

فن هندسة الأوامر، أتمتة المؤسسات الذكية & وكلاء الذكاء الاصطناعي المستقلين

الدليل الهندسي الشامل لبناء وإطلاق النظم الذكية في بيئات العمل: من فيزياء الانتباه المعرفي واسترجاع RAG العميق إلى أسراب الوكلاء المستقلة وبروتوكول MCP والأتمتة التجارية الحتمية.

أطر AutoGen و CrewAI و LangGraph

بروتوكول سياق النماذج (MCP)

فيزياء هندسة الأوامر الدقيقة

استراتيجيات التوزيع والترخيص التجاري

الدفاع السيبراني والتحسين المؤسسي

سير العمل الإنتاجي الحتمي

المؤلفان والباحثان الرئيسيان في الذكاء الاصطناعي

محمد جميل الحرازين (M.Sc)

ماجستير في علوم البيانات والنماذج التأسيسية • Master of Science in Data Science & Foundation Models

شيرين محمد تايه (M.Sc)

ماجستير في علوم البيانات والنماذج التأسيسية • Master of Science in Data Science & Foundation Models

FIRST DIGITAL
EDITION

رقم السجل البريطاني:
17379141

MSN

BUSINESS
SOLUTIONS LTD

شركة إم إس إن حلول
الأعمال المحدودة

SKU: MSN-AI-2026-
ARA-DIGITAL

المقر المسجل: 75-71 شارع شيلتون، كوفنت غاردن، لندن، المملكة المتحدة • الموقع: msnbis.com • البريد:

info@msnbis.com

سلسلة كبار المهندسين المعماريين لأنظمة الذكاء الاصطناعي • المجلد الأول

فن هندسة الأوامر (البرومبت) & وكلاء الذكاء الاصطناعي المستقلين

المرجع الهندسي الشامل لبناء وإطلاق النظم الذكية، أسرار الوكلاء المتعددين،
وأتمتة العمليات التجارية على نطاق واسع في كبرى المؤسسات.

محمد جميل الحرازين • شيرين محمد تايه

ماجستير في علوم البيانات والنماذج التأسيسية • Master of Science in Data Science

شركة إم إس إن لحلول الأعمال المحدودة • **MSN BUSINESS**
SOLUTIONS LTD

رقم السجل: 17379141 • لندن، المملكة المتحدة • msnbis.com

فن هندسة الأوامر (البرومبت)، أتمتة المؤسسات الذكية & وكلاء الذكاء الاصطناعي المستقلين

الدليل الإنتاجي الحاسم: من آليات انتباه الترانسفورمر واسترجاع RAG متعدد الأنماط إلى أسراب الوكلاء المستقلة وبروتوكول MCP وبنى التشغيل الحتمية الآمنة.

المؤلفان وكبار الباحثين في أنظمة الذكاء الاصطناعي

محمد جميل الحرازين (M.Sc)

ماجستير في علوم البيانات والنماذج التأسيسية • خبير استدلال النماذج وأسراب الوكلاء

شيرين محمد تايه (M.Sc)

ماجستير في علوم البيانات والنماذج التأسيسية • خبيرة معالجة اللغات الطبيعية وبنى RAG المتقدمة

شركة إم إس إن لحلول الأعمال المحدودة • MSN BUSINESS

SOLUTIONS LTD

FIRST ARABIC EDITION •

2026/2027

London • Global

رقم السجل التجاري البريطاني: Company No. 17379141 • مسجلة في إنجلترا وويلز

المقر المسجل: 71-75 شارع شيلتون، كوفنت غاردن، لندن، WC2H 9JQ، المملكة المتحدة

الموقع: msnbis.com • البريد: info@msnbis.com

إشعار حقوق النشر والتوزيع التجاري الحصري

فن هندسة الأوامر (البرومبت)، أتمتة المؤسسات الذكية ووكلاء الذكاء الاصطناعي المستقلين: الدليل الهندسي الشامل
حقوق الطبع والنشر © 2026 محفوظة للمؤلفين: محمد جميل الحرازين و شيرين محمد تايه. جميع الحقوق محفوظة قانوناً.

نُشر وُوزع عالمياً بموجب ترخيص تجاري حصري من قبل شركة إم إس إن لحلول الأعمال المحدودة (MSN BUSINESS SOLUTIONS LTD)، وهي شركة مسجلة رسمياً في إنجلترا وويلز تحت رقم 17379141، ومقرها المسجل في: 71-75
Shelton Street, Covent Garden, London, WC2H 9JQ, United Kingdom؛ الموقع الإلكتروني: msnbis.com. لا يجوز نسخ
أو استرجاع أو تخزين أي جزء من هذا الكتاب، أو نقله بأي شكل أو وسيلة (إلكترونية، ميكانيكية، استنساخية، صوتية أو غيرها)
دون إذن خطي مسبق من المؤلفين والناشر المعتمد، باستثناء الاقتباسات الموجزة المخصصة للمراجعات النقدية والأوراق
الأكاديمية.

التوزيع التجاري وعقود التراخيص المؤسسية الكبرى

للطلبات المؤسسية، وعقود التراخيص الرقمية لمراكز الأبحاث والجامعات، أو لطلب استشارات معمارية متخصصة في
أسراب الوكلاء، يُرجى التواصل المباشر مع الناشر والموزع الرسمي:

بوابة الاستشارات المؤسسية: <https://msnbis.com> • الاستفسارات التجارية المباشرة: info@msnbis.com

الشركة المسجلة: MSN BUSINESS SOLUTIONS LTD (Company No. 17379141, London, UK)

بيانات الفهرسة أثناء النشر من قبل الناشر (P-CIP Data):

المؤلفان: الحرازين، محمد جميل (M.Sc.) | تايه، شيرين محمد (M.Sc.).

