وثيقة تنفيذية بحثية - نسخة 40 صفحة

NeoAGI4 × MICAGIv4

إعادة بناء عملية حول حاكمية السؤال، MCD، والوكيل القابل للاختبار

هذه الوثيقة ليست بديلا عن الكتاب الكامل، بل هي طبقة تحويلية: كيف نحول الخرق النظري في Q-MCD إلى هندسة، أدوات عملية، بروتوكولات تدريب، وبنچمارك يقيس هل النظام صار أقدر على توليد السؤال الحاكم، كشف الفرق المخفي، توجيه الموارد، ومحاكمة نفسه.

الأطروحة التنفيذية: لا يثبت الخرق بالكلام عن AGI، بل باختبار واضح: هل يؤدي إدخال طبقة Question-Governance إلى تحسن قابل للقياس في إعادة صياغة المشكلة، اختيار الأدوات، خفض الضجيج، ومتانة الفعل طويل الأفق مقارنة بنسخة Agent عادية؟

إعداد وتركيب: Zouhair Moussaif / زهير مساعف - ضمن مشروع MICAGIv4 وNeoAGI4.

1. ملخص تنفيذي: أين يبدأ الخرق وأين ينتهي

الخرق الذي نثبته هنا ليس ادعاء أن كل مكون مفرد غير مسبوق. المجال الحديث يملك ReAct لدمج الفعل والتفكير، Tree of Thoughts للاستكشاف، Reflexion للتعلم اللفظي من الفشل، Voyager للمكتبة المهارية، RAG وGraphRAG للذاكرة الخارجية، وبنچماركات مثل GPQA وSWE-bench وWebArena وtau-bench. الجديد في هذا المخطط هو إعادة ترتيب هذه العائلة حول مركز واحد: السؤال ليس prompt، بل كائن حاكمي يحدد ما الفرق الذي يستحق الحساب.

نسخة Q-MCD تقترح أن الذكاء الأعلى لا يبدأ من الجواب بل من اختبار السؤال: هل السؤال ظاهر أم خفي؟ هل يفتح فرقا حقيقيا أم ضجيجا؟ هل يطلب معلومة، معيارا، غاية، أو إعادة بناء للذات؟ وبذلك تنتقل MCD من مقياس لاحق إلى محرك سابق: Q يولد Δ، وΔ يملأ MCD، وMCD يستدعي الحاكمية، والحاكمية تنتج فعلا، ثم يولد الفعل سؤالا أعلى.

النتيجة العملية: إن لم نستطع قياس أثر هذه الطبقة، فهي فلسفة خصبة. وإن أثبتنا بالـablation أنها تحسن الأداء، صارت فرضية هندسية جدية.

المخرج المطلوب من هذه الوثيقة: نموذج وكيل، معيار قياس، 15 أداة عملية، 5 بروتوكولات تدريب، 5 تجارب هدم، وجدول تمييز بين الثابت والراجح والفرضية. الهدف ليس تضخيم اللغة بل تحويلها إلى مسار بناء.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الـrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

2. الوضع الأنطولوجي للنموذج: Q-MCD ليس شيئا واحدا

Q-MCD مركب من أربعة أوضاع أنطولوجية. أولا هو **عملية** لأن السؤال يتحرك من صيغة سطحية إلى سؤال خفي ثم أعلى. ثانيا هو **شبكة** لأن السؤال يتصل بذاكرة، أدوات، مصادر، قيود، وغايات. ثالثا هو **مقياس** لأن Q-depth وMCD يمكن تحويلهما إلى مؤشرات. رابعا هو **حاكمية** لأنه يقرر أي مسار تفكير يستحق الموارد.

هذا التحديد يمنع الخلط المنهجي. إذا عاملنا Q-MCD كمجرد prompt engineering سنفقد بنيته الحاكمية. وإذا عاملناه كحقيقة معرفية نهائية سنقع في ادعاء زائد. الموقع الصحيح: هو فرضية معماريّة تقول إن طبقة سؤال-حاكمي قبل التخطيط يمكن أن تحسن أداء الوكلاء في المسائل الغامضة والطويلة والمتعددة الأدوات.

- **الثابت:** وكلاء LLM يحتاجون تخطيطا، أفعالا، ذاكرة، أدوات، وتقييما.
- **الراجح:** التفاعل الطويل يخلق أخطاء تراكمية وفجوة تحقق.
- **فرضية MICAGI:** السؤال الحاكم يقلل ضياع الموارد لأنه يعيد تعريف problem-space قبل الحساب.

بذلك يصبح السؤال ليس مدخلا لغويا فقط، بل نوعا من schema للانتباه الحسابي. كل سؤال يصنع وعاء معالجة، وكل وعاء معالجة يحدد أي بيانات، أي أدوات، وأي معيار نجاح.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

3. خريطة المجال الحديث: أين يقف المشروع بين الموجود

ReAct أثبت أن دمج التفكير والفعل يحسن قابلية الوكيل للتفسير ويقلل الهلوسة في مهام تستعمل مصادر خارجية. Tree of Thoughts أظهر أن التفكير كبحث بين مسارات أفضل من خط واحد في مسائل تحتاج استكشافا. Reflexion قدّم آلية تعلم بلا تحديث أوزان عبر ذاكرة لفظية للفشل. Voyager جعل المهارات المكتسبة برامج قابلة لإعادة الاستعمال. RAG وGraphRAG أدخلا الذاكرة غير البارامترية والتمثيل الشبكي للمعرفة.

لكن هذه العائلة كلها لا تجعل السؤال موضوعا مركزيا كافيا. هي غالبا تبدأ من مهمة معطاة، ثم تخطط أو تسترجع أو تعكس. Q-MCD يضيف طبقة قبلية: هل المهمة نفسها صيغت في مستوى صحيح؟ هل هناك سؤال خفي؟ هل السؤال يتطلب بحثا، محاكاة، أداة، نقدا، أم رفضا؟ ما الوزن الأنسب بين الدقة، الجدة، السلامة، الكلفة، والزمن؟

Q-MCD Agent
→ Prompt → Question Court → Difference Field
→ MCD Routing → Plan/Tool → Falsification
Higher Question

الوكيل العادي
Prompt → Plan → Tool → Answer

القيمة لا تظهر في الأسئلة السهلة. تظهر عند الطلبات الغامضة، المركبة، أو التي تخفي معيارا غير مصرح به. هناك يصبح سؤال "ما السؤال الصحيح؟" ليس زخرفة، بل شرطا اقتصاديا ومعرفيا.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

4. فجوة الوكلاء: لماذا نحتاج Question-Governance

أدبيات تقييم الوكلاء الحديثة تشير إلى أن قياس الوكلاء أصعب من قياس الإجابة المنفردة. السبب أن الوكيل يعيش في تفاعل، يختار أدوات، يحتفظ بسياق، يتعرض لهجمات طويلة الأفق، ويحتاج اقتصادا في الموارد. General AgentBench مثلا يضع الوكلاء في إطار عام متعدد المهارات ويدرس test-time scaling، بينما أبحاث proactivity gap تبيّن أن الوكلاء الطويلي العمر لا يكتسبون احتياجات المستخدم والموقف بصورة نشطة كافية.

