

<https://doi.org/10.62405/osiconf.2026.15>

Коломієць Євген,
Інститут загальної та неорганічної хімії
ім. В. І. Вернадського НАН України, Київ, Україна,
ORCID: 0000-0002-8941-663X, e-mail: kolomiets@ionc.kiev.ua

EX-ANTE ОЦІНЮВАННЯ ПРИКЛАДНИХ ДОСЛІДНИЦЬКИХ ТА ІННОВАЦІЙНИХ ПРОЄКТІВ: САМОАУДИТ ГЕЙТОВАНОЇ ПРОЦЕДУРИ З АГРЕГАЦІЄЮ ЗНИЖЕНОЇ КОМПЕНСАТОРНОСТІ

АНОТАЦІЯ. *Мета роботи* — задокументувати конструкцію гейтованої процедури ex-ante оцінювання прикладних дослідницьких та інноваційних проєктів і піддати її самоаудиту замість декларування переваг. *Методи:* аналітичне стрес-тестування правил агрегації та коригування оцінки на всьому діапазоні допустимих значень факторів; прогін процедури на калібрувальній вибірці з 22 історичних кейсів; зіставлення заявленого захисту від зміщення мовних моделей із наявними емпіричними бенчмарками. *Результати:* виявлено дві конструктивні вади. Зважене геометричне середнє знижує, але не усуває компенсаторність — фактор на нижній межі нормування 0,02 з вагою 0,20 дає індекс 0,457 за порога 0,35. Коригування оцінки за класом доказовості інвертується: за ваги довіри нижче 0,143 увесь досяжний діапазон індексу лежить вище порога, тож жодну ідею не може бути відхилено. *Висновки:* запропоновано виправлення обох правил, протокол сліпої позавибіркової перевірки та пре-реєстрований план порівняльного експерименту.

КЛЮЧОВІ СЛОВА: ex-ante оцінювання; композитний індекс; компенсаторність агрегації; референс-клас; підлабузництво мовних моделей.

ВСТУП. Ex-ante оцінювання дослідницьких і проєктних заявок — стадія, на якій ухвалюються рішення про розподіл ресурсів за відсутності даних і прототипу, — формалізоване значно слабше, ніж ex-post оцінювання результатів. Міжнародна реформа оцінювання, започаткована Лейденським маніфестом (Hicks et al., 2015), Сан-Франциською декларацією та Угодою про реформування оцінювання досліджень (CoARA, 2022), зосереджена насамперед на відмові від бібліометричних проксі при оцінюванні вже виконаної роботи. Процедури ранньої стадії лишаються переважно бальними.

До нових процедур раннього відбору зазвичай висувають три вимоги: згортка критеріїв не повинна дозволяти сильному критерію компенсувати фатальну ваду за іншим; підсумковий бал має бути інтерпретовним відносно калібрувальної вибірки; процедура має зберігати стійкість при залученні великих мовних моделей (далі — ВММ) як асистентів експерта.

Кожен складник цих вимог має власну літературу. Поетапний відбір із гейтами зростаючої строгості описано в моделі Stage-Gate (Cooper, 1990). Порівняння проєкту з класом історичних аналогів замість внутрішнього прогнозу — предмет прогнозування за референс-класом (Kahneman & Lovaglio, 1993; Flyvbjerg, 2006). Поведінку різних схем агрегації композитних показників, зокрема співвідношення адитивної та геометричної форм, систематизовано в посібнику ОЕСР і Спільного дослідницького центру (Nardo et al., 2008), де прямо зазначено, що геометрична агрегація знижує компенсаторність, але не усуває її. Систематичне зміщення ВММ у бік позиції співрозмовника — підлабузництво — показано експериментально (Sharma et al., 2023), а надійність ВММ саме як «першого гейта» методологічної придатності дослідницьких пропозицій оцінено на спеціалізованому бенчмарку (Ho et al., 2026).

