتاج إليه لإنجاز عمل نافع.

ولعل أجمل ما في هذا الاتجاه أنه يفتح المستقبل أمام أشكال أكثر تنوعاً من الابتكار.

فليس مطلوباً من كل جهاز أن يحمل عقلاً خارقاً، ولا من كل نموذج أن يعرف العالم كله، ولا من كل مؤسسة أن تشغل قدرة لا تحتاج إليها.

يكفي أحياناً أن تفهم الآلة مهمتها الصغيرة بعمق، وأن تجد حولها آلات أخرى تعرف مهماتها بالقدر نفسه، وأن تتعاون جميعاً دون أن تنسى حدودها.

وهذا ما جعلني أقرأ أعمال مسراج من زاوية مختلفة: لا بوصفها محاولة لتصغير العملاق، بل بوصفها محاولة لبناء بيئة من العقول المتخصصة، يمكن أن تتكاثر فيها القدرات وتتآزر كما تتآزر الكائنات في منظومة حية.

وربما يأتي يوم لا نعرف فيه اسم النموذج الذي أصلح لنا نصّاً، أو فهم وثيقة، أو ساعد طفلاً على القراءة، أو التقط خلافاً في آلة.

سنلاحظ فقط أن الأشياء أصبحت تعمل بصورة أذكى، وأن خصوصيتنا لم تعد الثمن الضروري لكل مساعدة، وأن الأجهزة الصغيرة تعلمت القيام بما كان يبدو فوق طاقتها.

ذلك اليوم لن يكون إعلاناً عن موت النماذج العملاقة.

سيكون إعلاناً عن نضج الذكاء الاصطناعي، حين يتوقف عن محاولة وضع العالم كله داخل عقل واحد، ويتعلم كيف يوزع المعرفة على عقول كثيرة.

فالنحلة لا تحتاج إلى أن تصبح بحجم الخلية لكي تخدمها.

والذكاء العظيم ليس دائماً أن تعرف كل الإجابات؛ بل أن تعرف أي عقل يملك الإجابة، وكيف تصل إليه، ومتى تكون الحكمة في ألا تسأل أحداً أصلاً.

المراجع والمصادر الأولية

1. [Stop Signals Provide Cross Inhibition in Collective Decision-Making by Honeybee Swarms](#) — Science / AAAS.
2. [Stop signals provide cross inhibition in collective decision-making by honeybee swarms](#) — PubMed record — U.S. National Library of Medicine.
3. [Mastering edge AI on Raspberry Pi with LiteRT and Gemma](#) — Raspberry Pi.
4. [Introducing the third generation of Apple Foundation Models](#) — Apple Machine Learning Research.
5. [La reconstruction du nid et les coordinations interindividuelles chez *Bellicositermes natalensis* et *Cubitermes* sp.; la théorie de la stigmergie](#) — Insectes Sociaux / Springer.
6. [Ant Colony Optimization](#) — About ACO — IRIDIA, Université libre de Bruxelles.
7. [Kuwait 1.5B: An Arabic SLM via Language Injection](#) — arXiv (preprint).
8. [Mutarjim: Advancing Arabic-English Translation with a Small Language Model](#) — arXiv (preprint).
9. [Sadeed: Advancing Arabic Diacritization through Small Language Model Fine-Tuning](#) — arXiv (preprint).
10. [Lahjawi: An Arabic Dialect Translation Suite](#) — Association for Computational Linguistics.
11. [Baseer: A Vision-Language Model for Arabic Document-to-Markdown OCR](#) — arXiv (preprint).
12. [Arabic AI Models: Language, OCR, Speech & Embeddings](#) — Misraj AI / Kawn Lab.

13. [Linguistic Specialization of Arabic in Text Embedding Models via Multi-Teacher Knowledge Distillation](#) — Misraj AI / Kawn Lab.
14. [Nabr — Quranic recitation ASR and Tajweed-aware mispronunciation detection](#) — Misraj AI / Kawn Lab.
15. [Baseer — Arabic document intelligence](#) — Misraj AI / Kawn Lab.
16. [Kawn Embed Light](#) — Misraj AI / Kawn Lab.
17. [Developer Solutions — Seamless API, Kawn Console, and deployment options](#) — Misraj AI.
18. [Workforces](#) — Misraj AI.
19. [FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance](#) — arXiv (preprint).
20. [RouteLLM: Learning to Route LLMs with Preference Data](#) — ICLR 2025.
21. [Mapping the Efficiency Landscape of Small Language Models](#) — IJCAI 2026.
22. [Energy demand from AI — Energy and AI](#) — International Energy Agency.
23. [Executive summary — Key Questions on Energy and AI](#) — International Energy Agency.