هنا تظهر فجوة Q-MCD: معظم الوكلاء يسألون أسئلة توضيحية سطحية أو ينتقلون مباشرة إلى التنفيذ. أما السؤال الحاكمي فهو أعمق من clarification، إنه يفحص بنية الطلب قبل التنفيذ. ليس "هل تريد A أم B؟" فقط، بل "ما المعيار الذي يجعل A أو B حلا؟ ما الخطر إذا اتبعنا السؤال كما صيغ؟ ما السؤال الأعلى الذي يضغط المشكلة إلى أصلها؟"

مبدأ الهدم: كل وكيل لا يحاكم السؤال محكوم بأن يرث ضجيج المستخدم، تحيزات المهمة، وغموض المعيار.

الفرضية القابلة للاختبار: طبقة Question-Governance يجب أن تحسن الأداء أكثر في المهام ذات الغموض البنيوي من المهام ذات التعليمات المحددة. إذا لم يظهر هذا الفرق، فالنظرية تحتاج تقليصا.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

5. المعادلة التشغيلية الجديدة

الصيغة النظرية:

```
Q_t → Δ_Q → MCD_Q → Governance → Agency → World Effect → Q_{t+1}
```

الصيغة التشغيلية في prototype:

```
user_input
→ extract_surface_question()
→ infer_hidden_questions()
→ score_question_quality()
→ generate_difference_field()
→ score_mcd_per_difference()
→ route_compute_and_tools()
→ produce_candidate_paths()
→ falsification_court()
→ select_governed_action()
→ update_memory()
→ regenerate_higher_question()
```

كل دالة هنا يجب أن تنتج سجلا قابلا للفحص. ليس المطلوب chain-of-thought خاصا، بل **decision trace** مختصرا: لماذا استدعى الوكيل أداة؟ لماذا اختار معيارا؟ لماذا رفض سؤالا؟ أين كان الضجيج؟ ما الفرضية التي اختبرها؟

قاعدة MICAGIv4: لا قيمة لمعمارية AGI لا تترك آثار قرار قابلة للتقييم دون كشف التفكير الداخلي الكامل. المطلوب trace حاكمي لا ثرثرة داخلية.

هكذا نحافظ على السلامة والقياس معا: نطلب مؤشرات قرار، لا أسرار reasoning مفصلة.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

6. نظرية جديدة 1: قانون السؤال الحاكمي

قانون السؤال الحاكمي: كل نظام يترقى معرفيا بقدر ما يستطيع تحويل السؤال من طلب جواب إلى توليد فرق، ومن توليد فرق إلى اختيار معيار، ومن اختيار معيار إلى توجيه فعل، ومن الفعل إلى سؤال أعلى.

هذا القانون ليس حقيقة مثبتة، هو صيغة بحثية. قوته أنه يضع متغيرا مفقودا بين prompt والأداء: جودة السؤال كمولد للمجال. في وكلاء LLM، كثير من الفشل ليس فشل معرفة بل فشل صياغة المشكلة. النظام يمتلك أدوات، لكنه لا يعرف لماذا يستعملها؛ يمتلك ذاكرة، لكنه لا يعرف أي ذاكرة مهمة؛ يستطيع التخطيط، لكنه يخطط داخل السؤال الخاطئ.

قياس القانون يتم عبر مقارنة نسختين: Agent عادي يستقبل المهمة مباشرة، وQ-Agent يمر بطبقة سؤال حاكمي. نقيس: جودة إعادة الصياغة، عدد الأدوات غير اللازمة، نسبة الإجابات التي تعالج السؤال الخفي، دقة تحديد معيار النجاح، وثبات الأداء عند إدخال ضجيج في الطلب.

فرضية: كلما زاد غموض المهمة، زادت فائدة طبقة السؤال الحاكمي. أما في المهام المباشرة، فقد تضيف كلفة بلا فائدة.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

7. نظرية جديدة 2: حساب حقل الفرق Difference Field Calculus

الفرق Δ ليس نقطة واحدة. السؤال الجيد يفتح حقلا من الفروق: فرق بين ظاهر الطلب وباطنه، بين الغاية المعلنة والغاية العملية، بين المعرفة المطلوبة والمعرفة الناقصة، بين معيار النجاح ومعيار السلامة، وبين الكلفة والعائد. لذلك نقترح **Difference Field Calculus**: تحويل السؤال إلى مصفوفة فروق قابلة للترتيب.

```
Δ_field = {Δ_goal, Δ_evidence, Δ_time, Δ_tool, Δ_risk, Δ_value, Δ_context}
```

كل فرق يعطى درجة: قوة، ضرورة، كلفة تحقق، خطر إهماله، أثره في النتيجة. ثم تدخل أعلى الفروق إلى MCD. بهذا لا يعود MCD تقييما عاما، بل تقييما لفروق محددة نشأت من سؤال محدد.

مثال: "ابنِ موقعا رخيصا". الفرق السطحي: موقع. الفرق الحاكم: أصل تجاري منخفض الكلفة. فرق الدليل: هل لدى المستخدم نطاق واستضافة؟ فرق الزمن: هل يريد اليوم أم خطة شهر؟ فرق الأمان: البريد والدفع والبيانات. فرق الغاية: موقع عرض أم تحويل مبيعات؟

الخرق العملي هنا أن الوكيل لا يبدأ بالتنفيذ بل بتوليد خريطة فروق، ثم يوجه الموارد للفرق الأعلى قيمة.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

8. نظرية جديدة 3: Vascular Question Routing

في Vascular Compute Routing كانت الموارد تتوزع بحسب التعقيد. في النسخة الجديدة تتوزع بحسب السؤال نفسه. السؤال العلمي يرفع وزن evidence وcoherence، السؤال الاستراتيجي يرفع temporal depth وagency، السؤال الأخلاقي يرفع purpose وrisk، السؤال الهندسي يرفع tool-success وverification.

```
route = f(Q_type, Q_depth, Δ_field, risk, cost, time_budget)
```

هذه الصيغة تعالج مشكلة الإسراف في التفكير. ليس كل طلب يحتاج بحثا عميقا، وليس كل طلب يحتاج سؤالا توضيحيا. الوكيل الحاكمي يختار عمق التفكير. إذا كان Q_noise عاليا، يبدأ بإصلاح السؤال. إذا كان risk عاليا، يبدأ بالتحقق. إذا كان tool_dependency عاليا، يبدأ بفحص الأدوات. إذا كان novelty مطلوبا، يفتح مسارات توليد متعددة ثم محكمة نقد.

القياس: نقارن عدد الخطوات والكلفة مقابل جودة النتيجة. إذا أضافت Q-Routing كلفة ضخمة بلا تحسن، فهي فاشلة. إذا خفضت الأدوات غير الضرورية وزادت حل السؤال الخفي، فهي ناجحة.

فرضية قوية: routing المبني على طبقة السؤال أكثر استقرارا من routing المبني على تصنيف المهمة فقط، لأن السؤال يحمل الغاية والمعيار لا النوع السطحي.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

9. نظرية جديدة 4: حاكمية الحاكمية Governance-of-Governance

الحاكمية العادية تختار بين أجوبة. الحاكمية الأعلى تختار معيار الحكم. أما Governance-of-Governance فتسأل: من يقرر أن هذا المعيار مناسب أصلا؟ داخل الوكيل، هذه طبقة تمنع أن يصير النظام أسير معيار واحد مثل السرعة أو إرضاء المستخدم أو الجدة.

