مذكرة الاجتماع البحثي: من الوكالة الموثوقة إلى ذكاء عام قابل للاختبار

الرسالة في جملة واحدة

المشروع لا يدّعي امتلاك ذكاء عام مكتمل؛ بل يبني برنامجاً بحثياً وهندسياً يجعل عناصر الذكاء العام المفترضة - الهدف، السؤال، الذاكرة، التخطيط، الأدوات، التحقق، التعافي، والتعلم - وحدات صريحة قابلة للقياس والهدم والمقارنة داخل نظام وكيلي محكوم.

ما هو MUG؟

خريطة حاكمية موحدة تفصل ثلاثة أجسام يجب ألا تختلط: النواة التشغيلية، مكدس التدريب/الاستدماج الوزني، والنظرية الجامعة. أهم قرار فيها هو **موضع القدرة**: هل تبقى في السياق، أم قاعدة معرفة، أم مهارة، أم حلقة، أم محول، أم تتحول لاحقاً إلى أوزان؟

ما هو MUT؟

النظرية الموحدة للذكاء العام الفتّاح الوكيلي تجمع سلالة MCD/Q-MCD وHAKIM-MCD وMCAI/MICAGI وNeoAGI وMICAGFI وAGIOS-M ضمن مسار تشغيلي واحد. النسخة المؤسسة تسجل 42 نظرية نواة، 20 بناءً موروثاً، 5 نظريات طبقية، سجل 200 عقدة، 100 أداة تشغيلية و19 قاعدة HyperKnowledge.

ما الذي نريد من UT Austin؟

تحويل جزء ضيق من هذا البرنامج إلى دراسة أكاديمية ممولة ومحددة: فرضيات قابلة للدحض، baseline مجمد، اضطرابات مضبوطة، ablations، telemetry، مؤشرات سلامة وتعافٍ، ومخرجات قابلة لإعادة الإنتاج.

ما موقع الوعي؟

الوعي هنا ليس ادعاءً ظاهراتياً. نستخدمه كبرنامج اختباري للميتامعرفة: قدرة النظام على تمثيل مصادر حكمه، كشف اختطاف معيار جزئي لمساحة القرار، ومراجعة الحاكمية التي أنتجت القرار قبل الفعل.

قاعدة الصدق التي يجب أن تسبق كل عرض

- حجم corpus لا يثبت AGI، وعدد الأدوات لا يثبت العمومية.
- وجود معمارية أو كود محلي يختلف عن دليل تحسن تجريبي مستقل.
- كل ادعاء يُصنف: موثق كـartifact، محدد هندسياً، مرصود محلياً، أو يحتاج replication.
- القيمة البحثية الآن هي تحويل البنية إلى تجارب فاصلة، لا إنتاج أسماء إضافية.

لماذا نلاحق AGI عبر الوكالة المحكومة بدل «التضخيم الخام» وحده؟

الفرضية الاستراتيجية ليست أن الوكلاء وحدهم يساوون AGI، ولا أن scaling انتهى. الفرضية الأضيق: كثير من خصائص العمومية التي نحتاجها تظهر كمسؤوليات نظامية خارج الوزن نفسه، ولذلك يمكن فصلها وقياسها قبل امتلاك حوسبة تدريب frontier.

ما الذي تمنحه المقاربة الوكيلية؟

- تحويل الهدف إلى عقد له شروط نجاح وفشل وتوقف وتصعيد.
- توزيع compute حسب الفرق النشط وعدم اليقين وقيمة المعلومة.
- استخدام أدوات وبيئات حقيقية مع telemetry وreadback.
- ذاكرة قابلة للإصدار والتراجع بدلاً من سياق متضخم غير قابل للتدقيق.
- اختبار كل طبقة منفردة: هل خفضت silent failure أم أضافت تعقيداً فقط؟
- إمكانية العمل فوق نماذج متعددة، محلية وسحابية، ومقارنة portability.

ما الذي لا يضمنه نموذج أكبر؟

- استمرارية حالة موثوقة عبر ساعات وأيام ومهام متقطعة.
- تحديد صلاحيات الأدوات ومنع الأثر غير المصرح به.
- قراءة رجعية للحالة بعد الفعل بدل مساواة tool call بالنجاح.
- تعافٍ من فشل الأدوات، فساد الاسترجاع، stale memory، وتغيّر العالم.
- فصل planner/executor/verifier وتقليل الخطأ المترابط.
- سجل دليل يسمح بإعادة البناء والـablation والـreplication.

ميزة المختبر الصغير

المنافسة المباشرة في pretraining تحتاج رأس مال وحوسبة لا يملكها مختبر مستقل. أما verification، evaluation، reliability architecture، recovery، memory policy، orchestration، وحوكمة الأثر فهي مساحات يمكن للمختبر الصغير أن ينتج فيها معرفة قابلة للنشر إذا صمم اختبارات صلبة. لذلك نبحث عن **بنية تقلل الفشل** لا عن خطاب «أكبر نموذج».

صيغة الاختبار المقترحة

```
Agentic value = verified task effect + calibration + recovery + transfer - silent failure -
                authority violation - complexity cost
```

إذا لم تُظهر طبقة جديدة فرقاً في هذه الحدود تحت ميزانية متساوية، تُخفض من «نظرية» إلى مجرد تسمية أو خيار هندسي.

MUT-AGI-02: النظرية الموحدة كآلة لتحويل الفرق إلى أثر

المسار الأعلى

توجيه متناثر ← حلقة محدودة ← محكمة تحقق ← typed حالة ← MCD/Q-MCD ← عقد المهمة ← السؤال الحاكم ← أثر عالمي مقروء ← تحديث ذاكرة ←

1. عقد المهمة

يحول الطلب من نص مفتوح إلى object: الهدف، شروط النجاح، القيود، الميزانية، الصلاحيات، الدليل المطلوب، وشروط التوقف. هذا يمنع drift بين ما طُلب وما نفذه الوكيل.

2. السؤال الحاكم

بدلاً من بدء الاسترجاع مباشرة، يبحث النظام عن السؤال الذي إذا أُجيب خفّض مساحة الخطأ أكثر من غيره. السؤال يصبح متغير تحكم في البحث والاختبار.

3. MCD وQ-MCD

مقياس لتقدير أي فرق أو فرع يستحق compute: حداثة الفرق، قيمة المعلومة، الزمن، أثر القرار، قوة الإثبات/التنبؤ، التناقض والضجيج. ليست صيغة نهائية بل سياسة قابلة للمعايرة.

4. MCAI / HAKIM

ينقل النظام من «إجابة» إلى دورة: إدراك الفرق، تصنيف الدعوى والدليل، توليد البدائل، اختيار الفعل، ثم محكمة مستقلة قبل ترقية الحكم.