Невирішена частина проблеми полягає не у відсутності складників, а в тому, що правила агрегації та коригування оцінки, з яких складається конкретна процедура, зазвичай не піддаються стрес-тестуванню на власному діапазоні допустимих значень. Заявки на «некомпенсаторність» і на

«стійкість до підлабузництва» найчастіше лишаються деклараціями, не перевіреними ані аналітично, ані експериментально.

Мета роботи — задокументувати конструкцію гейтованої процедури ex-ante оцінювання (робоча назва IDEA-GATE), виконати її самооудит за наперед визначеним протоколом, оприлюднити виявлені вади разом із виправленнями та окреслити межі застосовності. Робота свідомо не претендує на статус валідованого прогностичного інструмента.

МЕТОДИ. Архітектура. Процедура складається з трьох стадій, розділених гейтами; вартість кожної наступної стадії приблизно на порядок вища за попередню (години — дні — тижні). Дешеві перевірки виконуються першими, тож відхилення на ранній стадії зупиняє витрати.

Стадія 1 — бінарні гейти. Ідея канонізується у формі «механізм — ефект — застосування» і проходить п'ять перевірок: узгодженість із законами збереження та термодинамічними межами; наявність змістовної відповіді на питання, чому цього ще не зроблено і що змінилося; наявність попереднього рівня техніки; визначеність інкумбента, відносно якого рахуватиметься перевага; вимірюваність заявленого ефекту стандартними методиками. Кожен гейт дає бінарний результат, а їх добуток G належить множині $\{0, 1\}$. Дефолтний вердикт стадії — відхилити; регламент вимагає сформулювати найсильніший контраргумент до будь-яких аргументів на захист.

Стадія 2 — індекс перспективності. Обчислюється зважене геометричне середнє шести нормованих факторів, помножене на добуток гейтів: $PI = G \cdot \prod x_j^{w_j}$, де $\sum w_j = 1$. Фактори та ваги: перевага над прогнозованим інкумбентом Δ (0,20); масштаб задачі S (0,15); добуток необхідних технічних умов (0,20); добуток бар'єрів прийняття (0,20); капітальна ефективність, вимірювана вартістю й часом до доказу, а не до продукту (0,10); геометричне середнє захищеності та своєчасності (0,15). Усі фактори нормуються у відрізок $[0,02; 1]$, тобто нижня межа нормування становить 0,02.

Класи доказовості та коригування оцінки. Кожен вхід маркується класом А–Е за походженням значення: А — пряме вимірювання, Е — здогад. Класу відповідають вага довіри λ (1,00; 0,80; 0,50; 0,30; 0,10) і коефіцієнт примусового розширення інтервалу (1,0; 1,5; 2,5; 4,0; 6,0). Спільна вага довіри обчислюється як середнє арифметичне λ за вісьмома входами. Підсумкова оцінка коригується за правилом $PI = \lambda \cdot PI_{\text{raw}} + (1 - \lambda) \cdot PI_{\text{ref}}$, де $PI_{\text{ref}} = 0,405$ — медіана індексу на калібрувальній вибірці. Це евристичне стиснення оцінки; воно не виводиться з явної ймовірнісної моделі, а PI_{ref} є мірою центральної тенденції індексу, а не базовою ймовірністю успіху.

Стохастичне оцінювання. Виконується 20 000 реалізацій методом Монте-Карло; входи задаються PERT-розподілами за трійкою «мінімум — мода — максимум» із параметром форми 4. Виходом є розподіл індексу та ймовірність перевищення порога. Вплив кожного входу оцінюється абсолютною ранговою кореляцією Спірмена між його вибіркою та вибіркою індексу і подається діаграмою «торнадо».

Калібрувальна вибірка. Референс-клас містить 22 історичні технологічні кейси — 11 успіхів і 11 провалів, — оцінені ретроспективно за станом знань на рік початку проєкту. Успіхом вважається досягнення масового впровадження або створення ринку понад 1 млрд дол. на рік протягом п'ятнадцяти років від старту. Для довільної ідеї обчислюються перцентиль у вибірці та частка успішних серед семи найближчих за індексом сусідів.