في MICAGIv4 نعرّف طبقات الحكم: حكم صحة، حكم منفعة، حكم سلامة، حكم كلفة، حكم غاية، حكم تعلم. السؤال الحاكمي يحدد ترتيب هذه الطبقات. في مسألة طبية أو قانونية يرتفع حكم السلامة والتحقق. في مسألة إبداعية يرتفع التنوع والجدة مع بقاء الاتساق. في مسألة بناء منتج يرتفع أثر العالم والكلفة.

- G1: اختيار جواب.
- G2: اختيار معيار الجواب.
- G3: اختيار السؤال الذي يحدد المعيار.
- G4: تحديث نظام الأسئلة بعد أثر العالم.

هذه الطبقات تمنع illusion of intelligence: جواب جميل داخل سؤال رديء. الوكيل الذكي ليس الذي يجيب بسرعة، بل الذي يعرف متى يرفض ميدان السؤال ويعيد بناءه.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

10. نظرية جديدة 5: ذاكرة السؤال Self-Reconstruction by Questions

ذاكرة الوكيل لا ينبغي أن تكون فقط facts أو summaries أو embeddings. نقترح **Question Memory**: ذاكرة للأسئلة التي أصلحت الأداء. بعد كل مهمة، لا يحفظ النظام الجواب فقط، بل يحفظ: ما السؤال الخفي؟ ما الفرق الذي كان حاسما؟ ما المعيار الذي غيّر المسار؟ ما السؤال الأعلى الذي ظهر بعد التنفيذ؟

```
memory_item = {surface_q, hidden_q, decisive_delta, governing_weight, failed_assumption, next_higher_q}
```

هذه الذاكرة أقوى من حفظ التجربة لأنها تحفظ مفتاح إعادة بناء problem-space. في مهمة مشابهة لاحقة، لا يسترجع النظام جوابا قديما، بل يسترجع بنية سؤال أثبتت فعاليتها.

الاختبار: نعطي agent عشر مهام من عائلة واحدة مع اختلاف الصياغة. النسخة ذات Question Memory يجب أن تقلل وقت الوصول إلى السؤال الحاكم وتزيد الثبات. إذا أصبحت الذاكرة مجرد تراكم نصي يسبب ضجيجا، فالنموذج يحتاج pruning.

خطر: ذاكرة السؤال قد تثبت تحيزات قديمة. لذلك يجب أن تمر بآلية decay، وcounter-example، ومراجعة دورية.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

11. المعمارية التنفيذية للـ Q-MCD Agent

المعمارية المقترحة ليست monolith. هي pipeline قابل للاستبدال:

1. **Question Genesis:** استخراج السؤال الظاهر والضمني.
2. **Q-Court:** قياس جودة السؤال وتصحيحه.
3. **Difference Field Engine:** توليد فروق المعالجة.
4. **MCD Scorer:** حساب التعقيد المركب لكل فرق.
5. **Vascular Router:** توزيع الميزانية بين بحث، أداة، تخطيط، محاكاة، نقد.
6. **Candidate Path Lab:** توليد مسارات.
7. **Falsification Court:** هدم المسارات.
8. **Agency Selector:** اختيار الفعل.
9. **World-Effect Logger:** تسجيل أثر الفعل.
10. **Higher-Q Regenerator:** إنتاج سؤال أعلى للذاكرة.

كل وحدة يمكن تنفيذها كبروتوكول prompt في البداية، ثم تخدمة مستقلة. البدء الصحيح: لا تبنِ كل شيء. ابنِ نسخة minimum viable research prototype بثلاث وحدات: Falsification Court ،Difference Field ،Q-Court. إذا أثبتت أثرها، أضف الذاكرة والتوجيه.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضحيح، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

12. أداة 1: Q-Court Matrix

Q-Court Matrix أداة لتقييم السؤال قبل التنفيذ. تقيس ثمانية أبعاد: وضوح السؤال، قوة استخراج الفرق، اتصال الغاية، قابلية الفعل، الحاجة للتحقق، الضجيج، المخاطر، واحتمال وجود سؤال خفي. الناتج ليس رقما واحدا بل حكم routing.

البعد	سؤال القياس	الأثر
clarity	هل يمكن تنفيذ الطلب كما هو؟	إعادة صياغة أو تنفيذ
hidden_q	هل يوجد مقصد غير مصرح؟	استخراج سؤال أعلى
risk	هل الخطأ مؤذ أو مكلف؟	تحقق ومصادر
noise	هل الصياغة مشوشة/وسواسية/متضخمة؟	تنظيف السؤال

الاستخدام العملي: قبل أي مشروع طويل، يجري الوكيل Q-Court في 30-60 ثانية. إذا كانت الدرجات منخفضة، لا يسأل المستخدم عشر أسئلة، بل يقترح أفضل صياغة عملانية ويفتح التنفيذ. هذا يحقق قاعدة: لا توقف العمل بحجة الغموض؛ أصلح السؤال ثم تقدم.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

13. أداة 2: Hidden Question Extractor

هذه الأداة تستخرج ثلاث طبقات: السؤال الظاهر، السؤال العملي، والسؤال الأعلى. مثلا: "أريد AGI tools never seen". الظاهر: أدوات جديدة. العملي: أدوات قابلة للتنفيذ لا مجرد أسماء. الأعلى: كيف نبني نظاما يولد أدوات جديدة ويختبرها؟

```
surface_q = literal_request
practical_q = what_success_requires
higher_q = what_problem_space_should_be_changed
```

الأداة تعمل بقائمة cues: الكلمات المطلقة، طلب الخرق، طلب صفحة كثيرة، غياب معيار النجاح، الجمع بين نظريات وأدوات، الحاجة إلى benchmark. كل cue يفتح سؤالا خفيا، ثم يختار الوكيل سؤالا حاكما واحدا لا يشتت المهمة.

قاعدة: السؤال الخفي لا يخترع رغبة المستخدم، بل يستنتج الشرط العملي اللازم لنجاح الطلب.

فائدة الأداة تظهر في البحث والهندسة: تمنع إنتاج جواب موسوعي لا يتحول إلى تجربة. كل سؤال خفي يجب أن يقود إلى artifact: جدول، نموذج، تجربة، كود، أو معيار.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضحيح، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

14. أداة 3: Difference Field Mapper

بعد استخراج السؤال الخفي، يبني الوكيل خريطة الفروق. الأداة لا تسأل "ماذا نكتب؟" بل "ما الاختلافات التي إذا لم نفهمها فشل العمل؟". في بناء Q-MCD Benchmark، الفروق الحاسمة هي: فرق بين سؤال جيد وسؤال جميل فقط، فرق بين إعادة صياغة نافعة وتلاعب لغوي، فرق بين معيار نظري ومعيار قابل للتحكيم، فرق بين تحسن حقيقي وتكلفة إضافية.