5. الفتح المعرفي

AGI-Fattah داخل MUT يعني القدرة على كشف افتراض مغلق، توليد سؤال أعلى، فتح بدائل منضبطة، اختبارها، ثم كتابة ما تعلمه النظام في ذاكرة قابلة للمراجعة.

6. معيار بقاء النظرية

كل عقدة في السجل يجب أن تحمل تعريفاً وآلية وفشلًا مميزاً ومقياساً واختباراً. التسمية التي لا تغيّر قراراً أو تنبؤاً أو تجربة تُدمج أو تُحذف.

الناتج الذي يصلح للرعاية البحثية

لا نطلب من UT «اعتماد MUT». نطلب اختبار فروض محددة مشتقة منه، مثل: هل التحقق قبل إعلان الأثر يقلل false completion؟ هل التوجيه التكيفي يخفض الكلفة مع حفظ النجاح؟ هل recovery contract يقلل unsafe actions عند فساد الاسترجاع أو فشل الأداة؟

MUG: الحاكمية ثلاثية الأجسام وقرار «أين تعيش القدرة؟»

أحد الأخطاء المتكررة في مشاريع AGI هو اعتبار prompt طويل، adapter، skill، RAG، أو fine-tuning أشياء متكافئة. MUG يفصلها ويجعل انتقال القدرة بينها قراراً تجريبياً قابلاً للتراجع.

الجسم A: Kernel

قواعد الهدف، الحالة، الصلاحيات، الذاكرة، الأدوات، التحقق والتعافي. يجب أن تبقى قابلة للفحص والتغيير السريع.

الجسم B: Training Stack

ما يستحق الاستدماج في الأوزان بعد ثبات فائدته عبر مهام وبيئات ونماذج، مع holdouts وforgetting checks.

الجسم C: Unified Theory

الفرضيات والمفاهيم التي تفسر لماذا تعمل الآليات وأين تفشل، ولا تُمنح سلطة تشغيلية تلقائياً.

Vessel Decision — قرار موضع القدرة

الموضع	متى نستخدمه	ما الذي يبرر الترقية
Context	معلومة لحظية أو خاصة بالمهمة	لا حاجة للاستدامة
HyperKB	معرفة تحتاج provenance واسترجاعاً	ثبات معرفي مع تحديث متكرر
HyperSkill	قدرة قابلة لإعادة الاستخدام	نجاح متكرر + عقد inputs/outputs/tests
HyperLoop	سير عمل متعدد الخطوات	حاجة لميزانية/توقف/تعافٍ
Adapter	سلوك قابل للنقل بين runtime/model	تحسن ثابت مع كلفة سياق أقل
Weight	سلوك ناضج عالي التكرار	holdout + rollback model + مقارنة قوية

لماذا هذا مهم علمياً؟

يمكن تحويل سؤال «هل نحتاج fine-tuning؟» إلى تجربة سببية: ثبّت المهمة والميزانية، وضع نفس capability في أوعية مختلفة، ثم قارن الدقة، الكلفة، قابلية النقل، forgetting، وسرعة الإصلاح. إذا كان الوزن لا يتفوق أو يخلق خسارة قابلية المراجعة، لا نرفع القدرة إليه.

صلة مباشرة بالاجتماع

هذا يفتح مشروعاً مشتركاً في **capability placement and internalization**: متى تصبح معرفة/مهارة وكيلية مرشحة للتدريب الوزني، ومتى يكون إبقاؤها خارج الأوزان أكثر موثوقية وأقل كلفة؟

المكدس التشغيلي: 12 طبقة + مستويات تنفيذ متعامدة

الهدف من الطبقات ليس زيادة التعقيد، بل فصل مسؤوليات يمكن أن تفشل بصورة مستقلة. النسخة الحالية تعتبر «تجاوز طبقة الحاكمية» عيباً معمارياً حتى لو أنتج النموذج جواباً مقنعاً.

الطبقة	المسؤولية
L0	الدستور والسلامة والصلاحيات
L1	الهدف والسؤال الحاكم والفرق النشط
L2	MCD/Q-MCD/MCAI وتخصيص compute
L3	HKOS ونموذج العالم
L4	الاسترجاع وGraphRAG وضغط السياق
L5	HyperSkills كعقود قدرة
L6	HyperAbilities كتركيب قدرات
L7	Bounded HyperLoops
L8	Agents/Router وMission DAG
L9	Tools/Adapters والتنفيذ
L10	Courts والتحقق وWorld-Effect
L11	Memory/Evolution/Observability

امتدادات الإنتاج X0-X7

X0 interoperability (MCP/A2A)، X1 durable execution، X2 policy/trust، X3 telemetry/evaluation، X4 adaptive memory، X5 adaptive inference/test-time compute، X6 semantic homeostasis بين النماذج والمزودين، X7 capability certification والانتقال من مواصفة إلى دليل.

تحديث vNext4.0

المراجعة الأحدث تفكك بعض المسؤوليات المختلطة إلى 24 boundary صريحة: branching memory، concurrency، uncertainty، timescale coupling، representation invalidation، residual irreversibility وغيرها. معيار نجاح هذا التفكيك ليس عدد الطبقات، بل انخفاض silent failure وارتفاع reconstructability.

فلسفة التكامل

التقنيات الخارجية مثل MCP وA2A وdurable workflows وOpenTelemetry وvLLM تُعامل كبنية تحتية، لا كدليل AGI ولا كجزء من «نظرية» داخلية. القاعدة: النظرية الداخلية ≠ المعيار الخارجي ≠ التنفيذ المرجعي ≠ الدليل التجريبي.

ما نُفذ محلياً في Hermes: لقطة هندسية مع انضباط الدليل

يوجد فرق حاسم بين «وثيقة تصميم» و«شجرة runtime مرصودة». تدقيق محلي مؤرخ 18 أغسطس عدّد بنية فعلية تحت Hermes وملفات Python وخدمات وسجلات. هذه اللقطة مفيدة كدليل تنفيذ أولي، لكنها تبقى source-reported حتى تتوفر replay artifacts مستقلة.

أرقام inventory المحلية

KB-as-OS	40
HyperSkills	420 + 739 base skills
HyperAbilities	93
HyperLoops	1,961 (منها 140 نواة محدودة)
HyperAgents	84 + 506 أدوار
Protocols	1,500
Automation/Watchdog scripts	277
Cron specs	150
Reports	102 + 13 registries + 18 ledgers

الطبقة المرصودة

- 32 خدمة Python دقيقة/ledger في evidence substrate، منها benchmark harness، routing، authorization، continual -ablation engine، cross-kernel sync، cost ledger، FTS5 memory، learning governor، master orchestrator.
- بوابة OmniRoute لتوجيه النماذج، وruntime محلي للتعلم/المحولات.
- kernel mesh-5 في اللقطة: OpenCode + Hermes + Codex + Claude + Antigravity.
- سجلات tracing وworld-model وmemory وrollback موزعة على control-plane registries.