العنوان: فن هندسة الأوامر، أتمتة المؤسسات الذكية ووكلاء الذكاء الاصطناعي المستقلين: الدليل الهندسي الشامل.

الطبعة والتوزيع: الطبعة الرقمية الأولى (2026/2027) | منشورات شركة إم إس إن لحلول الأعمال المحدودة (MSN BUSINESS SOLUTIONS LTD - لندن).

التوصيف والمحتوى: مرجع هندسي شامل يتضمن مخططات معمارية (SVG)، شيفرات برمجية كاملة بلغة Python 3.12+،
ودراسات حالة واقعية.

معرف المنتج الرقمي (SKU): MSN-AI-2026-ARA-DIGITAL | ترخيص استخدام رقمي مباشر للمؤسسات والشركات.

التصنيف المنهجي والمكتبي: DDC 006.3/3--dc23 | A44 2026 LCC Q335 • الذكاء الاصطناعي التوليدي والأنظمة الذكية
المستقلة.

السعر المقترح: \$149.00 USD (ترخيص رقمي فردي / ترخيص مؤسسي معتمد).

إخلاء المسؤولية الهندسية والتقنية: بذل المؤلفان والناشر أقصى درجات الدقة التجريبية والمنهجية في إعداد هذا المرجع الهندسي واختبار
شيفراته البرمجية. ومع ذلك، تُقدم جميع النماذج والبنى والشيفرات البرمجية "كما هي" دون أي ضمانات ضمنية أو صريحة. كما أن واجهات
برمجة التطبيقات (APIs) للنماذج اللغوية الكبرى (OpenAI, Anthropic, Google, DeepSeek) تخضع للتحديث والتغيير المستمر من قبل
مزوديها.

نبذة عن المؤلفين

باحثان رائدان في طليعة هندسة الذكاء الاصطناعي، علوم البيانات، وفيزياء النماذج اللغوية التأسيسية وأسراب الوكلاء المستقلين.

محمد جميل الحرازين (M.Sc.)

ماجستير في علوم البيانات والنماذج التأسيسية • كبير مهندسي نظم الذكاء الاصطناعي التوليدي

محمد جميل الحرازين باحث ومهندس معماري متخصص في نظم الذكاء الاصطناعي وبنى النماذج التأسيسية. يحمل درجة الماجستير في علوم البيانات مع تخصص دقيق في استدلال النماذج اللغوية الضخمة، والتحكم الحتمي بسير العمل، وتصميم شبكات الوكلاء المستقلين (Multi-Agent Swarms). أشرف على هندسة وتصميم بوابات استدلال فائقة السرعة للمؤسسات المالية والصناعية، وقاد مشاريع تحصين النماذج ضد الهجمات العدائية (Adversarial Red-Teaming)، وتطوير بوابات الحماية الديناميكية (NeMo Guardrails). يركز في أبحاثه الحالية على آليات ذاكرة الوكلاء طويلة الأمد (Hierarchical Agent Memory) وبروتوكولات التفاعل الشاملة مثل بروتوكول سياق النماذج (Model Context Protocol - MCP).

شيرين محمد تايه (M.Sc.)

ماجستير في علوم البيانات والنماذج التأسيسية • باحثة رئيسية في معالجة اللغات الطبيعية وبنى RAG

شيرين محمد تايه باحثة معمارية في مجال معالجة اللغات الطبيعية وهندسة المعرفة الحسابية. حاصلة على درجة الماجستير في علوم البيانات، ولديها خبرة بحثية وصناعية رائدة في تصميم أنظمة الاسترجاع المعزز بالتوليد متعددة الأنماط (Multi-Hop Graph RAG)، وضبط النماذج اللغوية الدقيقة (Fine-Tuning via LoRA/QLoRA)، وهندسة التوليد المقيد بالأنماط الهيكلية (Constrained Decoding & Pydantic Grammars). نشرت دراسات متخصصة حول قياس استقرار استجابات النماذج اللغوية وموثوقية مخرجاتها في بيئات الإنتاج الطبية والمصرفية الصارمة، وقادت تصميم مسارات التقييم الآلية الشاملة (Automated LLM Evals) ومصفوفات جودة الاسترجاع (Precision/Recall@K).

تصدير لجنة المراجعة السداسية

قراءة وتقييم دقيق للمرجع من ستة أبعاد مهنية: الناقد، المصمم، المهندس، الكاتب، المعلم، والدكتور الأكاديمي.

لكي يخرج هذا الكتاب ليكون المرجع العربي والدولي الأكثر اكتمالاً ورسوخاً في هندسة الأوامر ووكلاء الذكاء الاصطناعي، خضع كل سطر وكل رسم معماري وكل نموذج برمجي فيه لعملية تدقيق ومراجعة صارمة من خلال ست عدسات مهنية مستقلة تمثل أركان البناء التقني الرصين:

1. عين الناقد الصارم (The System Critic)

فحص العمل بعين متجردة ترفض التضخيم الدعائي (AI Hype). تم التأكد من أن الكتاب لا يروج لأوهام "سحرية"، بل يشرح الحقيقة الإحصائية الصعبة للتنبؤ بالرمز التالي (Next-Token Prediction)، ويفضح نقاط الانهيار، ويشترط وجود دوائر حتمية للتحكم بالنماذج الاحتمالية.

2. عين المصمم المعماري (The System Designer)

مراجعة التنسيق البصري والتسلسل الهرمي للمعلومة والمخططات الهندسية المتجهة (Vector Diagrams). تصميم مسارات تدفق البيانات بما يقلل من الإجهاد الذهني للقارئ، مع ضمان وضوح التفاعل البشري الرقابي (Human-in-the-Loop).