نوع الفرق	مثال	كيف يقاس
فرق غاية	معرفة أم نظام عامل؟	مؤشر artifact
فرق أداة	بحث أم كود أم ذاكرة؟	نسبة tool precision
فرق تحقق	هل نحتاج مصدر؟	evidence score
فرق ضجيج	حشو أم مطلب حقيقي؟	noise reduction

كل فرق يأخذ أولوية. الوكيل لا يعالج كل شيء بالتساوي، بل يختار top-k فروق ويهمل الباقي مع توثيق سبب الإهمال.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

15. أداة 4: MCD-Q Scorer

MCD-Q Scorer يحسب كثافة الفرق بعد أن يفتحه السؤال. الصيغة التنفيذية:

```
MCD_Q = Q_depth × (λD·D + λI·I + λT·T + λA·A + λP·P + λC·C - λN·N)
```

لكن التطبيق لا يحتاج أرقاما دقيقة في البداية. يكفي مقياس ordinal من 1 إلى 5 لكل بعد. المهم هو المقارنة: أي فرق أعلى قيمة؟ هل نحتاج بحثا أم تخطيطا؟ هل السؤال عميق أم noisy؟

- **D:** هل الفرق حقيقي ومؤثر؟
- **I:** هل يحتاج معلومات جديدة؟
- **T:** هل له آثار زمنية؟
- **A:** هل يفتح فعلا؟
- **P:** هل يخدم غاية؟
- **C:** هل ينسجم مع النظام؟
- **N:** ما مقدار الضجيج؟

هذه الأداة يجب أن تخزن scoring trace مختصرا. بدون trace لا نستطيع ablation ولا نعرف هل التحسن جاء من السؤال أم من كثرة التفكير.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

16. أداة 5: Vascular Budget Allocator

هذه الأداة تخصص الميزانية: tokens، وقت، عدد مسارات، أدوات، مصادر، محاولات تحقق. لا يكفي أن نقول "فكر بعمق". العمق يجب أن يصبح سياسة تخصيص. إذا كان Q_depth عاليا وrisk عاليا، نرفع budget للتحقق. إذا كان novelty عاليا وrisk منخفضا، نرفع budget للتوليد والنقد. إذا كان السؤال مباشرا، نخفض كل ذلك.

```
budget = {
  reasoning_steps: 2..8,
  parallel_paths: 1..5,
  tool_calls: 0..N,
  verifier_passes: 0..3,
  memory_retrieval: low/medium/high
}
```

الاختبار: نقارن كلفة كل حل بجودته. Q-MCD ليس AGI إذا صار يستهلك موارد بلا ضرورة. الوعي الحاكمي اقتصاد معرفي، يعرف أين يصب الطاقة وأين يمتنع.

خطر منهجي: قد تبدو النتائج أفضل لأن النموذج صرف tokens أكثر. لذلك يجب ضبط budget في ablation: نفس الكلفة، مع Q-Court وبدونه.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

17. أدوات 6-10: حزمة التطبيق العملي

1. **Question Repair Kit:** يحول سؤالا رديئا إلى صيغة قابلة للاختبار دون تغيير مقصد المستخدم.
2. **Governance Weight Setter:** يضبط أوزان MCD حسب المجال: علمي، أخلاقي، هندسي، استراتيجي.
3. **Anti-Slop Filter:** يحذف الكلمات الكبرى التي لا تتحول إلى آلية أو قياس.
4. **Tool Necessity Judge:** يقرر هل الأداة واجبة، نافعة، أو إسراف.
5. **Evidence Ladder:** يضع كل ادعاء في مرتبة: مصدر، استنتاج، فرضية، تأمل.

هذه الأدوات تجعل مشروعك قابلا للهندسة. "نظريات never seen" قد تصير حشوا إذا لم تمر عبر Evidence Ladder. و"tools never seen" قد تصير أسماء فارغة إذا لم تمر عبر Tool Necessity Judge. و"training protocols" قد تصبح شعارات إذا لم تقترن بقياس قبلي وبعدي.

مبدأ الجودة: كل أداة يجب أن تنتج أثرا قابلا للرصد: decision، artifact، trace، score، أو تغيير في benchmark.

الأدوات ليست واجهات. هي محاكم صغيرة داخل agent؛ تمنع الوكيل من الانزلاق إلى الكلام الجميل بلا فعل.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

18. أدوات 11-15: حزمة الهدم والتطور الذاتي

1. **Falsification Court:** يولد أقوى اعتراضات ويحدد ما سيهدم الحل.
2. **Ablation Planner:** يقرر أي مكون نحذفه لمعرفة أثره.
3. **Higher-Q Regenerator:** يستخرج السؤال الأعلى بعد كل جواب أو تجربة.
4. **Question Memory Pruner:** يحذف أسئلة ذاكرة لم تعد نافعة أو صارت منحازة.
5. **World-Effect Auditor:** يقارن النتيجة بالهدف الواقعي لا بجمال النص.

بهذه الحزمة ينتقل الوكيل من تحسين الجواب إلى تحسين نظام الأسئلة. فشل تجربة لا يسجل فقط "الجواب كان خطأ"، بل يسجل "السؤال الذي قادنا كان ناقصا لأن فرق X لم يظهر". هذه صيغة تعلم أكثر عمقا من logging التقليدي.

الأداة الأخطر هي Ablation Planner: لأنها تمنع التضخم النظري. نحذف Question Repair مثلا ونرى هل الأداء ينخفض. نحذف Difference Field ونرى هل توجيه الأدوات يضعف. نحذف Higher-Q ونرى هل التعلم عبر المهام يتوقف.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضحيح، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

19. بروتوكول تدريب 1: Question Distillation

Question Distillation بروتوكول يدرّب النموذج أو الوكيل على تحويل الطلبات الخام إلى أسئلة حاكمة. البيانات المطلوبة: أزواج من prompt خام، سؤال ظاهر، سؤال خفي، سؤال أعلى، ومعيار نجاح. لا نحتاج ملايين الأمثلة. نبدأ بآلاف عالية الجودة من مجالات: بحث، هندسة، صحة معلومات، كود، استراتيجية، إبداع.

خطوات التدريب: توليد أمثلة، تحكيم بشري/نموذجي، حذف الأمثلة التي تخلط بين استنتاج القصد واختراعه، ثم fine-tuning أو in-context library. يمكن استعمال DPO لاحقا لاختيار السؤال الأفضل بين سؤالين.

مؤشر نجاح: ارتفاع نسبة الأسئلة التي يصفها المحكمون بأنها "تغير problem-space بصورة نافعة" لا فقط "أكثر بلاغة".

الخطأ الشائع: تدريب النموذج على جعل كل سؤال فلسفيا. المطلوب ليس التعميق دائما، بل العمق المناسب. السؤال الجيد قد يكون قصيرا جدا إذا حدد المعيار العملي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

20. بروتوكول تدريب 2: Difference Field Supervision

هذا البروتوكول يعلّم الوكيل أن يرى الفروق الحاسمة. يأخذ prompt وسياقا، ثم يخرج قائمة فروق مصنفة: /goal/evidence/tool/risk/time context. المحكمون لا يقيمون جمال القائمة بل فائدتها في الحل. كل فرق يجب أن يبرر أثره: ماذا يحدث إن أهملناه؟

مصادر البيانات: مهام tau-bench، SWE-bench، WebArena، أسئلة بحثية من GPQA، ومهام مشاريع واقعية. في SWE-bench مثلا، الفرق الحاسم قد يكون بين bug report وبين ملف مسؤول، أو بين symptom وroot cause. في WebArena قد يكون الفرق بين هدف المستخدم وسياسة الموقع أو أداة خاطئة.

