AI diplomacy needs a clear chain of responsibility

Foreign ministries should test whether AI improves approved work before expanding its role.

RESEARCH BRIEFING · STRATEGY · 11 NOVEMBER 2026

Foreign ministries can produce briefs faster with AI, but faster drafting does not establish better policy. A useful pilot measures the final approved document, the errors corrected and the staff time required to reach a decision.

Key points

• Separate source evidence, analytical assessment and authority to act.

• Keep conditions and revisions in the institutional record.

• Compare approved outputs on matched tasks before expanding a pilot.

Productivity depends on the task

An experiment with 758 consultants found gains on tasks suited to early GPT-4: speed improved by more than 25%, human-rated performance by more than 40% and completion by more than 12%, according to Harvard Business School’s 2023 summary. These are experimental task outcomes, not ministry-wide gains. [1]

The same caution applies to adoption surveys. Among generative-AI users in an OECD survey of more than 5,000 SMEs, 65% reported better employee performance but 26% reported higher revenue. The survey covered seven countries in late 2024; neither result measures diplomatic effectiveness. [2]

Assign three distinct responsibilities

A source reviewer establishes what a document supports. An analyst explains the inference and competing interpretations. An authorised official approves the position. One person may perform several roles, but the record should distinguish them.

For a material claim, keep its source, date and supporting passage. For an assessment, record uncertainty and the evidence that would change it. For an instruction, record the approver, scope and conditions. NIST identifies confabulation and human–AI interaction as areas requiring risk management. [3]

Preserve conditions as documents change

A searchable archive should show whether a proposal was accepted, rejected, amended or left open. Retain the source text beside any machine interpretation. Record successive instructions rather than overwriting the earlier position.

Test the system with a revised negotiating draft. Can a new team recover which concession depended on verification? Can it distinguish a minister’s public statement from an operative instruction? These are proposed evaluation questions, not reported test results.

Measure one complete workflow

Choose a recurring product, such as a weekly negotiation brief. Record case difficulty, source volume, preparation time, review time and material corrections. Compare assisted and unassisted cases under the same approval standard.

Count abandoned uses and failed citations. Report whether the final approved recommendation changed and why. A pilot should end with a documented choice to expand, revise or stop, supported by quality, effort and recurring-cost evidence.

Sources & research notes

01 / Harvard Business School AI Institute / Jagged Technological Frontier (2023)

https://aiinstitute.hbs.edu/navigating-the-jagged-technological-frontier/
758 consultants; experimental tasks and early GPT-4. Results are not firm-wide returns or estimates for current tools.

02 / OECD / Generative AI and the SME Workforce (2025)

https://www.oecd.org/en/publications/generative-ai-and-the-sme-workforce_2d08b99d-en.html
Survey of more than 5,000 SMEs in seven countries in late 2024. Reported benefits are respondents’ assessments, not causal estimates.

03 / NIST / Generative Artificial Intelligence Profile, AI 600-1 (2024)

https://doi.org/10.6028/NIST.AI.600-1
Voluntary risk-management guidance, including confabulation and human–AI interaction; not a certification.

Independent TACITUS analysis. Research checked 11 November 2026. Published findings retain their dates and limits. Proposed curricula, tests, schedules and worked calculations are labelled analytical examples, not client results.