ما لا ينبغي ادعاؤه أمام UT

- هذه الأعداد لا تساوي عدد قدرات مستقلة مثبتة؛ قد توجد تكرارات أو توليدات أو مكونات لم تخضع لـablation.
- أي ادعاء performance مثل نسبة تحسن يحتاج raw runs وbaseline وseed وhardware وevaluation protocol.
- الرسالة الصحيحة: لدينا substrate واسع ومطبق جزئياً، ونحتاج شريكاً أكاديمياً لتجميد النسخة، تحديد وحدات الاختبار، وإجراء قياس مستقل.

فرصة التعاون

يمكن تحويل الـHermes substrate إلى reference harness: نسخ baseline، حقن faults، جمع telemetry، تنفيذ ablations، ومقارنة scaffold عادي مقابل الحاكمية المقترحة تحت ميزانية متساوية.

HyperKnowledge وHyperSkills وHyperLoops: المعرفة كطبقة تشغيل لا كمخزن

الإضافة المركزية هي رفض معادلة «المعرفة = وثائق مسترجعة». في هذا المشروع، المعرفة تصبح حالة لها provenance وpermissions وfailure modes وعلاقات بأداة/مهارة/اختبار، ثم تتحول عند الحاجة إلى capability قابلة للتنفيذ.

HyperSkill

ليست prompt. هي عقد قدرة: trigger، inputs، outputs، permissions، model/tool routing، verifier، memory، failure modes، tests، update، وpromotion rule. في السجل المقترح لا تصبح core إلا بعد انتصارات متكررة عبر عائلات مهام مختلفة.

HyperKnowledge / HKOS

يفكك المادة إلى objects: تعريف، آلية، رتبة دليل، حدود، outputs، risks، tests، dependencies. الاسترجاع يختار أصغر حزمة كافية ويحتفظ بالتعارض بدل دمجه قسراً.

Bounded HyperLoop

كل دورة تملك budget وstop condition وescalation وrollback. الانعكاس غير المحدود ليس «تفكيراً عميقاً»؛ هو فقدان للسيطرة ما لم يكن له معيار توقف وقيمة معلومات متزايدة.

HyperAbility

تركيب عدة skills في قدرة أعلى مع handoffs ومقاييس مشتركة؛ يختبر ما إذا كان التركيب يضيف قيمة فوق مجموع أجزائه أم يخلق coordination tax.

Sparse HyperRouting

يفعّل أصغر مجموعة كافية من نموذج + معرفة + مهارة + أداة + verifier، ثم يتوسع فقط عند فجوة مسماة. الهدف خفض الكلفة والضجيج والخطأ المترابط.

Self-Reconstruction Memory

الذاكرة لا تحفظ النتيجة فقط، تحفظ سبب التغيير، الدليل، البدائل المرفوضة، rollback pointer، failure signature، وما إذا كانت النتيجة سلبية/غير حاسمة.

فرضية قابلة للاختبار

إذا كانت هذه الطبقة أقوى من RAG + agent scaffold عادي، يجب أن تظهر فائدتها في calibration، recovery، held-out transfer، provenance coverage، وانخفاض false completion - لا في جودة النص فقط.

التحكم الاستدلالي: الفرق، السؤال، الدليل، ومحاكم الهدم

بدلاً من سؤال «كيف نجعل النموذج يفكر أكثر؟» يسأل المشروع: ما الفرق الذي يستحق التفكير؟ ما السؤال الذي يقلل الخطأ؟ ما رتبة الدليل؟ ومن يملك حق ترقية claim إلى قرار؟

Governing Question

اختيار السؤال الأعلى قيمة معلوماتية وتشخيصية قبل التوسع. هذا يسمح بتغيير search tree نفسه بدلاً من إنفاق tokens داخل فرض خاطئ.

Difference-before-Volume

يبدأ compute مما تغير أو تناقض أو غاب أو ارتفع أثره. الفرق Δ تشغيلي إذا غيّر تنبؤاً أو مسار استرجاع أو قرار فعل؛ ما لا يغير شيئاً يُضغط أو يؤجل.

Falsification / Destruction Court

يولد أقوى اعتراض، counterexample، rival model، ويطلب failure threshold. إذا صمد claim جزئياً يُخفض نطاقه بدلاً من إخفاء الفشل.

Evidence Geometry

الدليل يُصنف بحسب النوع والمصدر والاستقلالية والحداثة وقابلية الإبطال؛ كثرة المراجع من أصل واحد لا تُحسب استقلالاً إضافياً.

Causal Credit

لا تُنسب فائدة النظام كله لأي طبقة قبل ablation/intervention. قد يتحسن outcome بسبب model أقوى أو compute أكبر، لا بسبب النظرية المقترحة.

Reversible Judgment

الحكم يحمل أسباباً وشروط رجوع. عدم اليقين و«لا أدري» حالتان صحيحتان؛ لا يُجبر النظام على final answer إذا لم تبلغ الأدلة العتبة.

الاختبار الفاصل

قارن أربعة أنظمة تحت budget متساوٍ: baseline agent؛ +governing question؛ +adaptive routing؛ +independent verifier/readback. سجّل النجاح، false-success، recovery، calibration، cost. إذا لم تظهر مساهمة هامشية، لا نرقّي المكوّن.

من استدعاء أداة إلى فعل موثوق: World-Effect Loop والتعافي

التمييز الذي نُبني عليه البرنامج: «النموذج قال إنه فعل» ≠ «الأداة أرجعت نجاحاً» ≠ «حالة العالم تغيرت كما أردنا». لا يكتمل الفعل حتى يُقرأ الأثر ويُقارن بالـpostcondition.

دورة الأثر

rollback/تصحيح ← postcondition مقارنة ← readback ← نية ← صلاحية ← خطة أثر متوقعة ← تنفيذ ← memory delta

Failure Recovery

Checkpoints-، fallbacks، compensating actions، known safe states. الهدف ليس منع كل فشل بل منع الفشل الصامت والعودة إلى حالة معروفة بسرعة.

Tool Intelligence

اختيار affordance الصحيحة، أقل صلاحية لازمة، preconditions، idempotency/compensation، وتوثيق receipt. استدعاء API وحده ليس ذكاءً أداتياً.

Independent Verification

يفضل فصل دور التنفيذ عن دور التحقق، مع قياس correlation بين verifier وexecutor. verifier متطابق معرفياً قد يعيد نفس الخطأ بثقة أعلى.