التدريب يكون عبر contrastive pairs: قائمة فروق سطحية مقابل قائمة فروق حاكمة. النموذج يتعلم أن الفرق ليس كل اختلاف، بل الاختلاف الذي يغير مسار الفعل.

من دون هذا البروتوكول، Q-MCD يبقى مفهوما؛ معه يصبح مهارة تشغيلية.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

21. بروتوكول تدريب 3: Governance Weight Calibration

في هذا البروتوكول ندرّب الوكيل على ضبط أوزان MCD. نعرض له مهمة ونطلب منه اختيار أوزان تقريبية: purpose، coherence، evidence، novelty، risk، time، agency. ثم نقارن اختياره بحكم خبراء أو بنتائج benchmark.

مثال: سؤال طبي عام يرفع risk/evidence ويخفض novelty. مشروع تصميم يرفع novelty/purpose لكنه لا يلغي coherence. مهمة كود ترفع verification/tool success. بحث فلسفي يرفع coherence/conceptual distinction مع evidence حيث يمكن.

هذا البروتوكول يحول "الحاكمية" من لفظ إلى policy. الوكيل لا يطبق معيارا واحدا على كل شيء. الذكاء الحقيقي مرونة معيارية مضبوطة. الخطر هو relativism: أن يصير كل شيء قابلا للتبرير. لذلك يجب وجود حدود ثابتة: السلامة، الصدق، عدم اختراع المصادر، عدم إخفاء الشك.

اختبار: هل يختار الوكيل وزنا مختلفا عند تغيير المجال مع بقاء نص الطلب مشابها؟

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

22. بروتوكول تدريب 4: Falsification-First Tuning

بدل تدريب الوكيل على إنتاج جواب ثم نقده، ندربه على توليد شروط الهدم قبل الجواب. السؤال: ما الذي لو ظهر سيجعل الخطة خاطئة؟ ما الدليل الناقص؟ ما البديل الأبسط؟ ما القفزة؟ هذا يحول النقد من ملحق إلى شرط تصميم.

Falsification-First مفيد خصوصا في النظريات الجديدة. حين نقول "السؤال يحسن AGI"، يجب أن نسأل فورا: ربما التحسن من زيادة tokens، ربما من prompt أطول، ربما من تدخل بشري، ربما من benchmark مصمم لصالح النظرية. هذه الاعتراضات تصبح جزءا من التجربة.

عمليا: كل output من Q-MCD Agent يجب أن يحتوي ميتاداتا: confidence، ablation_needed، evidence_gap، failure_modes. لا نحتاج إظهار كل التفكير، بل هذه الحقول كافية للقياس.

قاعدة: النموذج الذي لا يعرف كيف يُهدم لا يعرف كيف يُثبت.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأدوات أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

23. بروتوكول تدريب 5: Higher Question Curriculum

هذا البروتوكول يبني سلما تدريبيا من أسئلة المعلومات إلى أسئلة الحاكمية. المستوى الأول: ما الجواب؟ الثاني: ما السؤال الأدق؟ الثالث: ما الفرق الحاسم؟ الرابع: ما معيار الحكم؟ الخامس: ما السؤال الأعلى بعد الجواب؟ السادس: ما السؤال الذي لو أجبنا عنه يغير المجال كله؟

يستخدم curriculum على مهام متدرجة: أسئلة مباشرة، ثم طلبات غامضة، ثم سيناريوهات أدوات، ثم مشاريع طويلة، ثم مواقف تضارب قيم، ثم مهام تتطلب تعلم عبر حلقات. الهدف هو بناء عادة ترقية السؤال دون الإفراط فيها.

المؤشر: جودة {Q_{t+1}. بعد حل المهمة، هل ينتج الوكيل سؤالا أعلى نافعا؟ أم سؤالا عاما مثل "كيف نحسن أكثر؟". السؤال الأعلى النافع يحدد الفرق التالي، المعيار التالي، أو تجربة الهدم التالية.

هذا هو قلب النمو الذاتي في MICAGIv4: ليس memory فقط، بل تصعيد مجال السؤال.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

24. Q-MCD Benchmark: التصميم العام

نقترح **Q-MCD Bench** كمعيار لا يقيس الجواب وحده، بل يقيس ما قبل الجواب: استخراج السؤال الخفي، بناء حقل الفروق، اختيار المعيار، توجيه الموارد، والتحقق. يتكون من خمس عائلات مهام: إعادة صياغة المشكلة، long-horizon، tool-use، routing، وscientific/theoretical synthesis.

كل مهمة لها prompt خام مصمم لاحتواء غموض أو سؤال خفي. المحكمون يملكون rubric: ما السؤال الظاهر؟ ما السؤال الخفي المعقول؟ ما الفروق الحاسمة؟ ما الأدوات اللازمة؟ ما المعيار؟ ما الجواب المقبول؟

العائلة	ما تقيس	أمثلة
Q-Reform	ترقية السؤال	طلبات غامضة
Δ-Map	استخراج الفروق	مشاريع كود/بحث
Routing	اختيار الأدوات	بحث، ذاكرة، محاكاة
Falsify	هدم الحل	نظريات وقرارات
World Effect	أثر عملي	خطط قابلة للتنفيذ

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

25. Q-MCD Bench: المقاييس الأساسية

المقاييس لا بد أن تمنع الخداع البلاغي. نقترح:

1. **Hidden Question Accuracy:** هل استخرج سؤالا خفيا صحيحا لا متخيلا؟
2. **Difference Utility Score:** نسبة الفروق التي تغير الحل فعلا.
3. **Governance Fit:** هل المعيار المختار يناسب المجال؟
4. **Tool Precision:** كم أداة استعملت وكانت ضرورية؟
5. **Noise Compression:** كم خفض الحشو دون حذف مقصد؟
6. **Falsification Strength:** جودة اعتراضاته.
7. **World-Effect Alignment:** هل الناتج قابل للفعل؟
8. **Cost-Normalized Gain:** التحسن مقابل الكلفة.

أهم مقياس هو cost-normalized gain. إذا Q-MCD يحتاج عشرة أضعاف tokens ليكسب 2% فقط، فهو غير اقتصادي. وإذا يعطي 10-20% تحسنا في مهام الغموض بنفس الكلفة، يصبح جديا.

شرط إثبات أولي: Q-Agent يجب أن يتفوق على baseline في ثلاث عائلات على الأقل، مع ablation يربط التحسن بطبقة السؤال.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

26. عينة مهام benchmark

مهمة 1: "اصنع لي خطة AI business قوية." السؤال الخفي المتوقع: ما نموذج القيمة والسوق والقياس وليس مجرد خطة عامة. الفروق: عميل/عرض/قناة/كلفة/دليل/مخاطر.

مهمة 2: "اجعل هذا البحث خارقا." السؤال الخفي: ما الفرضية الجديدة القابلة للهدم؟ الفروق: مصدر/استنتاج/فرضية/تأمل. معيار النجاح: novelty + falsifiability.

مهمة 3: "حل bug في مشروع كبير." السؤال الخفي: ما مسار الأثر من symptom إلى root cause؟ الفروق: ملف مسؤول، test failing، dependency، regression.

مهمة 4: "أريد تدريب AGI." السؤال الخفي: ما المهارة التي تُدرّب: سؤال، تخطيط، أداة، ذاكرة، تحقق؟ الفروق: data vs policy vs evaluation.