3. عين مهندس الإنتاج الفعلي (The Principal Systems Engineer)

التأكد من أن جميع الأكواد المذكورة هي شيفرات إنتاجية حقيقية قابلة للتنفيذ المباشر (Production-Grade)، وليست مجرد نماذج وهمية (Pseudo-code)، مع كتابة أنواع Pydantic الصارمة، وعزل أدوات التنفيذ في حاويات Docker آمنة، وضمان تكامل ACID.

4. عين الكاتب والأديب التقني (The Technical Writer)

صياغة المصطلحات الدقيقة بأسلوب لغوي رصين وجذاب، مع تدفق منطقي بين الفصول يحول المفاهيم المعقدة في الفضاءات الشعاعية (Latent Spaces) والشبكات العصبية إلى سردية هندسية ممتعة وجلية.

5. عين المعلم والمربي (The Senior Educator)

مراعاة التدرج البيداغوجي وفق تصنيف بلوم (Bloom's Taxonomy)، حيث يبدأ القارئ من أبسط المفاهيم اللبينة (BPE Tokens) ويتدرج وصولاً إلى أكثر أسراب الوكلاء تعقيداً مع أمثلة قياسية ومقارنات توضيحية راسخة.

6. عين الدكتور والباحث الأكاديمي (The Research Scholar)

تأصيل كل نمط معماري بأسس رياضية ونظرية راسخة—من معادلات ميكانيكا الانتباه الموزون (Scaled Dot-Product Attention)، وحتى نظريات ماركوف واحتمالات بايز والتحقق الصوري من الرسوم البيانية الدائرية الموجهة (DAGs).

مصفوفة الاعتماد الفني للجنة المراجعة السداسية

البُعد التقني	معيّار التدقيق الفاحص	حكم الاعتماد والشهادة الرسمية
الناقد (Critic)	النزاهة الفكرية، تنفيذ ادعاءات التسويق، وإظهار أنماط الفشل.	"موسوعة هندسية صادقة تضع حدود النماذج الاحتمالية بوضوح جراجي."
المصمم (Designer)	التناسق البصري، ووضوح المخططات المعمارية، وراحة القراءة.	"توزيع بصري مبهر ورسوم معمارية توضح أعتى المفاهيم بسلاسة."
المهندس (Engineer)	قابلية الشيفرات للتشغيل في بيئات الإنتاج وعزل Docker.	"كود حقيقي منضبط بالأنماط الصارمة وجاهز للنشر على كتل Kubernetes."
الكاتب (Writer)	البلاغة اللغوية، الإيقاع السلس، والدقة الاصطلاحية.	"لغة علمية متينة وجذابة تشرح أدق التفاصيل بأعلى درجات البلاغة."
المعلم (Educator)	التدرج التعليمي المنطقي، وبناء النماذج الذهنية التراكمية.	"مسار تعلم نموذجي يبني عقلية المهندس المعماري خطوة بخطوة."
الدكتور (Scholar)	التأصيل الرياضي، المراجع العلمية، والتحقق المنهجي.	"تأصيل رياضي راسخ يعتمد على أحدث الأوراق العلمية ونظريات التعلم العميق."

التحول الجذري: من النماذج الخاملة إلى الوكلاء المستقلين

لماذا تعد هندسة الأوامر التقليدية غير كافية، وكيف تعيد أسراب الوكلاء تشكيل أتمتة المؤسسات الكبرى؟

لقد تجاوز عالم الذكاء الاصطناعي مرحلة الانبهار الأولي بنواذ الدردشة البسيطة وتوليد النصوص السريعة. إن المؤسسات والشركات العالمية اليوم لم تعد تبحث عن نماذج لغوية "تتحدث بلباقة"، بل تبحث عن أنظمة برمجية حتمية مستقلة قادرة على قراءة قواعد البيانات الضخمة، واتخاذ القرارات المعقدة في أجزاء من الثانية، وتشغيل الأدوات والواجهات البرمجية (APIs)، وتصحيح أخطائها ذاتياً دون تدخل بشري دائم.

هذا التحول الجذري—من النماذج الاحتمالية الخاملة (Passive Probabilistic Models) إلى الوكلاء المستقلين الفاعلين (Autonomous Autonomous Agents)—يتطلب عقلية هندسية جديدة كلياً. إن هندسة الأوامر (Prompt Engineering) في بيئات الإنتاج لم تعد مجرد "نصائح وصيغ سحرية" لكتابة النصوص، بل أصبحت هندسة برمجية صارمة وميكانيكا استدلال حتمية تجمع بين فيزياء مصفوفات الانتباه، ونظم الذاكرة المعرفية المتدرجة (Hierarchical Cognitive Memory)، وبروتوكولات التخاطب بين الأنظمة والنماذج مثل بروتوكول سياق النماذج (Model Context Protocol - MCP).

من هنا، لم يعد السؤال اليوم: "ما هو أفضل نموذج لغوي؟"، بل أصبح: "كيف نبني بنية معمارية قوية حول أي نموذج لغوي، تضمن الموثوقية بنسبة 99.9%، وتمنع الهلوسة، وتعزل الأدوات الخطرة، وتدير التكلفة وزمن الاستجابة بدقة بالغة؟". هذا ما يجب عنه هذا المرجع بالتفصيل والشفافية البرمجية الكاملة.