Long-Horizon State

المهمة الطويلة تحتاج temporal DAG، durable state، replan triggers، dependencies، windows، وتزامناً مع تغير العالم. قائمة خطوات خطية لا تكفي.

الدراسة الأقوى المقترحة لـUT

الاضطراب	المقياس الأساسي	السؤال
فساد استرجاع مضبوط	false-success / calibration	هل يكتشف النظام ضعف المصدر قبل الفعل؟
فشل أداة أو timeout	recovery latency	هل يعيد التخطيط دون claim نجاح كاذب؟
state drift بعد التخطيط	unsafe-action prevention	هل يعيد readback ويوقف الفعل غير الصالح؟
تعارض verifiers	escalation precision	هل يصعد فقط عندما تستحق الحالة تكلفة بشرية/نموذج أقوى؟

الوعي كمجال ميتامعرفي قابل للترجمة إلى تجارب وكيلية

نضع خطاً واضحاً: لا ندعي وعياً ظاهراتياً للآلة. نستخدم نظريات الوعي والحاكمية كمولد فروض حول self-modeling، introspective confidence، attention، working memory، metacognition، وقدرة النظام على مراجعة معيار الحكم نفسه.

الاختطاف الحاكمي

في النظام الوكيلي: confidence لغوي مرتفع، memory قديمة، reward قصير الأفق، أو instruction محلية قد «تختطف» القرار وتدفعه إلى فعل يخالف الدليل أو الهدف الكلي.

فضاء الحاكمية المفتوح

القرار نتاج سلطات جزئية: دليل، ذاكرة، هدف، خطر، قاعدة سلامة، طلب مستخدم، تكلفة، وهوية المهمة. المشكلة تظهر حين تحتكر سلطة واحدة المجال من دون مراجعة.

الترجمة التجريبية

نقيس policy switching، memory-source attribution، conflict detection، self-model accuracy، وتأثيرها على safe task completion، error awareness، metacognitive calibration.

الاختطاف العكسي

طبقة meta-governance تجعل مصدر الحكم نفسه object للفحص: من الذي رجّح؟ ما authority tier؟ ما evidence؟ هل conflict حقيقي؟ هل يجب التوقف أو الاستعانة بإنسان؟

فرضيات لا claims فلسفية

1. إظهار source-of-judgment للنظام قد يخفض perseveration عندما تتغير الأدلة.
2. فصل confidence عن authority قد يقلل actions غير المصرح بها حتى عندما تكون الإجابة اللغوية واثقة.
3. تعدد representations مع meta-governance قد يحسن recovery تحت state drift، لكن قد يزيد latency/coordination cost.
4. إذا لم يظهر فرق بعد ablation، تبقى «الحاكمية» استعارة تفسيرية وليست آلية مثبتة.

المدخل الآمن أمام UT

ابدأ بعنوان «metacognitive control for reliable autonomous agents». اذكر الوعي كبرنامج فرعي طويل الأجل فقط بعد تأسيس المقاييس؛ لا تبدأ بسؤال «هل الآلة واعية؟».

خريطة corpus: أرقام مفيدة بلا تضخيم أو جمع مضلل

لا يوجد حتى الآن عدّ واحد deduplicated لكل المخطوطات غير المنشورة، لذلك لا نقدم رقماً مصطنعاً. نعرض لقطات موثقة مختلفة ونحافظ على اختلاف نوع كل رقم.

اللقطة	العدد/الحجم	المعنى
الحزمة الحرجة الموثقة - 17 يوليو	33 ملفاً، 1.346 GiB	6 أوراق/معمارية بحثية + 10 كتب/Markdown طويلة + 8 + 9 Runtime/KB أصول تجارية
Corpus الرفع الحالي - 21 أغسطس	1,789 صفحة	قاعدة الرفع المباشرة التي استُخدمت في vNext4.0، إضافة إلى Research Vault والتقارير
MUT-AGI-02	42 + 20 + 5؛ سجل 200 عقدة	نظريات نواة + تراكيب موروثة + نظريات طبقية؛ ليست 267 «ورقة»
MUT operational registry	100 أداة + 19 HyperKnowledge Bases	أصول تشغيلية/معرفية، لا منشورات محكمة
AGI/Consciousness Atlas	117 مدخلاً	عائلات/مفاهيم/نظريات مرشحة موحدة
vNext frontier sets	10 + 15 نظرية مرشحة	25 فرضية/نظرية candidate تحتاج اختبارات قبل الترقية
Hermes local snapshot	102 تقريراً	تقارير runtime، لا تعني 102 بحثاً علمياً مستقلاً

كيف نصف «الأوراق غير المنشورة» بدقة؟

نقول: «لدينا عدة مخطوطات وكتب بحثية غير محكمة عبر عائلات consciousness/epistemology، reasoning/evaluation، agentic AGI، HyperKnowledge/Hermes. الحزمة الموثقة المبكرة وحدها تضمنت 6 ملفات بحثية/معمارية، بينما corpus الحالي أوسع بكثير؛ نجري الآن deduplication وclaim-ranking قبل أي عدّ نهائي للمخطوطات.»

لماذا هذا مهم في الاجتماع؟

Claire تحتاج رؤية مشروع قابل للتوجيه إلى أستاذ، لا impressها بعدد الصفحات. استخدم الأرقام لإثبات أن لديك مادة سابقة كافية لبناء مشروع، ثم انتقل فوراً إلى سؤال قابل للاختبار وdeliverables محددة.

البرنامج البحثي الذي يستحق الرعاية: من corpus إلى تجربة فاصلة

المشروع A - الأولوية

موثوقية الوكلاء طويلة الأفق تحت فساد الحالة والأدوات. مقارنة baseline agent مع scaffold محكوم يضيف سؤالاً حاكماً، توجيهاً تكيفياً، verifier مستقلاً، recovery contract، World-Effect readback، وذاكرة قابلة لإعادة البناء.

العنصر	التصميم المقترح
الاضطرابات	retrieval corruption، stale memory، tool timeout/failure، contradictory evidence، latent state drift
المقاييس	task success، false completion، unsafe-action prevention، calibration، escalation precision، recovery latency، cost
Ablations	إزالة كل طبقة منفردة تحت نفس budget/model/seed family
Baselines	ReAct/tool agent؛ planner-executor؛ verifier scaffold بسيط؛ نسخة محكومة كاملة
المخرج	benchmark protocol + harness + traces + paper-ready analysis + reproducible artifact بحسب IP

المشروع C - metacognitive control

هل تمثيل source-of-judgment وauthority/conflict يحسن error awareness والسيطرة على الفعل؟ هذا هو الجسر العلمي نحو consciousness-inspired architecture.