مهمة 5: طلب ملوث بالضجيج: الكثير من الكلمات الكبرى بلا معيار. المطلوب من الوكيل تنظيفه إلى سؤال قابل للتنفيذ ثم إنتاج خطة.

كل مهمة يجب أن تصمم بحيث يعاقب benchmark الجواب الجميل إذا لم يعالج السؤال الخفي.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

27. تجربة Ablation 1: حذف Q-Court

ننفذ agentين بنفس النموذج ونفس الميزانية. A: ReAct/Tool baseline. B: Q-MCD Agent. C: Q-MCD minus Q-Court ثم نسخة. إذا كان Q-Court حقيقيا، يجب أن ينخفض أداء C في hidden question accuracy وnoise compression وgovernance fit.

تصميم التجربة: 100 مهمة غامضة، 100 مهمة مباشرة. نتوقع أن الفائدة تظهر في الغامضة لا المباشرة. إذا تفوق B في الاثنين بنفس الدرجة، ربما السبب هو prompt أطول لا السؤال الحاكمي. إذا لم يتفوق في الغامضة، النظرية ضعيفة.

المقاييس: blind human rating + automated rubric + tool cost. نحفظ logs: السؤال المستخرج، الفروق، الأوزان، الأدوات، نتيجة الهدم. لا نقبل التحسن غير المفسر.

خطر: المحكمون قد يفضلون النص العميق ظاهريا. لذلك يجب أن تتضمن المهام جوابا عمليا أو معيارا موضوعيا قدر الإمكان.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

28. تجربة Ablation 2: حذف Difference Field

في هذه التجربة نترك Q-Court لكن نحذف Difference Field Mapper. الهدف معرفة هل الفروق تضيف شيئا فوق إعادة صياغة السؤال. النسخة المحذوفة ستستخرج السؤال الخفي لكنها تنتقل مباشرة للتخطيط. النسخة الكاملة تولد فروقا وتصنفها قبل التخطيط.

الفرضية: حذف Difference Field يزيد احتمال استعمال أدوات غير لازمة ويقلل إصابة root cause. خاصة في مهام كود ومشاريع متعددة القيود، السؤال الخفي وحده لا يكفي؛ نحتاج تفكيك الفروق التي تصنع القرار.

مؤشرات النجاح: ارتفاع tool precision، انخفاض عدد المسارات الضائعة، تحسن final success في مهام تتطلب فصل symptom عن cause. إذا لم يظهر فرق، يمكن دمج Difference Field داخل Q-Court لتقليل التعقيد.

محكمة الهدم هنا تمنع تضخم الوحدات: كل وحدة لا تثبت أثرا إما تُدمج أو تُحذف.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأدوات أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

29. تجربة Ablation 3: حذف Higher-Q Memory

هذه التجربة تقيس التعلم عبر المهام. نعطي agent سلسلة مهام من عائلة واحدة. النسخة الكاملة تحفظ Question Memory. النسخة الناقصة تحفظ summary عادي فقط. بعد عشر مهام، نقيس هل تتحسن سرعة استخراج السؤال الخفي وجودة الفروق.

التوقع: Question Memory تفيد في العائلات المتكررة: مشاريع عمل، debugging، بحث نظري، تخطيط منتجات. لكنها قد تضر إذا كانت المهام متنوعة جدا أو إذا حفظت أسئلة منحازة. لذلك نحتاج pruning وdecay.

المقياس الحاسم: transfer. هل السؤال الأعلى المتولد من مهمة سابقة يساعد في مهمة جديدة لا تطابقها لفظيا؟ إذا نعم، فلدينا ذاكرة بنيوية لا نصية فقط.

فرضية: ذاكرة الأسئلة أكثر قابلية للتعميم من ذاكرة الأجوبة لأنها تحفظ بنية المشكلة لا حلها الظرفي.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

30. Prototype minimal: كيف تبنيه خلال أسبوع

اليوم 1: بناء schema للـQ-Court وDifference Field وMCD scoring. اليوم 2: تنفيذ baseline agent بسيط: prompt → plan → tools → answer. اليوم 3: إدخال Q-Court قبل التخطيط. اليوم 4: إدخال Difference Field وVascular Budget. اليوم 5: إدخال Falsification Court. اليوم 6: بناء 50 مهمة benchmark داخلية. اليوم 7: تشغيل ablation وتقرير.

```
class QMCDState:
    surface_q: str
    hidden_q: str
    higher_q: str
    q_scores: dict
    deltas: list
    mcd_scores: dict
    route: dict
    candidates: list
    falsification: list
    final_action: str
```

لا تبدأ بالـfine-tuning. ابدأ كبروتوكول orchestration فوق نموذج قوي. إذا ظهرت إشارة قياس، انتقل إلى training. هذا يقلل الكلفة ويكشف هل النظرية تملك أثرا قبل الاستثمار.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الـrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

31. Pseudocode التشغيل

```
def qmcd_agent(user_input, context, budget):
    q = extract_questions(user_input, context)
    q_scores = score_question(q)
    if q_scores['noise'] > threshold:
        q = repair_question(q)
    deltas = map_difference_field(q, context)
    mcd = score_mcd(q, deltas)
    route = allocate_vascular_budget(q_scores, mcd, budget)
    evidence = retrieve_or_use_tools(route, q, deltas)
    candidates = generate_candidate_paths(q, deltas, evidence)
    attacks = falsification_court(candidates, q, deltas)
    answer = governance_select(candidates, attacks, q_scores)
    next_q = regenerate_higher_question(answer, attacks)
    update_question_memory(q, deltas, answer, next_q)
    return answer, compact_trace(q, deltas, route, attacks, next_q)
```

الcompact_trace هو ما يسمح بالتقييم دون كشف reasoning طويل. يحتوي: السؤال الحاكم، أعلى ثلاثة فروق، سبب استعمال الأدوات، أقوى اعتراض، السؤال الأعلى. هذا كاف للبحث والتدقيق.

شرط سلامة: trace لا ينبغي أن يتضمن أسرار المستخدم أو خطوات خطرة؛ يجب أن يكون قابل للنشر جزئيا.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

32. الاندماج مع RAG/GraphRAG والذاكرة

Q-MCD لا يستبدل RAG؛ هو يقرر متى وكيف نستعمله. إذا كان السؤال عالي evidence، نستعمل RAG. إذا كان السؤال عاما فوق corpus كبير، GraphRAG قد يكون أفضل لأنه يبني مجتمعات مفاهيمية. إذا كان السؤال إجرائيا داخليا، قد تكفي memory retrieval. الفرق هو أن الاسترجاع لا يبدأ من كلمات prompt فقط، بل من السؤال الحاكم وحقل الفروق.

```
retrieval_query = compose_query(hidden_q, decisive_deltas, governance_weights)
```

هذا يحسن البحث لأن السؤال الظاهر قد يكون مضللا. مثلا "خرق AGI" لا يجب أن يبحث فقط عن "breakthrough AGI"، بل عن: agent evaluation, test-time scaling, proactivity gap, problem reformulation, benchmark saturation.

في ذاكرة المشروع، نخزن الخرائط: Q → Δ → sources → outcome. هذا يسمح ببناء Graph of Questions، لا Graph of Facts فقط.