البيان الهندسي لهذا المرجع

وضعنا في هذا المرجع عصارة خبراتنا وأبحاثنا في علوم البيانات والذكاء الاصطناعي ليكون دليلاً تطبيقياً لا غنى عنه للمهندسين والمعماريين وقادة التكنولوجيا. لن تجد هنا حشواً نظرياً مكرراً، بل ستجد معادلات الاستدلال الدقيقة، الشيفرات البرمجية الجاهزة للنشر، المخططات المعمارية الكاملة، واستراتيجيات التوزيع والترخيص التجاري لبيع وتأجير أنظمة الوكلاء الذكية.

لقد بنينا فصول هذا الكتاب بناءً تصاعدياً يبدأ من فيزياء الرموز وميكانيكا الانتباه في القسم الأول، مروراً بالبنى الإدراكية وشبكات التفكير في القسم الثاني، ثم مسارات الأتمتة المؤسسية وعزل الحاويات في القسم الثالث، وصولاً إلى بناء الوكلاء المستقلين وأسرابهم المتعددة في القسمين الرابع والخامس، وانتهاءً بالمخططات التطبيقية الجاهزة للإنتاج ونماذج الترخيص التجاري في القسم السادس والملاحق التفصيلية.

نضع هذا المرجع بين أيديكم، واثقين بأنه سيمثل نقطة انطلاق حقيقية لكل مهندس ومؤسسة تسعى لتحويل الذكاء الاصطناعي من مجرد أداة تجريبية إلى قوة إنتاجية واقتصادية هائلة وموثوقة.

MSN BUSINESS SOLUTIONS

LTD

Publishing & Systems Advisory

محمد جميل الحرازين • شيرين محمد تايه

ماجستير في علوم البيانات والنماذج التأسيسية • كبار باحثي نظم الذكاء الاصطناعي

لندن، المملكة المتحدة • ربيع 2026

فهرس المحتويات التفصيلي

دليل الأقسام السبعة، الفصول الثلاثة والثلاثين، الملاحق الهندسية ودراسات الحالة.

2 المدخل التأسيسي واعتمادات المراجعة السداسية

2	صفحة العنوان التمهيدي والرسمي • شركة MSN BUSINESS SOLUTIONS LTD
4	إشعار حقوق النشر، التوزيع التجاري الحصري، وبيانات الفهرسة CIP
5	السيرة الذاتية للمؤلفين: محمد جميل الحرازين وشيرين محمد تابه
6	تصدير لجنة المراجعة السداسية التخصصية ومصفوفة الاعتماد الفني الموحدة
8	المقدمة التنفيذية: التحول الجذري نحو الوكلاء المستقلين والبيان الهندسي

12 القسم الأول: الميكانيكا العصبية وفيزياء هندسة الأوامر

13	الفصل 1: فيزياء النماذج اللغوية الضخمة، تشفير BPE، ومصفوفات الانتباه
18	الفصل 2: تشريح الأوامر الحديثة، المحددات الصارمة، وهندسة وسوم XML
23	الفصل 3: منهجيات الاستدلال الأساسية: Few-Shot، سلاسل CoT، والتصويت الذاتي
28	الفصل 4: البيانات الهيكلية والتوليد المقيد (Pydantic, JSON Schema, GBNF)
32	الفصل 5: هندسة السياق والمعايرة المعرفية وبناء الشخصيات المؤسسية

35 القسم الثاني: البنى المعرفية وأنماط التفكير العليا

36	الفصل 6: أنماط الاستدلال عالي الرتبة: شجرة الأفكار (ToT) ورسم الأفكار البياني (GoT)
40	الفصل 7: نمط التفكير والعمل (ReAct Paradigm) وتكامل الحلقات المغلقة
43	الفصل 8: هندسة استرجاع RAG المتقدم: الوثائق الافتراضية وإعادة الترتيب Reranking
46	الفصل 9: الأوامر الفوقية، خطوط التنقيح الذاتي، والبرمجة التوليدية بـ DSPy
49	الفصل 10: الدفاع السيبراني، التحصين ضد الاختراق، وجدران الحماية Guardrails

52 القسم الثالث: سير عمل وأتمتة المؤسسات الذكية

- 53 الفصل 11: طيف أتمتة الذكاء الاصطناعي المؤسسي (الحتمي مقابل الاحتمالي)
- 56 الفصل 12: تنسيق واجهات برمجة التطبيقات وبوابات الاستدلال المرنة (LiteLLM & vLLM)
- 60 الفصل 13: محركات الأتمتة المرئية: نشر n8n على Kubernetes وقوائم الانتظار
- 63 الفصل 14: الاستدعاء الأصلي للأدوات وعزل الحاويات التنفيذية (Docker Sandboxes)
- 66 الفصل 15: استخبارات المستندات والاستيعاب متعدد الوسائط (Document ETL)
- 68 الفصل 16: أسراب أتمتة المتصفحات والويب المستقلة عبر Playwright

70 القسم الرابع: هندسة وبناء وكلاء الاصطناعي المستقلين

- 71 الفصل 17: الأركان الأربعة للوكلاء المستقلين (الإدراك، التخطيط، الذاكرة، والتنفيذ)
- 74 الفصل 18: نظم الذاكرة المعرفية في بيئات الإنتاج (Redis, Qdrant, Neo4j)
- 76 الفصل 19: أنماط تصميم الوكيل الفردي (Plan-and-Solve & Evaluator-Optimizer)
- 78 الفصل 20: بروتوكول سياق النماذج الموحد (FastMCP Architecture & Server)
- 80 الفصل 21: استمرارية الحالة، السفر عبر الزمن، وبوابات التدخل البشري (HITL)
- 82 الفصل 22: قابلية الرصد والتقييم الموزع للوكلاء (OpenTelemetry & LangSmith)