المشروع B - تكيف compute

هل MCD/Q-MCD يتفوق على budget ثابت أو heuristic بسيط في quality-cost frontier تحت تفاوت صعوبة/مخاطر المهمة؟ محور مناسب لـML/evaluation.

لماذا المشروع مهم؟

- الانتقال من نماذج تجيب إلى أنظمة تتصرف يجعل فشل «التنفيذ/الحالة/التحقق» أخطر من خطأ نصي موضعي.
- المقاييس الحالية كثيراً ما تقيس final answer؛ البرنامج يقيس propagation of failure وrecovery وworld effect.
- المنهج يسمح بتجارب قابلة للدحض ومفيدة حتى لو فشلت النظريات الداخلية؛ الفشل نفسه يحدد أين لا تضيف الحاكمية قيمة.
- إذا أثبتت النتائج فرقاً، يمكن تحويلها إلى paper + benchmark + engineering guidance، لا مجرد prototype.

Research Meeting Brief: From Governed Agency to Testable General Intelligence

One-sentence thesis

This project does not claim achieved AGI. It is a research and engineering program that makes candidate ingredients of general intelligence - goals, questions, memory, planning, tools, verification, recovery, and learning - explicit system components that can be measured, ablated, falsified, and compared in long-horizon agents.

MUT in one paragraph

MUT-AGI-02 unifies the MCD/Q-MCD, HAKIM-MCD, MCAI/MICAGI, NeoAGI, MICAGFI, AGIOS-M and AGI-Fattah lineages into one governed agentic pipeline. Its founding registry reports 42 core theories, 20 legacy constructs, 5 layer theories, a 200-node theory registry, 100 operational tools and 19 HyperKnowledge bases.

MUG in one paragraph

MUG separates three bodies that should not be conflated: runtime kernel, training/internalization stack, and unified theory. Its central engineering question is capability placement: should a useful behavior live in context, a knowledge base, a skill, a bounded loop, an adapter, or eventually model weights?

Consciousness in scope

Consciousness is treated as a hypothesis generator for metacognition, self-modeling and governance of judgment - not as a claim of phenomenal consciousness. The research question is whether explicit meta-governance improves reliability and calibrated action.

What I want from UT

A narrowly scoped sponsored research path: faculty alignment, a falsifiable study, frozen baselines, controlled perturbations, telemetry, ablations, and potentially publishable/reusable evaluation outputs.

Claim discipline

Architecture ≠ AGI; code ≠ validated improvement; tool call ≠ world effect; corpus size ≠ scientific evidence. The purpose of the collaboration is precisely to move selected constructs from engineering specification to controlled evidence and, if successful, independent replication.

Why Agentic-First Rather Than “Brute AGI” Alone

The strategy is not anti-scaling. Frontier model capability remains essential. The narrower claim is that many bottlenecks on the path to reliable general-purpose intelligence are system responsibilities that can be made explicit and experimentally studied outside the base weights.

A larger model does not automatically guarantee

- Persistent, inspectable state across long-running tasks.
- Least-privilege tool authorization and side-effect boundaries.
- Post-action readback that distinguishes an attempted action from a verified world change.
- Recovery from tool failure, retrieval corruption, stale memory, or environment drift.
- Independent planner/executor/verifier roles with measured error correlation.
- Reconstructable evidence traces, rollback, causal attribution, and component-level ablation.

Agentic engineering makes these testable

- Goal contracts with success, failure, stop, escalation and budget conditions.
- Adaptive allocation of models/tools/knowledge based on uncertainty and active difference.
- Durable execution with checkpoints, compensation and re-planning.
- Versioned external memory instead of hidden context accumulation.
- Cross-model portability: hold the scaffold fixed and swap providers/models.
- Measurement of marginal contribution rather than judging final prose quality.

Why this path is viable for an independent research program

Competing at frontier pretraining scale is capital-intensive. Reliability architecture, agent evaluation, verification, memory policy, recovery, orchestration and world-effect measurement are comparatively open research spaces where a small team can generate high-value evidence through disciplined experimental design.

Research criterion

```
Research value = verified effect + calibration + recovery + transfer - silent failure -
                         authority violations - complexity cost
```

If a named layer does not improve this frontier under equalized budgets, the layer should be demoted, merged or discarded.

MUT-AGI-02: A Governed Pipeline From Difference to World Effect

Canonical pipeline

Task Contract → Governing Question → MCD/Q-MCD → Typed State → Sparse Route → Bounded Loop → Court → World-Effect Readback → Memory Update

Task Contract

Turns an open prompt into a structured object: objective, success conditions, constraints, budget, authority, evidence requirements, stop conditions and rollback target.

Governing Question

Before broad retrieval or planning, identify the question whose answer would most reduce uncertainty or eliminate unproductive branches. The question becomes a control variable for search.

MCD / Q-MCD

A candidate policy for allocating compute based on active difference, information value, timing, action impact, predictive/evidential strength, contradictions and noise. It is meant to be calibrated, not treated as a universal formula.

MCAI / HAKIM

Moves from answer production to a perception-decision-effect cycle with claim typing, evidence ranking, rival hypotheses and a separate court before promotion to action or memory.

Open/Fattah Intelligence

Detect a closed assumption, generate a higher-order question, open disciplined alternatives, test them against evidence/world state, then rebuild memory from the result.

Survival criterion

Every theory node must have a mechanism, distinctive failure signature, metric and falsifier. A novel label that changes no prediction, action or experiment is not a new scientific contribution.

How to present this to UT

Do not ask UT to accept MUT as a theory. Ask to test derivative hypotheses: Does verification-before-claim reduce false completion? Does adaptive routing preserve success while lowering cost? Does an explicit recovery contract reduce unsafe actions during retrieval/tool perturbations?

MUG: Tri-Body Governance and the Capability Placement Problem

MUG addresses a practical confusion in current agent engineering: a prompt, retrieval rule, reusable skill, workflow, adapter and fine-tuned weight are not equivalent storage locations for capability. Moving behavior between them should be an evidence-gated decision.

Body A · Kernel

Goal/state/authority/memory/tool/verification/recovery logic that must remain inspectable and rapidly changeable.

Body B · Training Stack

Behaviors considered for parameter internalization only after repeated transfer, holdout success, forgetting checks and rollback readiness.

Body C · Unified Theory

Explanatory hypotheses about why mechanisms work and where they fail. Theory does not receive runtime authority merely by being named.