Стадія 3 — вимірювання. Пріоритет вимірювань формується з верхівки торнадо; для кожного параметра складається паспорт із посиланням на стандартну методику й оцінкою метрологічної придатності, технічна готовність описується шкалою рівнів TRL (NASA, 2020), а пороги ухвалення рішення пре-реєструються до початку експерименту (Nosek et al., 2018).

Протокол самооудиту. Аудит виконано за чотирма наперед визначеними кроками: підстановка граничних допустимих значень факторів у формулу індексу та перевірка, чи здатна вона відхилити

ідею з одним майже нульовим фактором; розв'язання нерівності «мінімальний досяжний індекс перевищує поріг» відносно ваги довіри λ ; перевірка внутрішньої узгодженості наведених ілюстративних кейсів із задекларованими порогами; зіставлення заявленого захисту від підлабузництва з наявними емпіричними результатами. Обчислювальне ядро реалізовано мовою Python 3 без зовнішніх залежностей і містить режим самотестування.

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ. Розділення на калібрувальній вибірці та межі цього результату. За наївного означення переваги площа під ROC-кривою становить 0,959; за проєкції інкумбента вперед — 1,000. Оптимальний поріг за індексом Юдена (Youden, 1950) дорівнює 0,35; медіана вибірки — 0,405; для профілю дослідницької лабораторії застосовується знижений поріг 0,28.

Друге значення не є доказом прогнозної спроможності щонайменше з чотирьох причин. Значення факторів для кожного кейса виставляв автор, який знав його результат, — ретроспективна контамінація. Обсяг вибірки у 22 точки дає дуже широкий довірчий інтервал оцінки площі під кривою. Кейси відібрано з відомих і добре задокументованих, тоді як ідеї, що провалилися непомітно, у вибірку не потрапили, — це упередження виживання всередині самої калібрувальної вибірки. Нарешті, ідеальне розділення на калібрувальній вибірці є очікуваною поведінкою моделі з шістьма параметрами на 22 точках.

Окремо слід підкреслити, що штучний баланс 11/11 не відтворює реальної частоти успіху, тому частка успішних серед найближчих сусідів (29 % і 100 % у наведених нижче прогонах) є характеристикою вибірки, а не оцінкою ймовірності успіху нової ідеї. Позавибіркову перевірку не можна створити приховуванням назв тих самих кейсів: історичні приклади лишаються впізнаваними за описом. Потрібні інші кейси, не використані при конструюванні процедури, і незалежний оцінювач, який виставляє фактори до розкриття результатів.

Вада 1: компенсаторність знижено, але не усунуто. Твердження про «некомпенсаторність» геометричної згортки є неточним і узгоджується з відомим застереженням (Nardo et al., 2008). Нуль за будь-яким фактором справді обнуляє добуток, проте нуль недосяжний: нормування обмежує фактори знизу значенням 0,02. Підстановка граничного значення дає пряме спростування заявки: за одного фактора з вагою 0,20 на нижній межі та решти факторів, рівних одиниці, індекс становить $0,02^{0,20} \approx 0,457$, тобто перевищує поріг 0,35 попри фактично нульовий критерій.

Отже, строго некомпенсаторною є лише підсистема бінарних гейтів стадії 1, тоді як згортка стадії 2 забезпечує знижену компенсаторність. Виправлення полягає в тому, щоб винести неприпустимі значення з неперервної шкали у гейтову: для кожного фактора задається мінімально допустиме значення, недосягнення якого є окремою підставою для відхилення незалежно від індексу. Відповідно скориговано і формулювання назви процедури.

Вада 2: коригування оцінки інвертується за слабких доказів. Евристичне стиснення до PI_{ref} задумане як запобіжник проти високих балів на здогадах, і в цій частині воно працює: за низької ваги довіри верхня межа індексу знижується. Проте те саме правило одночасно підтягує вгору низькі оцінки. Мінімальний досяжний індекс становить 0,02 (усі фактори на нижній межі), тому за $\lambda = 0,10$, що відповідає класу доказовості E, увесь досяжний діапазон стискається до відрізка $[0,367; 0,465]$. Розв'язання нерівності $\lambda \cdot 0,02 + (1 - \lambda) \cdot 0,405 > 0,35$ відносно λ дає $\lambda < 0,143$: за ваги довіри нижче цього значення жодна ідея не може бути відхилена за порогом, хоч би якими слабкими були її фактори.