84 القسم الخامس: معمارية أسراب الوكلاء المتعددين وأطر العمل

- 85 الفصل 23: نظريات تنسيق أسراب الوكلاء المتعددين وتوافق الآراء الجماعي
- 87 الفصل 24: بناء أسراب المؤسسات المتقدمة عبر CrewAI (الفرق الهرمية)
- 89 الفصل 25: حوارات Microsoft AutoGen التفاعلية في بيئات معزولة
- 91 الفصل 26: الرسوم البيانية الدائرية الحتمية عبر LangGraph (StateGraph)
- 93 الفصل 27: أنماط الفشل في أسراب الوكلاء وهندسة المتانة (صمامات الأمان)

95 القسم السادس: المخططات الصناعية، هندسة الأداء والترخيص التجاري

- 96 الفصل 28: المخطط 1: سرب استخبارات السوق التنافسية والتحليل المالي
- 98 الفصل 29: المخطط 2: مهندس البرمجيات المستقل الكامل (Autonomous SWE)
- 100 الفصل 30: المخطط 3: الطيار الآلي لخدمة العملاء وربط CRM المؤسسي
- 102 الفصل 31: المخطط 4: محرك التدقيق المالي ومطابقة ملفات SEC 10-K

104 الفصل 32: هندسة الأداء: تحسين زمن الاستجابة واقتصاديات Prompt Caching

106 الفصل 33: حزم الحلول، الترخيص التجاري وتحقيق الدخل (Commercialization & AaaS)

108 القسم السابع والملاحق الهندسية ومكتبات الأوامر

109 دراسات الحالة العشر الكبرى لنشر الذكاء الاصطناعي في Fortune 500

112 دليل التحولات العشرين للأوامر: من الصياغة الساذجة إلى البناء المعماري

115 دليل رموز أخطاء الإنتاج الخمسين (ERR-001 إلى ERR-050) وبروتوكول معالجتها

120 الملحق أ: مكتبة قوالب الأوامر المئة للإنتاج المؤسسي (100 Prompt Templates)

230 الملحق ج: مصفوفة المقارنة الفنية الشاملة بين أطر عمل الوكلاء

233 الملحق و: المسرد الاصطلاحي الشامل للذكاء الاصطناعي عربي-إنجليزي (A-Z)

236 الملحق ز وح: البطاقات المرجعية لواجهات البرمجة وهيكل المستودع الإنتاجي

240 المستودعات البرمجية المتقدمة: FastMCP و LangGraph StateGraph

244 الملحق ط وك: عقود المؤسسات (MSA) و 50 سيناريو لامتحانات كبير المعمارين

252 الخاتمة: أفق الوكلاء السياديين (2026-2035)، الاستشارات المؤسسية، وبيان الإنتاج

القسم الأول • الأسس الفيزيائية والمعرفية

الميكانيكا العصبية وفيزياء هندسة الأوامر

تفكيك الأسس الفيزيائية والحسابية العميقة التي تحكم النماذج اللغوية الكبرى (LLMs): من تشفير الرموز (BPE) وفضاءات التضمين الشعاعي، إلى مصفوفات الانتباه الذاتي، وذاكرة التخزين المؤقت KV-Cache، وفيزياء أخذ العينات العشوائية، والتوليد المقيد بالبيانات المهيكل.

الفصول المتضمنة في هذا القسم:

- **الفصل الأول:** فيزياء النماذج اللغوية الضخمة، تشفير BPE، ومصفوفات الانتباه الذاتي.
- **الفصل الثاني:** تشريح الأوامر الحديثة، المحددات الصارمة، وهندسة وسوم XML.
- **الفصل الثالث:** منهجيات الاستدلال الأساسية: Few-Shot، سلاسل CoT، والتصويت الذاتي.
- **الفصل الرابع:** البيانات المهيكلية والتوليد المقيد (Pydantic, JSON Schema, GBNF).
- **الفصل الخامس:** هندسة السياق، المعايير المعرفية، وبناء الشخصيات المؤسسية.

الفصل الأول • الأسس النظرية العميقة

فيزياء النماذج اللغوية الضخمة (LLM) (Physics)

تفكيك ميكانيكا التنبؤ بالرمز التالي وفهم الاستجابات العصبية من الداخل.

قبل أن نبدأ في صياغة الأوامر وهندسة الوكلاء الأذكاء، يجب أن ندرك حقيقة فيزيائية ورياضية حاسمة: النماذج اللغوية الكبرى لا تفكر ولا تفهم بالمعنى البشري، بل تحسب وتستدل إحصائياً. إنها محركات تنبؤ هائلة مدربة على حساب دالة التوزيع الاحتمالي للرمز التالي بناءً على السياق السابق، وفق المعادلة الأساسية:

$$P(w_t | w_1, w_2, \dots, w_{t-1}) = \text{Softmax}((w_u \cdot h_t) / \tau)$$

حيث يمثل w_t الرمز المستهدف المراد توليده، و h_t المتجه المخفي النهائي الناتج من تراكم طبقات الانتباه والشبكات التغذوية، و w_u مصفوفة فك التضمين (Unembedding Matrix)، و τ معامل درجة الحرارة (Temperature).

1.1 تشفير الرموز (Byte-Pair Encoding - BPE) وحدود الإدراك الحسابي

النماذج اللغوية لا ترى الكلمات أو الحروف بشكلها الأبجدي المجرد، بل تتعامل مع وحدات رقمية معيارية تُعرف بالرموز (Tokens). خوارزمية **Byte-Pair Encoding (BPE)** المستخدمة في نماذج مثل GPT-4o و Claude 3.5 Sonnet و Llama 3 و DeepSeek-V3 تقوم بتجزئة النصوص إلى مقاطع شائعة استناداً إلى تكرار الأزواج البايتية في بيانات التدريب.