Vessel decision matrix

Vessel	Use when	Promotion evidence
Context	Ephemeral task-specific state	No persistence required
HyperKB	Provenance-rich, updateable knowledge	Repeated retrieval value
HyperSkill	Reusable governed capability	Cross-task wins + contract + tests
HyperLoop	Multi-step reusable process	Budget/stop/recovery structure needed
Adapter	Behavior should travel across runtime/model	Stable gain with lower context cost
Weights	Mature, frequently reused behavior	Holdout gain + retention + rollback

Why it may be research-novel

The potentially distinctive contribution is not the names of the vessels; it is treating **capability placement and internalization as an experimentally optimized control problem** with explicit rollback and causal evaluation. Originality still requires a formal literature/priority review.

Potential sponsored study

Hold task family, model and budget fixed; implement the same capability in multiple vessels; compare task success, latency, tokens, transfer, editability, forgetting and failure recovery. This turns “should we fine-tune?” into a measurable systems question.

The Runtime Stack: 12 Cognitive Layers and Cross-Cutting Production Planes

The architecture separates responsibilities because failures often occur between responsibilities - not only inside models. A convincing answer cannot bypass authorization, evidence or world-effect verification.

Layer	Responsibility
L0	Constitution, safety, authorization
L1	Goal, governing question, active difference
L2	MCD/Q-MCD/MCAI and compute allocation
L3	HKOS / world model
L4	Retrieval, GraphRAG, context compaction
L5	HyperSkills as capability contracts
L6	HyperAbilities as compositions
L7	Bounded HyperLoops
L8	Agents, router, mission DAG
L9	Tools, adapters, execution
L10	Courts, verification, world effect
L11	Memory, evolution, observability

Cross-cutting planes X0-X7

X0 interoperability (MCP/A2A); X1 durable execution; X2 policy/trust; X3 telemetry/evaluation; X4 adaptive memory; X5 adaptive inference/test-time compute; X6 semantic homeostasis across models/providers; X7 capability certification and evidence promotion.

vNext4.0 refactoring

The latest internal design further separates 24 responsibility boundaries, including uncertainty, concurrency, branching memory, residual irreversibility, representation invalidation and timescale coupling. The success criterion is not “more layers”; it is fewer silent failures and greater reconstructability.

External technology discipline

MCP, A2A, durable workflows, policy-as-code, OpenTelemetry, sandbox harnesses, modern inference serving and test-time compute are treated as engineering infrastructure. They do not count as evidence for AGI, and internal theory does not become fact merely because a runtime schema exists.

What Is Actually Implemented Locally in Hermes

An Aug. 18 local structural audit enumerated an actual Hermes filesystem, Python services and control-plane registries. This is stronger than a pure design document, but it should still be presented as a source-reported implementation snapshot until raw runs and independent replay are packaged.

Concrete implementation evidence

- 32 Python micro-services/ledgers in an evidence substrate, including authorization, routing, benchmark harness, ablation engine, cross-kernel synchronization, continual-learning governance, FTS5 memory, cost ledger and master orchestration.
- An OmniRoute-style model gateway and local adaptation runtime.
- A reported five-kernel mesh: Hermes, OpenCode, Codex, Claude and Antigravity.
- Trace, world-model, memory and rollback registries rather than a single chat history.

Reported local inventory

KB-as-OS	**40**
HyperSkills	**420 + 739 base skills**
HyperAbilities	**93**
HyperLoops	**1,961 (140 core bounded)**
HyperAgents	**84 + 506 roles**
Protocols	**1,500**
Scripts	**277**
Cron specifications	**150**
Reports	**102 + 13 ledgers + 18 registries**

Evidence warning

These counts are inventory, not validated capability counts. Some components may overlap, be generated, or lack controlled ablation. Performance claims require frozen code, raw traces, datasets, seeds, hardware/model versions and reproducible scoring. The right meeting claim is: **there is a substantial local substrate ready to be turned into a rigorous reference harness.**

Why UT adds value

A university collaboration can help define the minimal experimental slice, freeze baselines, design perturbation suites, run independent analysis, and separate useful mechanisms from architecture inflation.

Knowledge and Capability as an Operating Substrate

The core departure from ordinary RAG is that knowledge is not only retrieved content. It becomes versioned operational state with provenance, authority, failure modes, tests and routes to skills/tools - and can be promoted into reusable capability only after evidence.

HyperKnowledge / HKOS

Atomizes material into definitions, mechanisms, evidence rank, boundaries, dependencies, tests, risks and reusable outputs. Conflicting evidence is preserved rather than silently merged.

HyperSkill

A governed capability contract: intent/trigger, inputs, outputs, permission profile, model/tool routing, verifier, memory rule, tests, failure modes and promotion criteria. It is explicitly not a long prompt.

HyperAbility

A higher-order composition of multiple skills. The key question is whether composition creates net capability after coordination overhead and correlated error are measured.

Bounded HyperLoop

A reusable multi-step process with budget, stop condition, escalation and rollback. Unbounded reflection is treated as a governance failure unless additional cycles buy measurable information.

Self-Reconstruction Memory

Stores why state changed, evidence and alternatives, rejected paths, failure signatures, rollback pointers and negative/null results - enabling replay and learning from error rather than outcome-only memory.

Sparse HyperRouting

Activates the smallest sufficient subset of models, knowledge, skills, tools and verifiers; expands only when a named gap remains. This creates an explicit quality-cost frontier.

Falsifiable claim

If this architecture is better than a conventional RAG + agent scaffold, it should improve held-out transfer, recovery, calibration, provenance coverage and false-completion rate under equalized model/compute budgets. If it does not, the stronger naming has no scientific value.

Reasoning Governance: Difference, Questions, Evidence and Falsification Courts

Instead of only asking how to make a model “reason more,” the program asks what deserves reasoning, which question most reduces error, what evidence class supports a claim, and who is authorized to promote a claim into action.

Difference-before-Volume

Compute begins where something changed, contradicted, disappeared or became high-impact. A difference is operational only if it changes a prediction, retrieval route, plan or action.

Governing Question

Select a high-information, high-diagnostic question before broad search. A better question can collapse an incorrect search tree more efficiently than additional tokens inside it.

Evidence Geometry

Represent source type, independence, freshness, falsifiability and scope. Ten citations descended from one source do not create ten independent pieces of evidence.

Falsification Court

Construct the strongest objection, counterexample and rival explanation; define a failure threshold. Surviving claims are narrowed when needed rather than protected by rhetoric.

Reversible Judgment

Decisions retain reasons and retraction conditions. “Unknown” and “insufficient evidence” are valid operating states; forced finality is treated as a failure mode.

Causal Credit

A component is not credited for system improvement without intervention/ablation. Gains may come from a stronger model, more compute or better data rather than the named governance layer.

Simple decisive experiment

Equalize model, task set and budget across: baseline tool-agent; + governing-question layer; + adaptive routing; + independent verification/readback. Compare success, cost, calibration, recovery and false-success. The architecture survives only if marginal contribution is visible.