Виправлення двоскладове. По-перше, індекс не може використовуватися як правило «пройшов — не пройшов» за низької ваги довіри: якщо λ менша за наперед зафіксоване λ_{min} або дев'яностовідсотковий інтервал перетинає поріг, єдиним допустимим вердиктом є «вимірювати», а не «прийняти» чи «відхилити». По-друге, порівнювати слід величини одного масштабу: або

нескоригований індекс із порогом 0,35, або скоригований індекс зі скоригованим порогом $\lambda \cdot 0,35 + (1 - \lambda) \cdot PI_{ref}$. Змішування шкал і породжує описану інверсію.

Прогін конвєєра та внутрішня узгодженість ілюстрацій. Для відомого провалу (короткий преміальний відеоконтент, знання станом на 2018 рік) отримано медіану 0,373 за дев'яносто-відсоткового інтервалу [0,305; 0,463] і $\lambda = 0,51$; для відомого успіху (мРНК-вакцинна платформа, 2010 рік) — 0,468. Розрив у 0,095 малий, і це коректна поведінка: за входів класів С–Е стиснення притискає обидві оцінки до PI_{ref} , і процедура фактично повідомляє, що за наявних доказів не розрізняє випадки. Торнадо в обох прогонах вказує на бар'єри прийняття як на вхід із найбільшим рангом впливу.

Перевірка узгодженості виявила помилку в одній з двох ілюстрацій, які раніше подавалися як структурні перевороти рішення. Для глобального супутникового телефонного зв'язку (1990) індекс знижується з 0,320 до 0,201 при переході до проєктованого інкумбента. Обидва значення лежать нижче загального порога 0,35, тому опис цієї зміни як переходу «із зони успіху до зони відхилення» є хибним. Коректне формулювання: зміна перетинає лише знижений поріг лабораторного профілю 0,28 і, головне, змінює ранг кейса у вибірці. Далі цей приклад слід подавати як зсув перцентилів, а не як зміну вердикту.

Друга ілюстрація потребує уточнення правила, а не лише формулювання. Підвищення оцінки кремнієвої фотовольтаїки (1975) з 0,252 до 0,461 не впливає з правила в тому вигляді, як воно було записане: якщо вперед проєктується лише інкумбент, здешевлення самої фотовольтаїки на перевагу не впливає. Коректне правило вимагає проєктувати обидві альтернативи, кожен за власною кривою навчання. Крім того, закон Райта (Wright, 1936) пов'язує питомі витрати з накопиченим випуском, тож коефіцієнта навчання недостатньо: потрібен сценарій нарощування випуску до цільового року, а статистичні властивості таких прогнозів і їхні довірчі інтервали слід брати з відповідних оцінок (Nagy et al., 2013). Без цих уточнень приклад лишається правдоподібною історією, а не відтворюваним розрахунком.

Захист від підлабузництва є гіпотезою, а не результатом. Припущення, що інверсія дефолтного вердикту та обов'язковий контраргумент до захисту усувають зміщення ВММ, у цій роботі не перевірено. Ба більше, наявні дані свідчать проти простого варіанта цього припущення: на бенчмарку з 1099 дослідницьких пропозицій і дванадцяти сучасних моделей показано, що суворіші інструкції переважно зміщують помилки від необґрунтованого схвалення до необґрунтованого відхилення, а не усувають їх (Ho et al., 2026). Модель здатна побудувати слабкий контраргумент, спростувати його й погодитися з автором; дефолт «відхилити» водночас підвищує частку помилково відхилених придатних пропозицій.