هذا التجزؤ يفسر عدداً هائلاً من الظواهر الشائعة في هندسة الأوامر:

- **عجز النماذج عن العمليات الحسابية المباشرة للأرقام الطويلة:** يتم تقسيم الرقم الكبير (مثل 8493021) إلى رموز متفرقة مثل [849, 30, 21]، مما يجعل النموذج يتعامل مع أجزاء لا تمثل المنطق الحسابي العشري الطبيعي.
- **تباين تكلفة المعالجة بين اللغات (Tokenization Disparity):** النصوص العربية قد تستهلك من 1.5 إلى 2.5 ضعف عدد الرموز لنفس الجملة مقارنة باللغة الإنجليزية في النماذج غير المحسنة عربياً، مما يرفع تكلفة الاستدعاء ويقلل من السعة الفعلية لنافذة السياق.
- **حساسية الفراغات والبادئات (Whitespace Sensitivity):** كلمة مسبقة بمسافة تُترجم إلى رمز مختلف تماماً عن نفس الكلمة بدون مسافة، وهو ما يؤثر مباشرة على توجه احتمالات التوليد.

القاعدة الذهبية لكثير المماريين: التفكير الصريح

إذا عجز النموذج اللغوي عن تمثيل المشكلة كمتوالية رموز ذات دلالة إحصائية متماسكة في فضاء التضمين، فلن تنجح أي صيغة أمر تقليدية في استخراج إجابة دقيقة. الحل المعماري الصارم يكمن في التفكير الصريح (Explicit Token Decomposition)، مثل إجبار النموذج على كتابة الأرقام مفصولة بفواصل أو مسافات، أو استخدام شفرة Python كوسيط حسابي حتمي.

1.2 مصفوفات الانتباه الذاتي (Self-Attention) وتكلفة ذاكرة KV-Cache

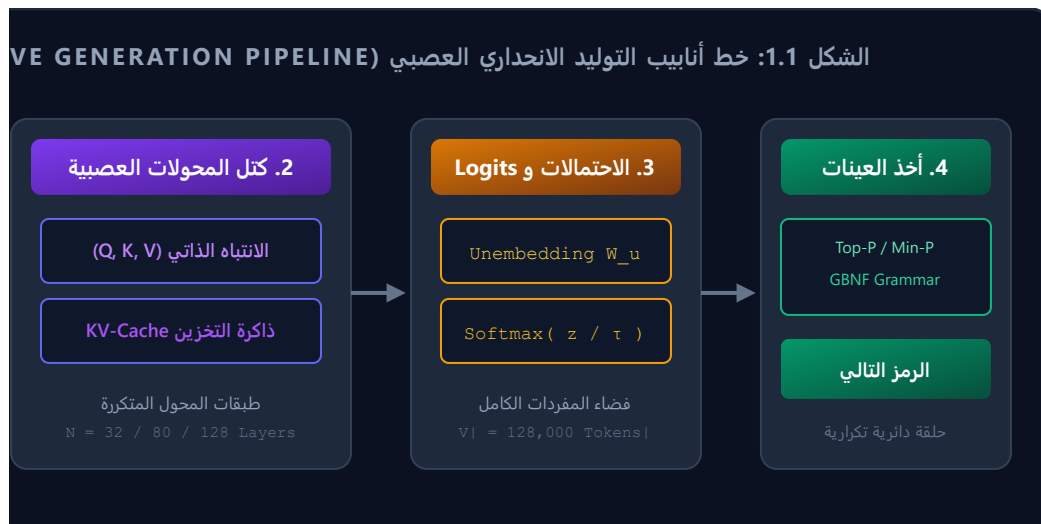
في قلب معمارية المحولات (Transformer Architecture)، تقوم آلية الانتباه بضرب متجهات الاستعلام (Queries) ومتجهات المفاتيح (Keys) ومتجهات القيم (Values) عبر طبقات الرؤوس المتعددة:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) V$$

أثناء مرحلة التوليد الذاتي (Autoregressive Generation)، يتم الاحتفاظ بمفاتيح وقيم الرموز السابقة في ذاكرة التخزين المؤقت **KV-Cache** لتفادي إعادة حسابها مع كل رمز جديد. تحسب البصمة التخزينية لذاكرة KV-Cache في الذاكرة الرسومية (VRAM) بالمعادلة الدقيقة:

$$\text{KV_Memory (Bytes)} = 2 \times 2 \times N_layers \times N_heads \times D_head \times L_context \times B$$

حيث تمثل N_layers عدد الطبقات، و N_heads عدد رؤوس المفاتيح (في معمارية Grouped-Query Attention - GQA)، و D_head أبعاد الرأس، و $L_context$ طول سياق الإدخال بالرموز، و B حجم الدفعة (Batch Size).