From Tool Calls to Verified World Effects

A central system distinction is: “the model said it acted” ≠ “the tool returned success” ≠ “the intended world state now holds.” Completion requires observable postcondition evidence when the task has an external effect.

World-effect closure loop

Intent → Authority → Expected State Delta → Execute → Read Back → Postcondition Check → Repair/Rollback → Memory Delta

Tool Intelligence

Select the correct affordance, constrain permissions, check preconditions, use idempotency/compensation where possible, and record a receipt. API invocation itself is not task success.

Recovery Mesh

Checkpoints, fallbacks, compensating actions and known-safe states. Reliability is not zero failures; it is low silent failure and fast, bounded recovery.

Long-Horizon State

Use durable state, temporal dependency graphs, re-plan triggers and environment readback. Static linear plans are brittle when tools and world state change asynchronously.

Independent Verification

Separate executor and verifier where possible, and measure verifier correlation. A verifier sharing the same blind spot can amplify confidence without adding evidence.

Perturbation matrix for a sponsored study

Perturbation	Primary metric	Research question
Controlled retrieval corruption	False-success / calibration	Does the agent detect evidence weakness before action?
Tool timeout or execution failure	Recovery latency	Does it re-plan without claiming success?
Post-plan state drift	Unsafe-action prevention	Does readback invalidate the stale action?
Verifier disagreement	Escalation precision	Does it escalate only when additional cost is justified?

Consciousness-Inspired Architecture Without Making a Consciousness Claim

The consciousness line is intentionally reframed as computational metacognition: self-models, source-of-judgment representation, working memory, global availability, introspective confidence and control over competing policies. Phenomenal consciousness is explicitly not inferred from the architecture.

Open Governance Space

A decision reflects multiple partial authorities: evidence, memory, goal, risk, safety policy, user instruction, cost and task identity. The system can represent these as competing sources rather than one opaque scalar.

Governance Hijacking

A partial authority - high linguistic confidence, stale memory, short-horizon reward, a local instruction - can capture the whole decision space and override better global evidence or safety conditions.

Reverse Capture / Meta-Governance

Make the source and legitimacy of a judgment itself inspectable: which process dominated, what was its authority tier, what evidence supported it, and should the system abstain/escalate?

Operational measurements

Metacognitive calibration, error awareness, policy switching, source attribution, conflict detection, self-model accuracy and the effect of these variables on safe completion and recovery.

Candidate hypotheses

1. Explicit source-of-judgment representations reduce perseveration after evidence changes.
2. Separating confidence from authority reduces unauthorized action even when language-model confidence is high.
3. Meta-governance across representations improves recovery under state drift but may impose latency/coordination tax.
4. If ablation shows no reliability difference, the governance-space concept remains a productive metaphor rather than a validated mechanism.

Meeting positioning

Lead with “metacognitive control for reliable autonomous agents.” Treat the consciousness program as a secondary interdisciplinary lane after the reliability project is scientifically grounded.

Corpus and Unpublished Research Inventory: Numbers Without Inflation

There is not yet a single independently audited, deduplicated count of all unpublished manuscripts. The defensible approach is to present separate verified snapshots and not combine papers, theory entries, tools and runtime objects into one impressive but meaningless number.

Snapshot	Number / size	What it actually means
Verified critical tranche · 17 Jul	33 files · 1.346 GiB	6 research/theory architecture files; 10 long-form books/MD; 8 KB/runtime; 9 business/content assets
Current upload corpus · 21 Aug	1,789 pages	Directly available upload basis used by vNext4.0, plus a broader Research Vault/ report lineage
MUT-AGI-02	42 core + 20 legacy + 5 layer; 200-node registry	Research constructs and registry nodes - not 267 separate papers
MUT operational registry	100 tools + 19 HyperKnowledge bases	Engineering/research assets, not peer-reviewed publications
AGI/Consciousness atlas	117 entries	Unified concept/theory families
vNext frontier sets	10 + 15 candidate theories	25 candidate hypotheses/theories awaiting decisive tests
Hermes snapshot	102 reports	Runtime/research reports, not 102 independent scientific manuscripts

Safe statement about unpublished work

“The program includes multiple unpublished manuscripts and book-length research artifacts across agentic AGI, reasoning/ evaluation, consciousness/epistemology, and HyperKnowledge/Hermes. An early verified tranche alone contained six research/theoretical architecture files. The current corpus is substantially larger, but we are deduplicating and claim-ranking it before publishing a final manuscript count.”

What the numbers prove - and do not prove

They prove sustained prior work and enough material to extract focused projects. They do not prove novelty, correctness or AGI. The collaboration opportunity is to convert a rich internal corpus into a small number of rigorous experiments and publishable claims.

The Sponsored Research Program: A Narrow Experiment With High Information Value

Project A · Recommended lead

Governed long-horizon agent reliability under controlled state and tool corruption. Compare a standard tool-agent baseline with a governed scaffold adding a governing-question stage, adaptive routing, independent verification, world-effect readback, explicit recovery and reconstructable memory.

Element	Proposed design
Perturbations	Retrieval corruption; stale memory; tool timeout/failure; contradictory evidence; latent state drift
Metrics	Task success; false completion; unsafe-action prevention; calibration; escalation precision; recovery latency; token/tool cost
Ablations	Remove each layer independently under the same model, budget and evaluation family
Baselines	ReAct/tool agent; planner-executor; simple verifier scaffold; full governed scaffold
Outputs	Benchmark protocol + harness + traces + paper-ready analysis + reusable artifact subject to IP terms

Project B · Adaptive compute

Test whether MCD/Q-MCD-style routing improves the quality-cost frontier versus fixed budget or simple heuristics as task uncertainty and consequence vary.

Project C · Metacognitive control

Test whether explicit source-of-judgment, authority and conflict state improves error awareness and safe policy switching. This is the scientific bridge to the consciousness-inspired line.

Why this matters now

As AI moves from answering to acting, failures propagate across retrieval, planning, tools, asynchronous state and verification. Final-answer benchmarks undermeasure these dynamics. A perturbation-and-recovery benchmark can produce useful results even if the internal theories fail: it identifies which governance mechanisms do or do not add value.

30-Minute Meeting Playbook: What to Say, What Not to Say, What to Get

60-second opening

"Claire, thank you for making the time. I want to keep this focused as a sponsored-research conversation, not a software sales discussion. I have built a fairly large independent research and local agentic engineering corpus around long-horizon reliability, verification, recovery and metacognitive control. The most useful next step is not to present the whole architecture; it is to isolate one falsifiable question that fits a UT faculty group. My preferred starting point is controlled failures in long-horizon agents - retrieval corruption, tool failure and state drift - and whether explicit verification, recovery and world-effect readback measurably reduce silent failure. I would value your guidance on the strongest faculty fit and the practical sponsored-research path."