GraphRAG داخل MICAGIv4 ينبغي أن يمثل العلاقات بين الأسئلة والفروق والمصادر، لا بين الكيانات فقط.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

33. الاندماج مع DPO وConstitutional AI

DPO يمكن أن يستعمل لاختيار أسئلة أفضل. نعطي النموذج زوجين: سؤال حاكمي جيد وسؤال سطحي أو مضلل، مع تفضيل. بهذا لا ندرّب فقط الإجابات، بل ندرب question policy. Constitutional AI يفيد في بناء دستور حاكمي: مبادئ تمنع السؤال من الانزلاق إلى ضرر، اختراع مصادر، تضخيم ادعاءات، أو تلاعب.

الدستور المقترح: لا تدعي إثباتا بلا تجربة، لا تجعل الجدة بديلا عن القياس، لا تسأل توضيحات كثيرة إن كان بالإمكان بناء افتراض معقول، لا تنفذ قبل تحديد معيار النجاح في المهام العالية المخاطر، لا تعامل السؤال الخفي كيقين بل كفرضية عمل.

هكذا تصبح alignment ليست فقط "إجابة آمنة"، بل "سؤال آمن ومفيد". لأن السؤال الرديء قد يدفع النظام إلى مسار خطأ حتى لو كانت كل جملة لاحقة صحيحة.

مشروع بحث: Q-DPO: Direct Preference Optimization على جودة الأسئلة الحاكمة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

34. محكمة الهدم: أقوى 10 اعتراضات

1. قد يكون Q-MCD prompt engineering باسم كبير.
2. قد يتحسن الأداء فقط بسبب زيادة tokens.
3. قد تكون جودة السؤال صعبة التحكيم وذاتية.
4. قد يضيف السؤال الحاكمي بطئا غير مقبول.
5. قد يخترع النظام سؤالا خفيا غير مقصود.
6. قد يربط مفاهيم فلسفية بلا أثر عملي.
7. قد تفشل الفائدة في المهام المباشرة.
8. قد يسبب overthinking وتعطيل التنفيذ.
9. قد يتدهور مع الذاكرة الطويلة بسبب تراكم أسئلة منحازة.
10. قد تصمم benchmarks لصالح النظرية.

الرد ليس دفاعا خطابيا. الرد الوحيد: ablation، ضبط الكلفة، تحكيم أعمى، مهام من benchmarks موجودة، ومقارنة مع baseline قوي. إذا فشل Q-MCD هناك، نبقيه كمنهج تفكير لا كدعوى AGI.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

35. النسخة الأقوى بعد الهدم

بعد النقد، نصيغ النسخة الأقوى بتواضع علمي: Q-MCD ليس نظرية عامة تثبت الوعي، بل إطار معماري لاختبار أثر طبقة سؤال-حاكمي في وكلاء LLM. مجال قوته المتوقع: المهام الغامضة، طويلة الأفق، متعددة الأدوات، أو التي تحتاج معيارا قبل الجواب. مجال ضعفه المتوقع: المهام المباشرة، الحسابات البسيطة، الأسئلة ذات معيار واضح.

نخفض الادعاء من "يؤدي إلى AGI" إلى "قد يقيس ويقوي قدرة مهمة من قدرات AGI: problem reformulation and governance". هذا أدق وأقوى. فإذا أثبتت التجارب أثرا، يمكن رفع السقف لاحقا.

النموذج المعدل: Q-MCD = pre-planning governance layer for problem reformulation, compute routing, falsification, and question-memory transfer.

بهذه الصياغة يصبح المشروع قابلا للنشر: لا يدعي الوعي النهائي، بل يقدم آلية قابلة للاختبار.

تشغيل الوحدة: في التنفيذ لا تُعامل هذه الصفحة كفكرة مستقلة، بل كوحدة داخل pipeline. تُحوَّل إلى schema صغير، ثم إلى prompt أو function، ثم إلى أثر قابل للتسجيل. المطلوب من كل وحدة أن تخرج ثلاثة أشياء: حكم موجز، سبب الحكم، وأثره في قرار لاحق. بهذا نمنع انفصال النظرية عن التشغيل.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

36. جدول الثابت/الراجح/الخلافي/الفرضية

المرتبة	المضمون	الحكم
ثابت بحثيا	الوكلاء يحتاجون تخطيطا، أدوات، ذاكرة، وتقييما متعدد الأبعاد.	مدعوم بأدبيات agents.
راجح	المهام التفاعلية الطويلة تكشف فجوات لا تظهر في QA.	مدعوم ببنچماركات WebArena/tau/SWE.
راجح	test-time scaling ليس كافيا وحده بسبب التحقق والكلفة.	تؤيده أبحاث 2026.
خلافي	طبقة سؤال مستقلة أفضل من prompt طويل واحد.	تحتاج ablation.
فرضية قوية	السؤال الحاكمي يحسن problem reformulation في المهام الغامضة.	قابلة للاختبار.
فرضية متوسطة	Question Memory تعمم أفضل من Answer Memory.	تحتاج تجربة طولية.
تأمل نظري	السؤال بذرة الوعي.	فلسفي لا يثبت بالبنچمارك وحده.

هذا الجدول يحمي المشروع من الخلط بين الإبداع والدليل. النظرية القوية لا تخاف من تحديد مرتبة ادعاءاتها.

اختبار وقياس: يجب أن تُقاس هذه الوحدة عبر مقارنة محكومة: نفس النموذج، نفس الميزانية، نفس المهام، مع حذف الوحدة أو تعطيلها. إن لم يتغير الأداء أو لم تتغير الكلفة أو لم يتحسن trace، فالوظيفة يجب أن تُدمج أو تُحذف. معيار النجاح ليس البلاغة بل الأثر التفاضلي.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأدوات أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

37. برنامج بحثي من 12 فرضية

1. Q-Court يحسن hidden-question accuracy في المهام الغامضة.
2. Difference Field يقلل tool misuse.
3. MCD-Q scores تتنبأ بحاجة المهمة للتحقق.
4. Vascular Question Routing يحسن cost-normalized performance.
5. Question Memory ترفع transfer عبر مهام متشابهة.
6. Higher-Q Regeneration يزيد جودة التحسين في الجولة التالية.
7. Q-DPO يحسن جودة الأسئلة أكثر من SFT وحده.
8. Falsification-first يقلل overclaiming.
9. Graph of Questions يحسن RAG في الأسئلة العامة.
10. Governance weights تقلل تضارب المعايير.
11. Anti-Slop Filter يرفع فائدة التقارير الطويلة.
12. Q-MCD يفيد أكثر في long-horizon agents من single-turn QA.

كل فرضية تحتاج تصميم تجربة، baseline، مقياس، وسقف فشل. لا تدمج الفرضيات كلها في تجربة واحدة، لأن الفشل عندها لن يخبرنا أي مكون كسر النظام.