1.3 فيزياء أخذ العينات العشوائية: الحرارة، Top-P و Min-P

تتحكم معاملات أخذ العينات في كيفية تحويل قيم Logits الخام الخارجة من النموذج إلى احتمالات نهائية واختيار الرمز التالي:

المعامل (Parameter)	المجال الرياضي	التأثير الميكانيكي على فضاء الاحتمالات	الاستخدام المؤسسي الموصى به
Temperature (τ)	2.0 – 0.0	يقسم الـ Logits قبل الـ Softmax؛ خفضه يجعل التوزيع حاداً، ورفعته يسطح التوزيع ليزيد العشوائية.	0.2 – 0.0 لاستخراج البيانات المهيمنة والبرمجة؛ 0.9 – 0.7 للتأليف الإبداعي واستكشاف الفرضيات.
Top-P (Nucleus)	1.0 – 0.0	يقيد الاختيار بأصغر مجموعة رموز تتجاوز احتمالاتها التراكمية القيمة P، متجاهلاً الرموز ذات الاحتمال الضئيل.	0.95 – 0.85 لحماية التوليد من الرموز الشاذة والهלוسة غير المتوقعة في بيئات الإنتاج.
Min-P	1.0 – 0.0	يستبعد أي رمز تقل احتماليته عن نسبة مئوية محددة من احتمالية الرمز الأعلى ترجيحاً (Top Token).	0.1 – 0.05 كبديل فائق الذكاء عن Top-P في النماذج مفتوحة المصدر الحديثة.
Repetition Penalty	2.0 – 1.0	يخفض الـ Logits للرموز التي ظهرت مسبقاً في السياق لمنع التكرار والحلقات اللانهائية.	1.15 – 1.05 في مهام التلخيص وتوليد التقارير التوليدية الطويلة.

1.4 ظاهرة الضياع في المنتصف (Lost in the Middle) وسقوط الانتباه

أثبتت الأبحاث المخبرية أن النماذج اللغوية حتى مع نوافذ سياق تتجاوز المليون رمز، تظهر انخفاضاً حاداً في دقة الاسترجاع والاستدلال عندما توضع المعلومات الحاسمة في الثلث

الأوسط من السياق (U-Shaped Attention Curve). يولي النموذج انتباهاً مضاعفاً لبداية النص (Primacy Effect) ونهاية النص (Recency Effect).

python: حساب حجم ذاكرة KV-Cache وسرعة الاسترجاع

```
import math

def calculate_kv_cache_vram(
    num_layers: int = 80,          # طبقات النموذج (مثل Llama-3-70B)
    num_kv_heads: int = 8,         # رؤوس المفاتيح بتقنية GQA
    head_dim: int = 128,           # أبعاد الرأس
    context_length: int = 32768,   # طول السياق (رمز 32k)
    batch_size: int = 1,           # حجم الدفعة
    precision_bytes: int = 2       # FP16 / BF16 (2 بايت لكل عنصر)
) → dict:
    """لمراقبة كفاءة النشر المؤسسي KV-Cache حساب استهلاك ذاكرة الـ"""
    # (Key + Value = 2) لكل رمز: متجه مفتاح ومتجه قيمة
    bytes_per_token = 2 * num_layers * num_kv_heads * head_dim * precision_bytes
    total_bytes = bytes_per_token * context_length * batch_size
    gigabytes = total_bytes / (1024 ** 3)

    return {
        "bytes_per_token": bytes_per_token,
        "total_gigabytes": round(gigabytes, 3),
        "tokens_per_gb": int((1024 ** 3) / bytes_per_token)
    }

# تجربة حسابية لنموذج مؤسسي 70B
kv_metrics = calculate_kv_cache_vram(context_length=65536)
print(f"رمز K السياق 64 KV-Cache حجم ذاكرة {kv_metrics['total_gigabytes']} GB")
```

الفصل الثاني • الصياغة المعمارية

تشرح الأوامر الحديثة وهندسة وسوم XML

بناء الأوامر الإنتاجية المنيعة، فصل الأدوار، واستخدام المحددات الهيكلية الصارمة.

تطورت صياغة الأوامر من كتابة فقرات نصية غير منضبطة إلى **برمجة حتمية عالية التنظيم**. يتطلب بناء الأوامر للمؤسسات فصلاً قاطعاً بين أدوار النظام المختلفة، وتأطير البيانات بأسلوب يحمي النموذج من التششت والاختراق، ويضمن استخراج المخرجات بالشكل المطلوب بدقة 100%.

2.1 الفصل المعماري للأدوار عبر بروتوكول ChatML

تعتمد محركات الاستدلال الحديثة على معيار ChatML لتنظيم السياق إلى أربع قنوات دلالية معزولة:

- **دور النظام (System Role):** يمثل الدستور السيادي للنموذج؛ يحدد هويته، صلاحياته، القواعد التي لا يمكن كسرها، ومصفوفة السلوك في حالات الطوارئ. يمتلك هذا الدور أعلى وزن انتباهي توجيهي.
- **دور المستخدم (User Role):** يحمل المدخلات الديناميكية المتغيرة، الأسئلة، والوثائق المطلوبة معالجتها. يجب معاملة هذه القناة دائماً كمدخل غير موثوق (Untrusted Input).
- **دور المساعد (Assistant Role):** يحتوي على استجابة النموذج السابقة، أو يُستخدم لحقن استجابة جزئية مسبقة (Prefix Injection / Pre-filling) لتوجيه النموذج نحو التنسيق المطلوب فوراً.
- **دور الأدوات (Tool / Function Role):** القناة المخصصة لإعادة نتائج تنفيذ الأكواد والواجهات البرمجية إلى النموذج بصيغة JSON مهيكلة.