Тому заявку переформульовано як гіпотезу з пре-реєстрованим планом перевірки. План передбачає три експериментальні умови — звичайний запит, запит із загальною вимогою критичності та повний протокол з інверсією дефолту, — фіксований перелік моделей із версіями, не менш ніж п'ять повторних прогонів на кожен пропозицію з фіксованими параметрами генерації, збалансований набір пропозицій із незалежно встановленою придатністю та роздільне звітування часток хибнопозитивних і хибнонегативних рішень. Пороги успіху фіксуються до збору даних. До виконання такого експерименту протокол слід описувати як конструктивну гіпотезу.

Уточнення окремих тверджень. Добуток імовірностей семи необхідних умов правомірний лише за їх незалежності; за наявності залежності потрібні умовні ймовірності, і приклад із $0,8^7 \approx 0,21$ слід читати як ілюстрацію принципу, а не як розрахунок. Абсолютна рангова кореляція Спірмена є мірою рангового впливу входу, а не часткою поясненої дисперсії, тому діаграму торнадо коректно описувати як упорядкування впливів. Верхівка торнадо не тотожна очікуваній цінності вибіркової інформації: остання залежить також від вартості вимірювання, його точності та здатності змінити рішення, тож торнадо є попереднім фільтром для розрахунку цінності інформації, а не її

замінником. Нарешті, зіставлення історичних витрат із гіпотетичною вартістю ранньої перевірки є контрфактичним сценарієм, а не встановленим ефектом; наводити його як досягнуту економію неприпустимо.

Межі застосовності. Критерії переваги над інкумбентом, бар'єрів прийняття, капітальної ефективності та захищеності орієнтовані на комерціалізацію технологій. Дослідження може бути науково цінним без ринкового конкурента й патентної захищеності, а вимога вимірювати ефект наявними стандартними методиками здатна відсіювати роботи, метою яких є створення нового методу вимірювання. Тому назву й висновки звужено до раннього оцінювання прикладних дослідницьких та інноваційних проєктів. Ваги факторів є експертним судженням і потребують аналізу чутливості; референс-клас домінований апаратними й енергетичними технологіями, тож для програмних і соціальних інновацій потрібна власна калібрувальна вибірка; процедура оцінює ідею, а не її виконання.

Що лишається внеском. Після виправлень внеском роботи є не сама композиція відомих елементів, а результат її самоаудиту: показано, за яких саме значень параметрів заявлені властивості процедури перестають виконуватися, і сформульовано умови, за яких композитний індекс не має використовуватися як вирішальне правило. Ці умови переносяться на інші бальні процедури незалежно від решти методики.

Висновки

1. Зважене геометричне середнє з ненульовою нижньою межею нормування не забезпечує некомпенсаторності: фактор на межі 0,02 з вагою 0,20 дає індекс 0,457 за порога 0,35. Строго некомпенсаторною лишається лише підсистема бінарних гейтів, а неприпустимі значення окремих факторів слід виносити в гейтову шкалу як окремі мінімально допустимі значення.
2. Евристичне стиснення оцінки до медіани калібрувальної вибірки інвертується за слабких доказів: за ваги довіри нижче 0,143 увесь досяжний діапазон індексу лежить вище порога і жодну ідею не може бути відхилено. Індекс не має застосовуватися як правило «пройшов — не пройшов» за низької ваги довіри; порівнювати слід величини однакового масштабу.
3. Площа під ROC-кривою, що дорівнює одиниці на калібрувальній вибірці з 22 ретроспективно оцінених кейсів, не є свідченням прогнозної спроможності; штучний баланс успіхів і провалів не дозволяє тлумачити частку успішних сусідів як імовірність. Позавибіркова перевірка потребує інших кейсів і незалежного оцінювача.
4. Твердження про архітектурну протидію підлабузництву мовних моделей є гіпотезою; наявні бенчмарки свідчать, що суворіші інструкції зміщують структуру помилок, а не усувають їх. Наведено пре-реєстрований план порівняльного експерименту.
5. Практичне значення роботи полягає у переносних вимогах до бальних процедур раннього відбору: виносити неприпустимі значення в окремі гейти; фіксувати ваги до обчислення; декларувати клас доказовості входів; не змішувати шкали при порівнянні з порогом; не подавати контрфактичні оцінки економії як установлений ефект.

