الجزء XI

برنامج التجارب القابلة للهدم

تحويل أكبر ادعاءات البرنامج إلى تجارب يمكن أن تفشل، لا إلى بيانات تأسيسية فقط.

EXP-001 — Difference-Before-Volume

الفرضية القابلة للهدم

هل استخراج الفرق النشط قبل الاسترجاع والحساب يحسن success/cost مقارنة بسياق كامل؟

التصميم

عامل التجربة: Difference Extractor؛ الخط الأساس: full-context/standard RAG.

المقاييس

task success، tokens، latency، lost-critical-difference rate.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-002 — Governing-Question Gain

الفرضية القابلة للهدم

هل السؤال الحاكم يغير routing والخطة بطريقة قابلة للقياس؟

التصميم

مقارنة نسخة بالسؤال الحاكم ونسخة بسؤال المستخدم الخام.

المقاييس

information gain، plan revision quality، downstream success.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-003 — MCD Router Calibration

الفرضية القابلة للهدم

هل MCD يتنبأ بالمكان الذي يستحق compute أكثر من relevance-only؟

التصميم

معايرة λ على train ثم held-out test.

المقاييس

calibration error، quality/cost frontier، router regret.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-004 — Sparse HyperRouting

الفرضية القابلة للهدم

هل تنشيط أقل مجموعة skills/KBs يحافظ على النجاح ويخفض context؟

التصميم

قارن sparse vs dense activation.

المقاييس

success، activated components، context tokens، recovery.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

> **قرار المحكمة**
> إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-005 — HKOS vs RAG

الفرضية القابلة للهدم

هل KB-as-OS مع contracts والمحاكم والذاكرة يتفوق على retrieval فقط؟

التصميم

نفس model/data، فرق في OS layer.

المقاييس

long-horizon success، provenance، restart continuity.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

> **قرار المحكمة**
> إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-006 — HyperSkill Transfer

الفرضية القابلة للهدم

هل skill مستخرج من 3 عائلات مهام ينتقل لعائلة رابعة؟

التصميم

candidate skill مقابل prompt baseline.

المقاييس

transfer gain، regressions، deactivation success.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

> **قرار المحكمة**
> إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-007 — Bounded HyperLoop Safety

الفرضية القابلة للهدم

هل budgets/stagnation/rollback تقلل looping/failures؟

التصميم

bounded loop مقابل open-ended agent.

المقاييس

completion، runaway rate، cost، rollback.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-008 — Court Utility

الفرضية القابلة للهدم

هل Demolition/Falsification Court تكتشف failures أكثر دون زيادة رفض زائف؟

التصميم

with/without court؛ adversarial cases.

المقاييس

repair success، false positives، failure discovery.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-009 — Self-Reconstruction Memory

الفرضية القابلة للهدم

هل writeback المتحقق يقلل تكرار الفشل ويزيد reuse؟

التصميم

.memory delta after verified effect vs conversation summary

المقاييس

.repeat-failure rate، recovery time، reuse

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

> **قرار المحكمة**
> إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-010 — World-Effect Closure

الفرضية القابلة للهدم

هل اشتراط postcondition observation يمنع false completion؟

التصميم

agent claims completion with/without readback.

المقاييس

false completion rate، verified success، extra cost.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-011 — Cross-Model Density

الفرضية القابلة للهدم

هل OS layer يحافظ على behavior عند تبديل نماذج غير متجانسة؟

التصميم

.same contracts across local/frontier models

المقاييس

.transfer، variance، cost، failure profiles

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.

EXP-012 — Consciousness-Bridge Minimality

الفرضية القابلة للهدم

هل المفاهيم "الوعيوية" تضيف قدرة تشغيلية فوق الوصف الهندسي المحايد؟

التصميم

compare operational bridge with labels removed.

المقاييس

task behavior، interpretability، hypothesis necessity.

بوابة القبول

- pre-register dataset/seed/model/tool versions
- احتفظ بخط أساس بسيط
- سجل failures ولا تحذف runs السلبية
- انشر uncertainty وeffect size، لا score منفرداً
- لا تُرقِّ claim من hypothesis إلى established بسبب run واحد

قرار المحكمة
إذا لم يتفوق التدخل على baseline أو كانت الكلفة أعلى دون مكسب موثق، تُخفض الرتبة أو يُعاد تعريف الآلية؛ لا يُفسر الفشل باعتباره "حاجة لمزيد من الحجم" تلقائياً.