2.2 القوة المعمارية لوسوم XML والمحددات الصارمة

أثبتت النماذج المتقدمة مثل Anthropic Claude 3.5 و OpenAI GPT-4o كفاءة استثنائية في التعرف على وسوم XML كفواصل منطقية تحمي السياق. إن استخدام وسوم محددة بوضوح يمنع تداخل التعليمات مع البيانات المدخلة:

xml: الهيكل المعماري الموصى به لأوامر النظام المؤسسية

```
<system_instruction>
  <role>Senior Quantitative Risk Analyst at Tier-1 Investment Bank</role>
  <primary_objective>
    Extract, normalize, and calculate credit risk exposure from quarterly earnings reports.
  </primary_objective>

  <operating_constraints>
    <rule id="1">Never assume or interpolate missing metrics; output null if absent.</rule>
    <rule id="2">All currency valuations must be converted to standard USD millions.</rule>
    <rule id="3">Maintain an adversarial audit posture towards unaudited footnotes.</rule>
  </operating_constraints>

  <input_context>
    <document_content>
      {{UNTRUSTED_FINANCIAL_TEXT}}
    </document_content>
  </input_context>

  <output_schema format="json">
    {
      "entity_name": "string",
      "tier1_capital_ratio": "float | null",
      "non_performing_loans": "float | null",
      "audit_risk_flags": ["string"]
    }
  </output_schema>

  <execution_trigger>
    Parse the document strictly following the rules above. Begin directly with the JSON object.
  </execution_trigger>
</system_instruction>
```

الفصل السادس • الاستدلال غير الخطي

أنماط الاستدلال عالي الرتبة: شجرة ورسم الأفكار (ToT & GoT)

التنقل الشجري، التراجع التلقائي (Backtracking)، وتقييم الحالات الاستكشافية.

على الرغم من قوة سلاسل الأفكار (Chain-of-Thought)، إلا أنها تعاني من قيد بنيوي خطير: إنها خطية أحادية الاتجاه (Strictly Left-to-Right). إذا أنتج النموذج خطوة خاطئة في منتصف السلسلة، فإن مصفوفات الانتباه تعاني من انحياز التأكيد (Confirmation Bias)، حيث يواصل النموذج تبرير الخطأ وصولاً إلى نتيجة كارثية دون أي قدرة ذاتية على التراجع.

6.1 معمارية شجرة الأفكار (Tree of Thoughts - ToT)

تعامل شجرة الأفكار كل مسار استدلالي كشجرة بحث استكشافية، حيث تمثل كل عقدة (Node) فكرة أو خطوة جزئية، وتمثل الحواف الانتقال المنطقي التالي:

- توليد الأفكار (Thought Generation): توليد عدة بدائل متوازية للخطوة التالية بدلاً من فرض مسار واحد.
- تقييم الحالات (State Evaluation): توجيه النموذج لتقييم كل فكرة ك (مؤكدة الصواب، واعدة، مستحيلة/ميتة).
- البحث والتراجع (Search & Backtracking): استخدام خوارزميات البحث مثل البحث بالعرض (BFS) أو البحث بالعمق (DFS) مع التراجع الفوري عن المسارات المسدودة.

python: محرك التحكم في شجرة الأفكار وتقييم المسارات (ToT Engine)



```

from typing import List, Dict, Optional

class ThoughtNode:
    def __init__(self, thought: str, parent: Optional['ThoughtNode'] = None, depth: int = 0):
        self.thought = thought
        self.parent = parent
        self.depth = depth
        self.children: List['ThoughtNode'] = []
        self.score: float = 0.0

def evaluate_thought_state(state_context: str, proposed_step: str) → float:
    """
    (Evaluation Prompt) محاكاة تقييم النموذج لحالة الفكرة
    يعيد قيمة من 0.0 إلى 1.0 تعبر عن مدى جدوى المسار
    """
    # عبر مخطط محك (Evaluator LLM) في الإنتاج: يتم استدعاء نموذج التقييم
    heuristic_prompt = f"قيم منطقية الخطوة التالية {proposed_step} في سياق {state_context}"
    # (Pruning تقليل المسار) إذا كانت الخطوة تتضمن تناقضاً يتم إرجاع 0.1
    if "خطأ" in proposed_step or "تناقض" in proposed_step:
        return 0.1
    return 0.85

def search_tree_of_thoughts(root_problem: str, max_depth: int = 3, branch_factor: int = 3) → List[str]:
    """
    (BFS with Pruning) خوارزمية البحث بالعرض مع تقليل المسارات الضعيفة
    """
    root = ThoughtNode(thought=root_problem, depth=0)
    current_level = [root]

    for d in range(max_depth):
        next_level = []
        for node in current_level:
            # توليد 3 أفكار بديلة لكل عقدة
            candidates = [f"للخطوة_{d+1}_مسار_{i+1}" for i in range(branch_factor)]
            for cand in candidates:
                child = ThoughtNode(thought=cand, parent=node, depth=d+1)

```

```

        child.score = evaluate_thought_state(node.thought, cand)
        if child.score ≥ 0.5: # تقليم الأغصان الفاشلة
            node.children.append(child)
            next_level.append(child)

        current_level = sorted(next_level, key=lambda x: x.score, revers
e=True)[:branch_factor]
        if not current_level:
            break

        # استخراج أفضل مسار متكامل
        best_leaf = max(current_level, key=lambda x: x.score) if current_level
    else root
        full_path = []
        curr = best_leaf
        while curr:
            full_path.append(curr.thought)
            curr = curr.parent
        return list(reversed(full_path))

path = search_tree_of_thoughts("تحسين هيكل رأس المال لشركة ناشئة", max_depth=2)
print(f"أفضل مسار استدلالي معتمد: {' -> '.join(path)}")

```

6.2 رسم الأفكار البياني (Graph of Thoughts - GoT)

بينما تقتصر الأشجار على التفرع والتراجع، يسمح رسم الأفكار البياني (GoT) بعمليات شبكية متقدمة: دمج مسارين مستقلين (Aggregation)، وإنشاء حلقات تغذية راجعة لتنقيح فكرة سابقة (Feedback Loops)، مما يعكس طريقة عمل العقول البشرية في فرق العمل متعددة التخصصات.