ПОДЯКИ. Дослідження виконано у вільний від роботи в Інституті загальної та неорганічної хімії імені В. І. Вернадського Національної академії наук України час. Додаткового зовнішнього фінансування не залучалося.

ДОСТУПНІСТЬ ДАНИХ. Калібрувальна вибірка з 22 історичних кейсів із ретроспективними оцінками факторів, правила нормування, означення успіху та обчислювальне ядро з режимом самотестування депонуються у відкритому репозиторії Zenodo з присвоєнням DOI; посилання буде наведено у фінальній версії. Депонований набір містить лише публічні історичні кейси. Текстові протоколи стадій 1 і 3, а також оцінки конкретних неопублікованих ідей автора до набору не входять, оскільки не потрібні для відтворення наведених числових результатів.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

- CoARA. (2022). Agreement on reforming research assessment. Coalition for Advancing Research Assessment. <https://www.coara.org/agreement/>
- Cooper, R. G. (1990). Stage-gate systems: A new tool for managing new products. *Business Horizons*, 33(3), 44–54. [https://doi.org/10.1016/0007-6813\(90\)90040-I](https://doi.org/10.1016/0007-6813(90)90040-I)
- Flyvbjerg, B. (2006). From Nobel Prize to project management: Getting risks right. *Project Management Journal*, 37(3), 5–15. <https://doi.org/10.1177/875697280603700302>
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429–431. <https://doi.org/10.1038/520429a>
- Ho, S.-T., Liu, M., Nghiem, H., & Huang, F. (2026). SoundnessBench: Can your AI scientist really tell good research ideas from bad ones? *arXiv*. <https://doi.org/10.48550/arXiv.2605.30329>
- Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, 39(1), 17–31. <https://doi.org/10.1287/mnsc.39.1.17>
- Nagy, B., Farmer, J. D., Bui, Q. M., & Trancik, J. E. (2013). Statistical basis for predicting technological progress. *PLoS ONE*, 8(2), e52669. <https://doi.org/10.1371/journal.pone.0052669>
- Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
- NASA. (2020). Technology readiness assessment best practices guide (NASA/SP-20205003605). National Aeronautics and Space Administration. <https://ntrs.nasa.gov/citations/20205003605>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., ... Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv*. <https://doi.org/10.48550/arXiv.2310.13548>
- Strathern, M. (1997). «Improving ratings»: Audit in the British University system. *European Review*, 5(3), 305–321. [https://doi.org/10.1002/\(SICI\)1234-981X\(199707\)5:3<305::AID-EURO184>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4)
- Wright, T. P. (1936). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, 3(4), 122–128. <https://doi.org/10.2514/8.155>
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)

Kolomiets Yevhen,

V. I. Vernadsky Institute of General and Inorganic Chemistry,
National Academy of Sciences of Ukraine, Kyiv, Ukraine,
ORCID: 0000-0002-8941-663X, e-mail: kolomiets@ionc.kiev.ua

**EX-ANTE EVALUATION OF APPLIED RESEARCH AND INNOVATION PROJECTS:
A SELF-AUDIT OF A GATED PROCEDURE WITH REDUCED-COMPENSABILITY
AGGREGATION**

ABSTRACT. *The aim is to document the design of a gated ex-ante evaluation procedure for applied research and innovation projects and to subject it to a self-audit rather than assert its advantages. **Methods:** analytical stress-testing of the aggregation and score-correction rules across the full admissible range of factor values; a run on a calibration set of 22 historical cases; and comparison of the claimed safeguard against language-model*

bias with existing benchmarks. **Results:** two design defects were found. The weighted geometric mean reduces but does not eliminate compensability — a factor at the normalisation floor of 0.02 with weight 0.20 still yields an index of 0.457 against a threshold of 0.35. The evidence-class correction inverts: below a confidence weight of 0.143 the whole attainable range lies above the threshold, so no idea can be rejected. **Conclusions:** corrections to both rules, a blind out-of-sample protocol and a preregistered experiment are proposed.

KEYWORDS: ex-ante evaluation; composite indicator; compensability of aggregation; reference class; language-model sycophancy.