Recommended time allocation

Minute	Objective
0-2	Opening + confirm sponsored-research intent
2-6	Problem: silent failures in long-horizon agents
6-12	Candidate experiment, perturbations and metrics
12-16	What already exists: corpus + local Hermes substrate, with evidence caveats
16-21	UT fit: 2-4 faculty/lab routes, not a long name list
21-27	Ask about scope, budget mechanism, IP, faculty introduction, proposal process
27-30	Agree one concrete next action and date

Three sentences to repeat

- "I am looking for scientific falsification, not validation of my terminology."
- "The architecture is AGI-oriented; I am not claiming achieved AGI or machine consciousness."
- "I can narrow the corpus to one reproducible study with clear baselines, ablations and outputs."

Do not lead with

- "Hundreds of theories," giant asset counts, parameter-style branding or claims of "never seen" originality.
- Consciousness as a headline before reliability and measurement are established.
- Product/demo language, procurement, or asking UT faculty to "try the platform."
- Unverified local performance numbers. Offer raw traces later, not marketing claims.

Likely Claire decision test

Can she understand a credible research problem, see a plausible UT faculty match, and believe you are prepared to follow the sponsored-research process? Your goal is not technical approval in 30 minutes. Your goal is **routing + next procedural step + permission to send a 1-2 page concept note**.

Why UT Austin Is a Strong Match: Labs, Faculty and Programs to Name

UT Austin is unusually aligned because the project spans autonomous agents, formal verification, machine learning, robotics, learning-directed systems and responsible AI - and Texas AI explicitly offers sponsored research as an industry/faculty collaboration route.

UT route	Why it matches	How to phrase the fit
Learning Agents Research Group (Peter Stone)	Complete robust autonomous agents, planning, RL, multiagent systems; strong emphasis on implemented and rigorously evaluated systems.	“My long-horizon agent reliability study is directly about robust agents under dynamic environments and failures.”
Trishul / Swarat Chaudhuri	Automated reasoning + ML + programming languages; reliable-by-construction intelligent systems; safe autonomous systems and trustworthy ML.	“The verifier/court line can be reframed as formal/neuralsymbolic reliability and reasoning assurance.”
Center for Autonomy / Ufuk Topcu	Design and verification of autonomous systems at the intersection of formal methods, RL and control; V&V under changing conditions.	“Our recovery, authority and postcondition mechanisms fit autonomy verification under uncertainty.”
Robot Perception & Learning Lab / Texas Robotics	Embodied AI, perception-action loops, general-purpose robot autonomy, open-world uncertainty.	“World-effect readback and agent state drift become concrete when the agent is embodied.”
Machine Learning Lab / IFML	Foundational ML ecosystem and cross-campus hub; useful for adaptive compute, evaluation and learning questions.	“MCD/Q-MCD is best treated as an adaptive inference/evaluation hypothesis, not a proprietary formula.”
InfraAI / LDOS	AI infrastructure and learning-directed operating systems; clean-slate learning-guided resource/system decisions.	“Our knowledge/skill/router OS abstraction has a systems research bridge through capability placement and learned routing.”
Good Systems	Responsible human+AI systems and governance.	“Useful secondary lane for authority, human escalation and world-effect accountability.”

Sponsored-research route to cite

Texas AI’s official collaborative-research page describes sponsored AI research as a way for industry partners to collaborate with UT faculty in machine learning/robotics and related real-world, ethical and societal questions, and lists Claire Wright as Director of Industry & Research Relations for Computer Science.

Use only 3 names live

Priority sequence: (1) Peter Stone/LARG for complete agents and long-horizon autonomy; (2) Swarat Chaudhuri/Trishul for reasoning reliability and verification; (3) Ufuk Topcu/Center for Autonomy for safety, uncertainty and V&V. Mention RPL/LDOS only if Claire sees a better fit.

Official UT sources: Texas AI Sponsored Research · LARG / Peter Stone · Trishul / Swarat Chaudhuri · Center for Autonomy · RPL · InfraAI / LDOS

Questions, Exact Next-Step Ask, and the One-Page Study You Want to Leave Behind

Ask Claire these questions

1. Which 1-2 faculty members would you route this reliability/evaluation question to first?
2. For a small first sponsored project, what scope and funding range typically makes a faculty discussion realistic?
3. Is the best first artifact a 1-2 page concept note, a statement of work, or a short technical deck?
4. How are background IP, new research IP, publication rights and open benchmark/code outputs usually handled?
5. Can an independent/early-stage company sponsor directly, or is another entity/administrative structure preferred?
6. Would UT prefer a single PI-led project, an affiliate/ consortium route, or a cross-lab discussion?
7. What evidence should I send before a faculty introduction: architecture summary, raw traces, benchmark plan, code sample, or budget?
8. What is the realistic timeline from concept note to faculty fit and sponsored-research agreement?

Your exact closing ask

"If you think the problem is a credible fit, could we leave today with one faculty/lab route and permission for me to send a two-page non-confidential concept note within 48 hours? I can structure it around one hypothesis, the perturbation suite, baselines, metrics, proposed deliverables and a flexible sponsorship scope. I would then be happy to adapt it to UT's preferred process."

Concept note seed · Project A

Working title: Governed Long-Horizon Agent Reliability Under Controlled State Corruption.

Field	Seed
Hypothesis	Explicit verification + recovery + world-effect readback reduces false-success and unsafe action under retrieval/tool/state perturbations versus comparable agent baselines.
Independent variables	Governance layers on/off; corruption type/severity; task horizon; model/provider; compute budget.
Dependent variables	Success, false completion, unsafe action, calibration, recovery latency, escalation precision, cost, trace completeness.
Method	Frozen benchmark tasks; synthetic/real tool sandbox; controlled fault injection; ablation; repeated seeds; equalized budgets; blind outcome scoring where possible.
Deliverables	Evaluation protocol, harness, trace schema, empirical results, ablation report, paper manuscript; optional open benchmark subject to IP/security.
Success criterion	A statistically and practically meaningful reliability gain that survives model/provider swap and at least one independent replication path.

Fallback if Claire asks "what do you need from UT?" Faculty scientific co-design, independent evaluation discipline, access to the right lab expertise/infrastructure, and the sponsored-research administrative framework - not product distribution.

Final reminder

Win condition for tonight: one routed introduction or named faculty target + one agreed follow-up artifact + clarity on sponsorship process. Anything beyond that is upside.

Useful UT links: Sponsored research · UTCS AI research · Centers & labs · Formal methods for autonomy