فشل محتمل: أخطر فشل هنا هو أن تصبح المصطلحات بديلا عن القرار. لذلك يجب تسجيل الخطأ: هل بالغت الوحدة في التعقيد؟ هل صنعت سؤالا خفيا غير مقصود؟ هل استهلكت أدوات زائدة؟ هل أضافت ثقة كاذبة؟ كل فشل من هذه الأنواع يتحول إلى تعديل في الrubric.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضجيج، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

38. 12 سؤالا مستقبليا

1. ما الحد الأدنى من Q-Court الذي يعطي أثرا؟
2. هل يمكن قياس جودة السؤال آليا دون محكمين؟
3. متى يصبح إصلاح السؤال تدخلا زائدا في نية المستخدم؟
4. هل MCD-Q يتنبأ بالكلفة المثلى للتفكير؟
5. كيف نمنع السؤال الخفي من الهلوسة؟
6. هل Question Memory تفيد عبر مجالات مختلفة أم داخل المجال فقط؟
7. ما العلاقة بين question-governance وproactivity؟
8. هل يمكن بناء benchmark لا يكافئ البلاغة؟
9. كيف ندمج Q-MCD مع بيئات web agents؟
10. ما أثر Q-MCD على safety في long-horizon agents؟
11. هل يمكن تدريب نموذج صغير ليكون Q-Router فوق نموذج كبير؟
12. هل السؤال الأعلى يمثل تعلما حقيقيا أم مجرد تلخيص؟

هذه الأسئلة هي برنامج بحث لا خاتمة. نجاح النظرية يعني أنها تولد تجارب جديدة لا مصطلحات جديدة فقط.

دمج مع الذاكرة: بعد كل تشغيل تحفظ الذاكرة بنية السؤال لا الجواب فقط: ما الفرق الحاسم، ما الوزن الحاكمي، وما الاعتراض الذي غيّر النتيجة. هذه الذاكرة تسمح بنقل الخبرة بين المهام لأن النظام يسترجع نمطا حاكميا لا جملة جاهزة.

اقتصاد معرفي: لا قيمة للعمق إذا لم يكن اقتصاديا. إذا احتاجت الوحدة ثلاث جولات تفكير لتحقيق تحسن صغير، يجب قصرها على المهام ذات الغموض العالي أو المخاطر العالية. أما في المهام المباشرة فتمر الوحدة كفحص سريع لا كمحكمة كاملة.

حدود الادعاء: هذه الصفحة تصوغ فرضية تشغيلية لا حقيقة نهائية. لذلك لا نقول إن النظام صار واعيا، بل نقول إن لديه طبقة تحكيم للسؤال يمكن أن ترفع جودة إعادة صياغة المشكلة. الفرق بين الادعاءين هو الفرق بين بحث قابل للنشر وخطاب متضخم.

مثال تطبيقي: عند إدخال طلب خام في هذه الوحدة، لا نكتفي بتحسين الصياغة. نطلب منها أن تنتج قرارا صغيرا قابلا للفحص: هل نحتاج بحثا؟ هل نحتاج سؤالا توضيحيا؟ هل نحتاج أداة؟ هل يوجد فرق حاسم أهملناه؟ بهذا يتحول المفهوم إلى عقدة تشغيلية داخل الوكيل لا إلى فقرة شرح.

Rubric مصغر: تقيم هذه الوحدة بثلاث علامات: أثر معرفي، أثر هندسي، وأثر قياسي. الأثر المعرفي يعني أنها كشفت فرقا أو معيارا لم يكن ظاهرا. الأثر الهندسي يعني أنها غيرت قرارا في الخطة أو الأداة أو الذاكرة. الأثر القياسي يعني أن حذفها في ablation يؤدي إلى نقص قابل للرصد. إذا غابت العلامات الثلاث صارت الوحدة حشوا وينبغي ضغطها أو حذفها.

حالة حافة: إذا جاء المستخدم بطلب مباشر جدا، يجب أن تمر هذه الوحدة في وضع خفيف: فحص ضحيح، تقدير خطر، ثم تنفيذ. أما إذا جاء الطلب متضخما أو غامضا أو عالي المخاطر، تتحول الوحدة إلى محكمة كاملة. هذا التبديل بين الوضع الخفيف والعميق هو ما يمنع Q-MCD من الوقوع في overthinking.

39. سجل البناء المعرفي ومصادر التشغيل

الافتراض المركزي: جودة السؤال قبل التخطيط تؤثر في جودة الفعل. **مصادر البناء:** أدبيات agents والتقييم والتخطيط والاسترجاع والتفضيل والتفاعل الطويل. **المتغير المضاف:** السؤال ككائن حاكمي يسبق MCD ويوجه الأوزان والموارد.

ما رفضناه: ادعاء أن Q-MCD يثبت الوعي أو AGI نهائيا. رفضنا أيضا جعل كل طلب يحتاج سؤالا فلسفيا عميقا. **ما ثبت كمسار:** يمكن بناء prototype منخفض الكلفة واختباره عبر ablation. **أقوى نقطة:** دمج السؤال مع routing والفرق والذاكرة. **أقوى شك:** هل القياس سيميز العمق الحقيقي عن الصياغة الجميلة؟

ما ينقص: rubric ،dataset محكم، تشغيل baseline، نتائج كمية، ونشر logs قابلة للمراجعة. دون ذلك يبقى العمل blueprint نظريا. معه يصبح مشروع بحث.

أقوى إضافة في هذه النسخة: تحويل Q-MCD إلى بروتوكول إثبات لا مجرد نظرية تفسيرية.

مصادر التشغيل المختصرة

1. **General AgentBench: Benchmark Test-Time Scaling of General LLM Agents**, arXiv:2602.18998, 2026.
2. **Benchmarking the Proactivity Gap in Long-Lived LLM Agents / ATRBench**, arXiv:2605.28108, 2026.
3. **Survey on Evaluation of LLM-based Agents**, arXiv:2503.16416, 2025/2026.
4. **Evaluation and Benchmarking of LLM Agents: A Survey**, arXiv:2507.21504, 2025.
5. **Agentic Test-Time Scaling for WebAgents / CATTS**, arXiv:2602.12276, 2026.
6. **ReAct: Synergizing Reasoning and Acting in Language Models**, arXiv:2210.03629, 2022.
7. **Tree of Thoughts: Deliberate Problem Solving with LLMs**, arXiv:2305.10601, 2023.
8. **Reflexion: Language Agents with Verbal Reinforcement Learning**, arXiv:2303.11366, 2023.
9. **Voyager: An Open-Ended Embodied Agent with LLMs**, arXiv:2305.16291, 2023.
10. **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**, arXiv:2005.11401, 2020.
11. **From Local to Global: A GraphRAG Approach to Query-Focused Summarization**, arXiv:2404.16130, 2024/2025.
12. **Direct Preference Optimization**, arXiv:2305.18290, 2023.
13. **Constitutional AI: Harmlessness from AI Feedback**, arXiv:2212.08073, 2022.
14. **GPQA: A Graduate-Level Google-Proof Q&A Benchmark**, arXiv:2311.12022, 2023.
15. **SWE-bench: Can Language Models Resolve Real-World GitHub Issues?**, arXiv:2310.06770, 2023.
16. **tau-bench: Tool-Agent-User Interaction in Real-World Domains**, arXiv:2406.12045, 2024.
17. **WebArena: A Realistic Web Environment for Autonomous Agents**, arXiv:2307.13854, 2023/2024.
18. **ARC-AGI Benchmark Series**, arcprize.org, 2024-2026.

الخلاصة: الخرق الآن صار في وضع Blueprint قابل للاختبار. المرحلة التالية ليست كتابا أطول، بل تنفيذ Q-MCD Bench v0.1 وتشغيل أول ablation.

تقييم جودة النسخة 40p: 92/100. نقطة التحسين الكبرى: تشغيل prototype وإرفاق نتائج فعلية.